Karlsruhe Institute of Technology

Enhanced sampling methods applied to chemical reactions in proteins.

zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

von der KIT-Fakultät für Chemie und Biowissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Denis Mario Maag

1. Referent: Prof. Dr. Marcus Elstner
2. Referent: Prof. Dr. Alexander Schug
Tag der mündlichen Prüfung: 27. April 2022

Zusammenfassung

Der erste Teil dieser Arbeit untersucht die mechanistischen Details von Protonentransferreaktionen in zwei Rhodopsinen, Bacteriorhodopsin (BR) und Histidin-Kinase-Rhodopsin (HKR). Beide Proteine sind Transmembran-Photorezeptoren mit einem kovalent gebundenen Retinal, das über eine Schiffsche Base mit einem Lysin verbunden ist. Nach Absorption eines Photons isomerisiert das Retinal und ein Photozyklus wird in Gang gesetzt. Hierbei kommt es zu einer Kaskade von strukturellen Veränderungen die zu mehreren Protonentransfers (PTs) führen.

BR nutzt die Energie des Lichts, um ein Proton aus der Zelle zu transportieren. Obwohl die Forschung seit mehr als 45 Jahren andauert, blieb der Mechanismus des letzten PT Schritts weitgehend unklar. Wir zeigen, dass die Reaktion über einen Protonen-Loch-Mechanismus abläuft, der durch einen hochkonservierten Arginin-Rest stabilisiert wird. Freie Energie Rechnungen, die mit quantenmechanischen/molekularmechanischen (QM/MM) Simulationen durchgeführt wurden, weisen eine gute Übereinstimmung mit den verfügbaren experimentellen Daten auf.

HKR besitzt zwei stabile Zustände, den blauabsorbierenden Rh-Bl- und den ultravioletabsorbierenden Rh-UV-Zustand, die durch Isomerisierung des Retinals reversibel photokonvertiert werden können. Bei jedem Übergang findet ein Protonentransfer zwischen der Schiff-Base und einem nahegelegenen Aspartat-Rest statt. Über die Funktion und physiologische Rolle von HKR ist wenig bekannt. Des Weiteren gibt es keine Kristallstruktur. Unser Ziel ist es, ein zuvor erstelltes Homologiemodell von HKR zu validieren, indem wir die berechneten Anregungsenergien mit denen aus Experimenten vergleichen. Darüber hinaus wird das freie Energieprofil des ersten PTs mit QM/MM-Simulationen berechnet.

Im zweiten Teil der Thesis werden Berechnungen der freien Energie mit einem neuartigen Verfahren durchgeführt, das Mulliken-Ladungen in der Dichtefunktionalen Tight-Binding-Methode (DFTB) als Reaktionskoordinaten verwendet. Der Ansatz wird auf protonengekoppelte Elektronentransfers in einem Modellsystem angewandt und ausgiebig getestet, indem verschiedene Ladungszustände und Konformationen des Systems in der Gasphase und in wässriger Lösung betrachtet werden.

Der dritte Teil befasst sich mit den strukturellen Merkmalen und der Energetik von Thiol-Disulfid-Austauschreaktionen. Die Auswirkung sterischer und elektrostatischer Beiträge auf die Regioselektivität in einem Protein unter mechanischem Stress wird in 334 QM/MM-Simulationen mit einer gesamten Simulationsdauer von 5.7 µs untersucht. Wir zeigen, dass die molekulare Umgebung die Reaktion beeinflusst, indem sie die Nukleophilie und Elektrophilie der beteiligten Schwefelatome modifiziert. Dies wird zusätzlich in einem Modellsystem demonstriert, in dem verschiedene externe elektrostatische Potentiale angelegt wurden.

Darüber hinaus wird die Genauigkeit der DFTB Methode zur Beschreibung von Thiol-Disulfid-Austauschreaktionen überprüft. Quantitative und qualitative Fehler werden mit einem künstlichen neuronalen Netz und einer speziellen Reaktionsparametrisierung (SRP) korrigiert. Beide Ansätze werden miteinander verglichen, indem der Thiol-Disulfid-Austausch in einem Modellsystem und Blutprotein mit QM/MM-Metadynamik simuliert wird. Wir zeigen, dass die Methoden bei geringen Rechenkosten hochpräzise freie Energieflächen erzeugen.

Abstract

The first part of this thesis aims to provide mechanistic details about proton transfer reactions in two rhodopsins, bacteriorhodopsin (BR) and histidine kinase rhodopsin (HKR). Both proteins are transmembrane photoreceptors with a covalently bound retinal, that is linked to a lysine via a Schiff base. Upon photon absorption, the retinal isomerizes and triggers a photocycle in which a cascade of structural rearrangements takes place, accompanied by proton transfers (PTs).

BR uses the energy of the light to release one proton to the extracellular side. Despite more than 45 years of research, the mechanism of the final long-range PT remained largely unclear. We show that the reaction proceeds via a proton-hole mechanism that is stabilized by a highly conserved arginine residue. The free energy profile, obtained with extensive quantum mechanical/molecular mechanical (QM/MM) free energy simulations, is in good agreement with available experimental data.

HKR exhibits two stable states, blue absorbing Rh-Bl and ultraviolet absorbing Rh-UV, which can be reversibly photoconverted by isomerization of the retinal. Each transition features a proton transfer between the Schiff base and a nearby aspartate residue. Little is known about the function and physiological role of HKR. Furthermore, there is no crystal structure available. We aim to validate homology model of HKR by comparing excitation energies obtained from the model with those from experiments. In addition, the free energy profiles of the first PT is calculated with QM/MM free energy simulations.

In the second part, free energy calculations are performed with a novel framework that uses Mulliken charges in the density-functional tight binding (DFTB) method as reaction coordinates. The approach is applied to proton-coupled electron transfers in a model system and extensively tested by considering different charge states and conformations of the system in the gas phase and in aqueous solution.

The third part addresses the structural features and energetics of thiol-disulfide exchanges. The effect of steric and electrostatic contributions on the regioselectivity in a protein under mechanical stress is investigated in 334 QM/MM simulations with a total sampling time of 5.7 µs. We show that the molecular environment directs the reaction by modulating the nucleophilicity and electrophilicity of the involved sulfur atoms. This is further demonstrated in a model system with varied artificial external electrostatic potentials.

In addition, the accuracy of DFTB for the description of thiol-disulfide exchange is reviewed. Quantitative and qualitative errors are corrected with an artificial neural network and a special reaction parameterization (SRP). Both approaches are benchmarked with QM/MM metadynamics of thiol-disulfide exchange in a model system and blood protein. We show that the methodologies generate highly accurate free energy surfaces at low computational cost.

Contents

List of Figures

List of Tables

Part I.

Introduction

1. Introduction

1.1. Proton transfer in Rhodopsins

Rhodopsins are photoreceptors that consist of seven transmembrane helices and a light-absorbing retinal, that is covalently bound to a lysine via a Schiff base. After absorption of the light, the retinal isomerizes which leads to conformational changes in the protein. These changes are required to fulfill a biological task, such as pumping protons across the membrane. In this thesis, two rhodopsins are investigated, bacteriorhodopsin and histidine kinase rhodopsin.

1.1.1. Bacteriorhodopsin

Bacteriorhodopsin (bR) was discovered in the early seventies in *H. salinarum*, a halophilic archaea living in environments with extreme salt concentrations.[1] When the salt concentration is reduced or removed, the membrane of *H. salinarum* disintegrates into fragments of deep purple color due to the covalently bound retinal with an absorption maximum at 570 nm.[2,3] The energy of the absorbed light is used to pump a proton across the cell membrane during a so-called photocycle. This creates a proton gradient, which is converted into chemical energy as adenosine triphosphate (ATP) by ATP synthase.[4,5] Only a few years later, the crystal structures of bR was obtained in 1975, making it the first membrane protein ever for which structural information was available.[6] The low-resolution structure (7 Å) revealed that three monomeric bR units form a trimer in the membrane with a peptide-to-lipid ratio of 3:1. Each monomer has a molecular weight of 24 kDa and consists of 248 residues which form seven transmembrane α-helices spanning the membrane almost perpendicular.

In the years that followed, many other members of the bacterial rhodopsin family were discovered, such as halorhodopsin which is an inward directed chloride pump,[7] and bR soon became a model system for several applications in structural biology[8] due to its high stability and relative structural simplicity. In the past 45 years, the combined effort of many research groups using site-directed mutagenesis, vibrational spectroscopy, X-ray crystallography, solid-state and molecular dynamics simulations lead to a widely accepted model of the photocycle, which involves a sequence of spectral states and five sequential proton transfer (PT) steps between several key residues. Although many of these states have been characterized with crystallography using cryotrapping techniques, more than 153 X-ray crystallographic structures with resolutions up to 1.25 Å are available in the protein data bank,[9] there is still no consensus on the complete reaction mechanism.[8,10,11]

Especially the mechanism of the long range PT during the final transition back to the ground state (**O→bR**) remains largely unclear. In **chapter 7** we explore the structural

features of the different states involved in the transition and compute the three-dimensional free energy surface of the PT with ~ 0.1 µs QM/MM multiple walker metadynamics. We show that the PT proceeds via a proton hole mechanism that is coupled to a structural rearrangement of a nearby arginine.

1.1.2. Histidin Kinase Rhodopsin

Histidine kinase rhodopsin (HKR) is a microbial rhodopsin found in the green alga *Chlamydomonas reinhardtii*.[12,13] It exhibits two stable isoforms: the blue-absorbing Rh-Bl state with an absorption maximum of λ_{max} = 486 nm, and the UVA-absorbing Rh-UV state with λ_{max} = 379 nm. They are stable for longer than 24 h and can be converted into each other by illumination with blue or UVA light, respectively. Both states and additional intermediates of the photocycle were characterized by laser flash photolysis data and pump-probe experiments.[13–15]

In RH-Bl, the Schiff base is protonated and the retinal in the 13-*cis*,15-*anti* configuration. Upon absorption of blue light, the retinal isomerizes to 13-*cis*,15-*syn* and an intermediate state (Int1) is formed within 0.6 and 5 ps. The Int1 state decays to the second intermediate P550 with a time constant of 453 ps, followed by the third intermediate P570 after additional ~2.5 ms. The absorption maxima of all intermediates are red shifted compared to the Rh-Bl state. Subsequently, the Rh-UV state is formed when the proton of the Schiff base is transferred to D239 within 27 ms, which corresponds to a barrier of ~60 kJ/mol with transition state theory.

When the Rh-UV state is illuminated by UVA light, a intermediate Rh-UV' state forms within 5 and 60 ps. Then, within a time constant of 0.1 and 3.4 ms, the retinal isomerizes back to 13-*cis*,15-*anti* and the Schiff base is reprotonated by D239.

The stability of the Rh-UV state may be explained by comparing the amino acid sequence with that of bacteriorhodopsin. In HKR, the highly conserved bR residues D85 and D96 are substituted by a methionine and a leucin. D85 serves as a counterion in bR and is protonated by the Schiff base at the beginning of the photocycle. D96 reprotonates the Schiff base and is subsequently reprotonated from the intracellular side in the photocycle of bR. Due to the absence of D85 and D96 homologues in HKR, D239 (D212 in bR) serves as counterion and proton acceptor and the deprotonated Rh-UV state might be stabilized since the Schiff base might no get reprotonated by the intracellular side. The exact function of HKR and its physiological role remains still largely unclear. In addition, no crystal structure is available, which complicates computational studies.

Thus, a homology model of HKR was built[16] which we aim to validate in **chapter 8**. QM/MM simulations of the Rh-Bl to Rh-UV transition are performed and the excitation energies of the intermediates calculated. Moreover, the free energy profile of the PT from the Schiff base to D239 is obtained by QM/MM free energy calculations.

1.2. Proton-coupled electron transfer

In numerous biological processes proton transfers are coupled to electron transfers in so-called proton-coupled electron transfer (PCET) reactions, which can occur in a sequential

or concerted manner.[17,18] The sequential pathway can proceed via two mechanisms: first an electron transfer (ET) followed by a proton transfer (PT), or a PT followed by an ET. In the concerted proton-electron transfer (CPET), both transitions occur in one step. Both mechanisms can proceed either orthogonal or collinear. In the first case, the electron donor-acceptor pair and the proton donor-acceptor pair are the same and in the latter they are different.

The time scales of the ET and PT depend significantly on the transfer distance as well as on the environment, e.g. water or protein. A prime example for PCET reactions in a protein is the reduction of ribonucleotides in the enzyme ribonucleotide reductase (RNR).[19,20] The catalytic reaction features a long-range ET over ~32 Å, via several PCETs along a pathway of redox-active amino acids. Most of the involved amino acids are tyrosines of which one (Y731) has to dynamically rearrange in order to facilitate the ET and PT.[21] Thus, conformational changes are from great importance for the correct description of PCETs.

PCET reactions have been extensively studied over the last decades and many simulations protocols have been developed, for example by Cukier[22–24] or Hammes-Schiffer and co-workers[25–27]. There are many great reviews and perspectives which describe and summarize the variety of available methods and their applications.[28–34] The general goal of all methods is to obtain the correct reaction energy profiles of the studied systems.

A new concept of free energy calculations was introduced by Gillet *et al.*[35] in 2018, where Mulliken charges in the density-functional tight binding (DFTB) method were used as reaction coordinates in biasing potential simulations. The method performed very well for QM systems, however, failed to describe QM/MM systems accurately due to certain missing derivatives. Moreover, the method was only available for second order DFTB.

The framework was extended to third order DFTB and the missing gradients were implemented. The approach is thoroughly tested in **chapter 9** by simulating PCET reactions in 32 setups of two tyrosines, which differ in their charge state, conformation and environment.

1.3. Disulfide bonds and thiol-disulfide exchange

Disulfide bonds in proteins fulfill many important tasks. Formed between two cysteins, which may be on the same or on different peptide strands, they stabilize the tertiary and quaternary structure and assist protein folding. Moreover, so-called allosteric disulfide bonds can trigger or inhibit the function of proteins, either by being cleaved or formed.[36,37] In other cases, disulfide bonds can act as catalysts for certain reactions. For example, the disulfide bonds of the redox protein families thioredoxin, glutaredoxin and protein disulfide isomerase reduce the disulfide bonds of their substrates by thiol-disulfide exchange.[38,39]

Thiol-disulfide exchange is an S_N2 reaction between a thiolate anion $R_1–S^-$ and a disulfide bond $R_2–S–S–R_3$, leading to a newly formed disulfide bond, either $R_1–S–S–R_2$ or $R_1–S–S–R_3$.[40] The nucleophilic thiolate can be a non-protein molecule (for example glutathion) or a cysteine located on a different or the same protein. Such inter- or intramolecular rearrangements can be triggered by shear stress, for example from flowing blood. The mechanical force acting on the protein stretches the disulfide bonds which

leads to a decrease of the activation energy for a nucleophilic attack.[41,42] Consequently, disulfide bonds can rearrange dynamically and are not necessarily static and stable.

The free energy profile of a thiol-disulfide exchange heavily depends on the environment. If the reaction takes place in a polar environment (water or protein), the negative charge is localized on the thiolate and a nucleophilic attack on the disulfide bond proceeds over a transition state in which the sulfur atoms are aligned nearly linear. The stability of the reactant and products as well as the height of the energy barriers are determined by the local environment. In gas phase, the free energy profile is inverted because the charge is delocalized between the the sulfurs. Thus, the linear "trisulfide" complex is no longer the transition state but a minimum.[43–45]

In **chapter 10** we investigate the intramolceular thiol-disulfide exchange in a small protein under mechanical stress. We perform 334 QM/MM force-clamp simulations of the protein with a total simulation time of ~5.7 µs and analyze what the prerequesites for a successful thiol-disulfide exchange are. We also perform QM/MM metadynamics of the two possible attacks to obtain the potentials of the mean force. In addition, we investigate the impact of electrostatics on the barrier heights of thiol-disulfide exchange with QM/MM metadynamics simulations of a small model system.

In **chapter 11** the accuracy of the density-functional tight binding method (DFTB) regarding thiol-disulfide exchange reactions is evaluated. In DFTB, the S–S bonds in the transition states are too long and the transition state is a local minimum on the free energy surface instead of a saddle point We correct these errors by reparameterizing the S–S potentials and also by applying a machine learned energy correction. Both approaches are tested with QM/MM metadynamics simulations of a small model system and are then applied to intramolecular thiol-disulfide exchange in a blood protein.

In **chapter 11** the reparamterized parameters are used to perform a series of QM/MM metadynamics simulations of intramolecular thiol-disulfide exchange in a protein to validate experimental results.

Part II.

Theoretical Background

2. Computational Chemistry

Computational chemistry is a subfield of theoretical chemistry that uses computer calculations or simulations to solve chemical problems. For example, in materials science, computational chemistry is used in the search for more efficient catalysts for chemical reactions or in the development of new materials. In the pharmaceutical industry, computer calculations are used to find potential drugs by predicting the preferred orientations and binding energies of drug candidates in the active site of an enzyme. In biophysics and structural biology, the motions of biomolecules such as proteins and nucleic acids are simulated to explain atomic-level phenomena that cannot be directly observed in experiments. In general, such applications require the calculation of molecular geometries and energies and the calculation of interactions between molecules. The most common used methods can be divided into four broad classes with different accuracy and computational efficiency:

(I) With ***ab initio*** ("from first principles") quantum chemistry, the system is described by a wave function. Properties of the system, e.g., the energies, are obtained by solving the Schrödinger equation. Popular correlated *ab initio* methods like CCSD(T) and MP4 yield highly accurate results when compared to experiments, however they scale up to N^7 with system size. Thus, they are restricted to small system sizes and short simulation times.

(II) **Density Functional Theory (DFT)** is also based on the Schrödinger equation but uses functionals of the electron density instead of a wave function. This reduces the computational cost; DFT usually scales with N^3 which allows the calculation of larger systems over a longer time scale.

(III) **Semiempirical quantum chemistry methods (SE)** use parameters that are fitted to experimental or sophisticated QM methods. This significantly speeds up calculations, for example, the semiempirical Density-Functional Tight Binding (DFTB) method is 2-3 orders of magnitude faster than DFT, although less accurate.

(IV) In **Molecular Mechanics (MM)**, molecules are modeled as balls (atoms) connected by springs (bonds) and the interactions between the atoms are described by a so-called force field. The computational effort of force field methods is low and thus very large systems can be simulated up to several µs. However, this comes at the expense of accuracy. It is not possible to simulate chemical reactions in which bonds are broken or formed. The system size and time scale can be even more increased with coarse-grained methods where molecules are represented by pseudo-atoms which are formed by a group of atoms, for example, four heavy atoms and their associated hydrogens (four-to-one mapping).

The feasible system sizes and simulation time scales of the described methods are summarized in Fig. 2.1. A more detailed description of *ab initio* methods, DFT and the semiempirical DFTB method is presented in chapter 3. The theory of all-atom force field methods and molecular dynamics is described in chapter 4. When a process or mechanism of interest exceeds the time scale of the available methods, *enhanced sampling* techniques can be employed to accelerate rare events. Popular enhanced sampling methods are described in chapter 5. Another way to overcome the time scale problem is the application of *machine learning* (ML) algorithms. ML approaches are widely used in computational chemistry nowadays and a short introduction is given in chapter 6. If not stated otherwise, all information presented are found in textbooks such as Ref. [46–49]

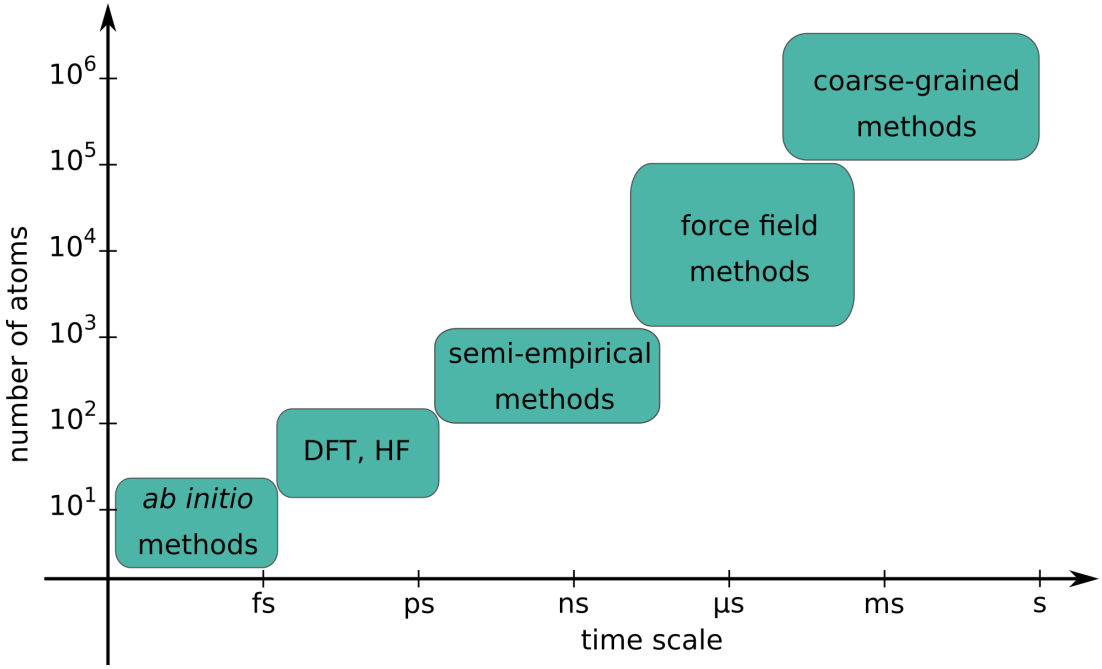

Figure 2.1.: Multiscale hierarchy of computational chemistry methods.

3. Quantum Chemistry

Schrödinger Equation
Phenomena like blackbody radiation and the photoelectric effect showed that the absorption and emission of light is quantized. Classical physics failed to describe such effects which lead to the development of quantum mechanics (QM). In QM, particles, such as atoms and electrons, are described by a many-body wave function $\Psi(r,t)$ that depends on the coordinates r of the particles. Their evolution in time is described by the time-dependent Schrödinger equation:

$$\begin{aligned} \hat{H}(r,t)\Psi(r,t) &= i\hbar\frac{\partial\Psi(r,t)}{\partial t} \\ \hat{H}(r,t) &= \hat{T}(r) + \hat{V}(r,t) \end{aligned} \tag{3.1}$$

where $\hat{H}(r,t)$ is the Hamiltonian operator, $\hat{T}(r)$ the kinetic energy operator, $\hat{V}(r,t)$ the potential energy operator, and $\hbar$ the reduced Planck constant. Solving the Schrödinger equation yields the wave function that describes the system as a function of time. The square of the absolute value of the wave function gives the probability of observing the particles at position r and time t:

$$P(r,t) = |\Psi(r,t)|^2 \tag{3.2}$$

When the potential energy operator is time-independent ($\hat{V}(r,t) = \hat{V}(r)$), the Hamilton operator becomes time-independent and the total energy of the system is obtained as:

$$\hat{H}(r)\Psi(r,t) = E(r)\Psi(r,t) \tag{3.3}$$

which has the solution:

$$\Psi(r,t) = \Psi(r)e^{(-iEt)} \tag{3.4}$$

For time-independent problems the exponential function can be separated from the spatial wave function and the time-independent Schrödinger equation is given as:

$$\hat{H}(r)\Psi(r) = E(r)\Psi(r) \tag{3.5}$$

Born-Oppenheimer Approximation
For a system that consist of M atoms and N electrons, the Hamilton operator that represents the total energy of the system consists of five contributions:

$$\hat{H} = \underbrace{-\frac{1}{2}\sum_{A}^{M}\frac{1}{M_A}\nabla_A^2}_{\hat{T}_N} \underbrace{-\frac{1}{2}\sum_{i}^{N}\nabla_i^2}_{\hat{T}_e} \underbrace{-\sum_{A}^{M}\sum_{i}^{N}\frac{Z_A}{r_{Ai}}}_{\hat{V}_{Ne}} + \underbrace{\sum_{i<j}^{N}\frac{1}{r_{ij}}}_{\hat{V}_{ee}} + \underbrace{\sum_{A<B}^{M}\frac{Z_A Z_B}{r_{AB}}}_{\hat{V}_{NN}} \tag{3.6}$$

where $\hat{T}_N$ is the kinetic energy of the nuclei, $\hat{T}_e$ the kinetic energy of the electrons, $\hat{V}_{Ne}$ the attractive nucleus-electron interactions, $\hat{V}_{ee}$ the repulsive electron-electron interactions and $\hat{V}_{NN}$ repulsive nucleus-nucleus interactions. **R** and **r** are the coordinates of the nuclei and electrons, respectively.

The Schrödinger equation can be further simplified by considering the difference in masses between the nuclei and the electrons. A proton, which is the lightest of all nuclei, weighs 1838 times more than an electron and thus moves much slower than electrons. Electrons respond almost instantaneously to changes of nuclear positions and therefore can be considered as moving in a field of fixed nuclei, known as the Born-Oppenheimer approximation. As a consequence of the fixed nuclei positions, the kinetic energy of the nuclei is zero and the nucleus-nucleus repulsion is constant for given geometries. The Hamilton operator reduces to the so-called electronic Hamiltonian and the electric Schrödinger equation is given as:

$$\hat{H}_{elec}\Psi_{elec} = \left(\hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee}\right)\Psi_{elec} = E_{elec}\Psi_{elec} \tag{3.7}$$

that only depends on the coordinates **r** of the electrons. The total energy of the N-particle system is obtained as:

$$E_{tot} = E_{elec} + E_{nuc} = E_{elec} + \sum_{A<B}^{N} \frac{Z_A Z_B}{r_{AB}} \tag{3.8}$$

In the following, only the electronic Schrödinger equation is considered and therefore the subscript “elec” is omitted.

Variational Principle
Due to the many-body interactions of the electrons, the Schrödinger equation can only be solved for one-electron systems, for example $H_2{}^+$. For many-electron systems the variational principle is used to find the best possible approximation of the many-body wave function Ψ. The variational principle states that any trial wave function Ψ_{trial} yields an energy E_{trial} that is an upper bound to the exact ground state energy E_0:

$$\langle\Psi_{trial}|\hat{H}|\Psi_{trial}\rangle = E_{trial} \geq E_0 = \langle\Psi_0|\hat{H}|\Psi_0\rangle \tag{3.9}$$

3.1. Hartree-Fock

Hartree Product and Slater Determinant.
Still, finding a solution of the Schrödinger equation is very complex and can be simplified by introducing the independent-particle model. In this approximation, the motion of each electron is considered to be independent of the dynamics of all other electrons and their interactions are included as an averaged effect. Thus, the N-electron wave function is constructed as the product of N one-electron wave functions, known as a Hartree Product:

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = \phi_1(1)\phi_2(2)\ldots\phi_N(N) \tag{3.10}$$

However, the Hartree product violates the antisymmetry principle for fermions (spin = 1/2), which states that the wave function Ψ must change the sign when the coordinates of any two electrons are interchanged. The antisymmetry is achieved by using linear combinations of the Hartree product, given as a Slater determinant:

$$\Phi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \dots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \dots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \dots & \phi_N(N) \end{vmatrix} \tag{3.11}$$

where $\frac{1}{\sqrt{N!}}$ is the normalization factor. Coordinates of the electrons are given along the rows, the columns are the ortho-normalized molecular orbitals (MOs) ϕ_i:

$$\left\langle \phi_i \middle| \phi_j \right\rangle = \delta_{ij} \tag{3.12}$$

which are a product of a spatial orbital and a spin function with an intrinsic spin coordinate α or β.

Hamiltonian.
The energy of the Slater determinant is obtained by splitting the molecular Hamiltonian $\hat{H}$ into individual one-electron operators $\hat{h}_i$ and two-electron operator $\hat{g}_{ij}$:

$$\begin{aligned} \hat{H} &= \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} + \hat{V}_{NN} \\ &= \sum_i^N \hat{h}_i + \sum_{i<j}^N \hat{g}_{ij} + \sum_{A<B}^M \frac{Z_A Z_B}{r_{AB}} \\ \hat{h}_i &= -\frac{1}{2}\nabla_i^2 - \sum_A^M \frac{Z_A}{r_{Ai}} \\ \hat{g}_{ij} &= \frac{1}{r_{ij}} \end{aligned} \tag{3.13}$$

where $\hat{h}_i$ describes the motion of electron i in the field of all nuclei, and $\hat{g}_{ij}$ the electron-electron repulsion.

The energy can then be written as:

$$E = \sum_i^N h_i + \sum_i^N \sum_{i<j}^N \left(J_{ij} - K_{ij}\right) + V_{NN} \tag{3.14}$$

The Coulomb integral J represents the electron-electron repulsion between two charge distributions. The exchange integral K has no classical analogy and originates from the antisymmetry of the Slater determinant.

In the next step, the slater determinant that yields the lowest energy of the system has to be found with the variational principle under the constraint that the MOs remain ortho-normalized. This is achieved by Lagrange multipliers ϵ_i which leads to Hartree-Fock equations:

$$\hat{f}\phi_i = \epsilon_i \phi_i \qquad i = 1, 2, \dots, N \tag{3.15}$$

where ϵ_i are the eigenvalues of the Fock operator $\hat{f}$. The eigenvalues can be interpreted as orbital energies. The Fock operator is an effective one-electron energy operator:

$$\hat{f}_i = \underbrace{-\frac{1}{2}\nabla_i^2 - \sum_A^M \frac{Z_A}{r_{Ai}}}_{\hat{h}_i} + \underbrace{\sum_j^N \left(\hat{J}_j - \hat{K}_j\right)}_{V_{HF}(i)} \tag{3.16}$$

where the first term is the kinetic energy and the second term the attractive electron-nucleus potential energy. The Hartree-Fock potential $V_{HF}(i)$ is the average repulsive potential of the electron experienced by all other electrons. The Fock operator depends on all occupied MOs via $\hat{J}$ and $\hat{K}$, therefore a specific MO can only be determined if all other occupied MOs are already known. Hence, the Hartree-Fock equations have to be solved iterative in a Self-consistent Field (SCF) procedure where the MOs are constructed from an initial guess and adjusted in each iteration until a convergence criterion of the total energy is met.

Basis Set Approximation.
The MOs are usually represented as a linear combination of atomic orbitals (LCAO):

$$\phi_i = \sum_\alpha^{M_{basis}} c_{\alpha i}\chi_\alpha \tag{3.17}$$

where $c_{\alpha i}$ are the expansion coefficients and χ_α the basis functions, for example Slater type or Gaussian type orbitals. With this approach, the Hartree-Fock equations can be written in a matrix notation leading to the Rotahn-Hall equations:

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{3.18}$$

where **F** is the Fock matrix that contains the Fock matrix elements, **S** the matrix that contains the overlap elements between the basis functions, **C** the $N \times N$ matrix that contains the expansion coefficients and **e** the $N \times N$ that contains the orbital energies. In order to determine the unknown MO coefficients, the Fock matrix has to be diagonalized. Since the Fock matrix requires that the coefficients are known, the Roothaan-Hall equation is solved iterative starting from an initial guess of the coefficients until the obtained set of coefficients is self-consistent within a certain threshold.

Electron Correlation
The accuracy of the HF method improves when the basis set size is increased. However, even with very large basis sets the HF method can only account for ≈99% due to the mean field approximation where every electron is treated independently, i.e., moving under the influence of an averaged electrostatic field induced by all other electrons. This leads to the neglect of electron correlation which often is very important for the description of chemical phenomena accounting for the remaining ≈1% of the total energy. One way to improve the HF results is the construction of a wave function that consists of more than one Slater determinant for which various so-called post-Hartree Fock methods have

been developed. Common methods are Configuration Interaction (CI), Møller–Plesset perturbation theory (MP) or Coupled Cluster (CC).

Configuration Interaction
In the CI ansatz, the trial wave function is constructed as a so-called configuration state function (CSF). The CSF is a linear combination of Slater determinants which only differ in their excitation states:

$$\Psi_{\mathrm{CI}} = c_0 \Phi_0^{\mathrm{HF}} + \underbrace{\sum_{i}^{\mathrm{occ}} \sum_{a}^{\mathrm{vir}} c_i^a \Phi_i^a}_{\mathrm{S}} + \underbrace{\sum_{i>j}^{\mathrm{occ}} \sum_{a>b}^{\mathrm{vir}} c_{ij}^{ab} \Phi_{ij}^{ab}}_{\mathrm{D}} + \cdots \tag{3.19}$$

They are constructed from a single reference Slater determinant where one or more electrons are promoted from an occupied MO to an unoccupied (virtual) MO. If only single excitation Slater determinants (S) are considered, the method is referred to as CIS. If the double excitation Slater determinants (D) are also included, the method is referred to as CISD. Since the CI ansatz scales poorly with the number of electrons ($O(N!)$), higher excitation Slater determinants are usually not considered. The expansion coefficients are obtained by minimizing the energy of the system with the variational principle.

The correlation can be improved with the Multi-Reference Configration Interaction (MRCI) method where more than one reference determinant is considered. The reference determinants are taken from a multiconfigurational SCF (MCSCF) calculation where the MOs within the determinants are optimized via the coefficients in the LCAO expansion (Eq. 3.17). Usually, only a subset of the MCSCF expansion space is used for MRCI calculations due to computational limitations.

Coupled Cluster
The CC theory uses an exponential ansatz for the construction of the wave function:

$$\Psi_{CC} = \left(1 + \hat{T} + \frac{1}{2!}\hat{T}^2 + \frac{1}{3!}\hat{T}^3 + \dots\right)\Phi_{HF} = e^{\hat{T}}\Phi_{HF} \tag{3.20}$$

where $\hat{T} = \hat{T}_1 + \hat{T}_2 + \dots$ are the excitation operators. The singly excitation operator $\hat{T}_1$ puts one electron into a virtual MO, the doubly excitation operator $\hat{T}_2$ puts two electrons into virtual MOs etc. Depending on how many excitation operators are included one obtains different methods:

$$\begin{aligned} \Psi_{\mathrm{CCD}} &= e^{\hat{T}_2}\Phi_{HF} \\ \Psi_{\mathrm{CCSD}} &= e^{(\hat{T}_1+\hat{T}_2)}\Phi_{HF} \\ \Psi_{\mathrm{CCSDT}} &= e^{(\hat{T}_1+\hat{T}_2+\hat{T}_3)}\Phi_{HF} \end{aligned} \tag{3.21}$$

where CCD stands for *coupled cluster doubles*, CCSD for *coupled cluster singles and doubles*, and CCSDT for *coupled cluster singles, doubles and triples*. The CCSDT method (N^8 scaling) is already very demanding and thus the CCSD(T) method (N^7 scaling) may be used instead where only approximate triples are used. More complicated coupled cluster methods,

for example CCSDTQ, which scales with N^{10}, are usually only used for high-accuracy calculations of small molecules. In comparison, HF formally scales with N^4 which can be reduced to N^1 in practice.

3.2. Semiempirical Wave Function Based Methods.

Semiempirical wave function methods are based on the Schrödinger equation and solved by diagonalizing a Fock matrix. Several approximations are employed to reduce the computational cost compared to *ab initio* calculations. First, only valence atoms are considered and a minimal basis set is used. Second, not all integrals are calculated, but rather taken as zero depending on the used method. This means, the *differential overlap* of two-, three-, and four-center integrals are neglected to some extend. In the *zero differential overlap* (ZDO) only two-electron integrals of the type $(\mu\mu \mid \nu\nu)$ are calculated and all three- and four-center two electron integrals are set to zero:

$$(\mu\lambda \mid \nu\sigma) = \delta_{\mu\lambda}\delta_{\nu\sigma}(\mu\mu \mid \nu\nu) \tag{3.22}$$

In the *complete neglect of differential overlap* (CNDO) the ZDO approximation is used and the one-center integrals on the same atom A and all two-center integrals between atoms A and B are parameterized. The *intermediate neglect of differential overlap* (INDO) goes beyond CNDO by also parameterizing two-electron integrals on the same atom. The *neglect of differential-diatomic overlap* (NDDO) further includes two-electron integrals $(\mu\nu \mid \lambda\sigma)$ where μ and ν are located on atom A and λ and σ are located on atom B, which greatly improves accuracy compared to the previously described methods.

Third, the overlap matrix is taken as a unit matrix ($\mathbf{S} = \mathbf{1}$) and the Roothaan-Hall equation transforms to

$$\mathbf{FC} = \mathbf{C}\epsilon \tag{3.23}$$

which is a standard eigenvalue problem.

Orthogonalization model 2
In NDDO a time consuming orthogonalization procedure of the Fock matrix is avoided by assuming a priori the Fock matrix **F** to be orthogonal. However, this assumption may lead to severe errors. Hence, orthogonalization corrected methods such as the orthogonalization model 2 (OM2) might be used instead. Combined with MRCI, OM2/MRCI can be used to obtain accurate excitation energies at low computational cost.

3.3. Density-Functional Theory

Electron Density
From the probability interpretation of the wave function, the electron density $\rho(\mathbf{r})$ can be obtained. It is defined as a multiple integral over the spin coordinates of all electrons and the spatial variables of all electrons but one:

$$\rho(\mathbf{r}) = N \int \cdots \int |\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)|^2 ds_1 d\mathbf{r}_2 \ldots d\mathbf{r}_N \tag{3.24}$$

The multiple integral represents the probability of finding a particular electron with arbitrary spin within the volume element $\mathbf{r}_1$ while all other $(N-1)$ electrons have arbitrary spins and positions. However, electrons are indistinguishable and therefore the probability of finding any of the N electrons within the volume element $\mathbf{r}_1$ is N times the probability for one particular electron. It follows, that the electron density integrates to the total number of electrons:

$$N = \int \rho(\mathbf{r}) d\mathbf{r}_1 \tag{3.25}$$

Due to the attractive forces between the electrons and the nuclei, $\rho(\mathbf{r})$ exhibits maxima (that are cusps) at the positions of the atoms.

Pair Density and Correlation Effects

The probability of finding two electrons within two volume elements $\mathbf{r}_1$ and $\mathbf{r}_2$ is the pair density $\rho_2(\mathbf{r}_1, \mathbf{r}_2)$. If the electrons were classical, non-interacting particles, the pair density would reduce to the product of the individual probabilities:

$$\rho_2(\mathbf{r}_1, \mathbf{r}_2) = \frac{N-1}{N} \rho(\mathbf{r}_1)\rho(\mathbf{r}_2) \tag{3.26}$$

with $\frac{N-1}{N}$ as prefactor since electrons are identical and not distinguishable. However, two effects have to be considered: (i) the exchange (or Fermi) correlation and (ii) the Coulomb correlation. The exchange correlation prevents that two electrons with the same spin are found at the same point in space. Due to the antisymmetry of a Slater determinant, this effect is already included in the Hartree-Fock approach. The Coulomb correlation denotes the electrostatic repulsion between the electrons which prevents that electrons move to close to each other. Thus, it is convenient to separate the pair density into a product of independent densities and a product of correlated densities:

$$\rho_2(\mathbf{r}_1, \mathbf{r}_2) = \rho(\mathbf{r}_1)\rho(\mathbf{r}_2)[1 + f(\mathbf{r}_1; \mathbf{r}_2)] \tag{3.27}$$

with $[1 + f(\mathbf{r}_1; \mathbf{r}_2)]$ as the correlation factor that accounts for the Fermi and Coulomb effects. It follows that the correlated probability of finding an electron at $\mathbf{r}_2$ is:

$$\rho(\mathbf{r}_2) f(\mathbf{r}_1; \mathbf{r}_2) = \frac{\rho(\mathbf{r}_1, \mathbf{r}_2)}{\rho(\mathbf{r}_1)} - \rho(\mathbf{r}_2) = h_{XC}(\mathbf{r}_1; \mathbf{r}_2) \tag{3.28}$$

where the first term $\rho(\mathbf{r}_1, \mathbf{r}_2)/\rho(\mathbf{r}_1)$ is the conditional probability, i.e., the probability of finding any electron at $\mathbf{r}_2$ if there is already an electron at $\mathbf{r}_1$ in the coordinate-spin space, and the second term $\rho(\mathbf{r}_2)$ the uncorrelated probability of finding an electron at $\mathbf{r}_2$. Their difference $h_{XC}(\mathbf{r}_1; \mathbf{r}_2)$ is called the exchange-correlation hole. In other words: each electron sits in a hole where the probability of finding another electron is close to zero due to electron correlation.

Hohenberg-Kohn Theorems

The advantage of using the electron density is that it only depends on three spatial variables whereas the wave function depends on 4N variables (one spin variable and three spatial variables for each electron). Already in 1927, Thomas and Fermi formulated an

expression for the energy of an atom, which depends solely on the electron density. The nuclear-electron and electron-electron contributions were treated classically and the simple expression for the kinetic energy was based on a fictional model system of constant electron density (uniform electron gas). They *assumed* that the ground state of the system depends on the electron density, however proof was not provided until about 40 years later by Hohenberg and Kohn in 1964.

The two Hohenberg-Kohn theorems state:

1. The external potential $V_{ext}(\mathbf{r})$ is a unique functional of $\rho(\mathbf{r})$, i.e., there cannot be two different $V_{ext}(\mathbf{r})$ that yield the same ground state electron density $\rho_0(\mathbf{r})$. Since $V_{ext}(\mathbf{r})$ fixes the Hamiltonian $\hat{H} = \hat{T} + \hat{V}_{ee} + \hat{V}_{ext}$, the ground state energy (and all other properties) are a functional of the ground state electron density $\rho_0(\mathbf{r})$:

$$E_0[\rho_0(\mathbf{r})] = T[\rho_0(\mathbf{r})] + E_{ee}[\rho_0(\mathbf{r})] + E_{Ne}[\rho_0(\mathbf{r})]$$

2. The energy that is obtained from any trial density $\rho(\mathbf{r})$ is an upper bound to the true ground state energy:

$$E[\rho(\mathbf{r})] \geq E[\rho_0(\mathbf{r})]$$

   Thus, the variational principle can be used to approximate the ground state electron density $\rho_0(\mathbf{r})$.

Kohn-Sham Approach

The first density functionals, such as the Thomas-Fermi model, were not very successful due to their low accuracy. Most of the problems resulted from the poor representation of the kinetic energy. Hence, Kohn and Sham (KS) suggested to split the kinetic energy functional into two parts: (i) the kinetic energy T_S of a reference system of non-interacting electrons with the same electron density as the real system, for which orbitals have to be re-introduced; (ii) the exchange-correlation energy E_{XC} which is the remainder of the exact kinetic energy that has to be treated approximately. The general DFT energy expression can be written as:

$$\begin{aligned} E_{DFT}[\rho(\mathbf{r})] &= T_S[\rho(\mathbf{r})] + J[\rho(\mathbf{r})] + E_{XC}[\rho(\mathbf{r})] + E_{Ne}[\rho(\mathbf{r})] \\ &= -\frac{1}{2}\sum_i^N \langle\phi_i|\nabla^2|\phi_i\rangle + \frac{1}{2}\sum_i^N\sum_j^N \iint |\phi_i(\mathbf{r}_1)|^2 \frac{1}{r_{12}} \left|\phi_j(\mathbf{r}_2)\right|^2 d\mathbf{r}_1 d\mathbf{r}_2 \\ &+ E_{XC}[\rho(\mathbf{r})] - \sum_i^N \int \sum_A^M \frac{Z_A}{r1A} |\phi_i(\mathbf{r}_1)|^2 d\mathbf{r}_1 \end{aligned} \tag{3.29}$$

where $T_S[\rho(\mathbf{r})]$ is the kinetic energy of non-interacting electrons, $J[\rho(\mathbf{r})]$ the classical Coulomb interaction, $E_{XC}[\rho(\mathbf{r})]$ the exchange-correlation energy, and $E_{Ne}[\rho(\mathbf{r})]$ the attractive nuclei-electron energy. $E_{XC}[\rho(\mathbf{r})]$ is the only term without an explicit form and contains everything that is unknown, i.e., all non-classical corrections including a self-interaction correction.

The electron density is obtained from the Kohn-Sham orbitals:

$$\rho(\mathbf{r}) = \sum_{i}^{N} |\phi_i|^2 \tag{3.30}$$

In order to find the electron density that minimizes the energy expression, the variational principle is applied under the orthonormality condition. This yields a set of one-electron equations:

$$\left[\frac{1}{2}\nabla^2 + \left(\int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 + V_{XC}(\mathbf{r}_1) - \sum_{A}^{N} \frac{Z_A}{r_{1A}}\right)\right]\phi_i = \left[\frac{1}{2}\nabla^2 + V_{eff}(\mathbf{r}_1)\right]\phi_i = \epsilon\phi_i \tag{3.31}$$

known as the Kohn-Sham equations. Analogously to the HF equations, the Kohn-Sham equations have to be solved iteratively with the LCAO approach, since $V_{eff}(\mathbf{r}_1)$ already depends on the density, and therefore on the orbitals, through the Coulomb term. The term V_{XC} is defined as the functional derivative of E_{XC} with respect to the electron density:

$$V_{XC}(\mathbf{r}_1) = \frac{\delta E_{XC}\rho(\mathbf{r})}{\delta\rho(\mathbf{r})} \tag{3.32}$$

If both exchange-correlation terms were known, the Kohn-Sham approach would yield the exact energy. However, since they are unknown, they have to be approximated which is the key challenge in DFT. Many functionals have been proposed, such as the local density approximation, the generalized gradient approximation and hybrid functionals.

Local Density Approximation

In the Local Density Approximation (LDA), the density is treated as a uniform electron gas. The E_{XC} term can be written in a simple form:

$$\begin{aligned} E_{XC}^{LDA}(\rho) &= \int \rho(\mathbf{r})\epsilon_{XC}[\rho(\mathbf{r})]d\mathbf{r} \\ \epsilon_{XC}[\rho(\mathbf{r})] &= \epsilon_X[\rho(\mathbf{r})] + \epsilon_C[\rho(\mathbf{r})] \end{aligned} \tag{3.33}$$

where ϵ_{XC} is the exchange-correlation energy per particle that can be split into exchange and correlation contributions. The exchange part ϵ_X of an electron in a uniform (homogeneous) electron gas can be determined analytically and yields:

$$\epsilon_X = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \rho(\mathbf{r})^{1/3} \tag{3.34}$$

The correlation energy ϵ_C can be determined by quantum Monte Carlo methods. LDA often performs well for solid-state systems, however fails to describe molecules where the density undergoes changes.

Generalized Gradient Approximation

In order to account for density fluctuations, the Generalized Gradient Approximation (GGA) includes the first derivative of the density $\nabla\rho(\mathbf{r})$ as a variable:

$$E_{XC}^{GGA}(\rho) = \int \rho(\mathbf{r})\epsilon_{XC}[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})]d\mathbf{r} \tag{3.35}$$

Most GGA functionals add correction terms on top of the LDA functional, such as the B88 functional by Becke or the LYP functional by Lee, Yang and Parr. For this, parameters have to be determined by fitting to reference data. Another possibility is to derive the parameters from certain conditions, which has be done for the popular Perdew-Burke-Ernzerhof (PBE) functional.

Hybrid Functionals
The GGA method can be further improved by including HF exchange. In the famous B3LYP hybrid functional, the exchange-correlation energy is given as a combination of density-functional exchange and correlation and HF exchange:

$$E_{XC}^{B3LYP} = (1-a) + aE_X^{HF} + bE_X^{B88} + (1-c)E_C^{LSDA} + cE_C^{LYP} \tag{3.36}$$

with $a = 0.20$, $b = 0.72$ and $c = 0.81$, which are determined by fitting to experimental data; LSDA stands for Linear Spin Density Approximation.

3.4. Density-Functional Tight-Binding

Derivation from DFT
DFT is significantly faster than ab initio quantum mechanical methods, however, still limited to small system sizes (~100 atoms). Larger system sizes are accessible with semiempirical (SE) methods such as the Density-functional Tight-Binding (DFTB) method that is is based on DFT and introduces certain approximations and element-specific parameters.

Starting point of DFTB is the DFT total energy which is expanded in a Taylor series around a chosen density $\rho(\mathbf{r})$. The density is approximated by a reference density $\rho^0(\mathbf{r})$, calculated from a superposition of neutral atomic densities, that is perturbed by a density fluctuation $\delta\rho(\mathbf{r})$:

$$\rho(\mathbf{r}) = \rho^0(\mathbf{r}) + \delta\rho(\mathbf{r}) = \sum_A \rho_A^0(\mathbf{r}) + \delta\rho(\mathbf{r}) \tag{3.37}$$

Expanding the exchange-correlation energy functional up to the third order yields the DFTB total energy:

$$\begin{aligned}
E^{\mathrm{DFTB}} &= E_{xc}[\rho_0(\mathbf{r})] + E_{NN} - \int V_{xc}[\rho_0(\mathbf{r})]\rho_0(\mathbf{r})d\mathbf{r} - \frac{1}{2}\iint \frac{\rho_0(\mathbf{r})\rho_0(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' \\
&\quad + \sum_i^{occ}\langle\psi_i\underbrace{\left|-\frac{1}{2}\nabla^2 + V_{ext} + \int\frac{\rho_0(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + V_{xc}[\rho_0(\mathbf{r})]\right|}_{\hat{H}^0[\rho_0(\mathbf{r})]}\psi_i\rangle \\
&\quad + \frac{1}{2}\iint\left(\frac{\delta^2 E_{xc}[\rho_0(\mathbf{r})]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')} + \frac{1}{|\mathbf{r}-\mathbf{r}'|}\right)_{\rho_0(\mathbf{r})}\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')d\mathbf{r}d\mathbf{r}' \\
&\quad + \frac{1}{6}\iiint\left.\frac{\delta^3 E_{xc}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'')}\right|_{\rho_0(\mathbf{r})}\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'')d\mathbf{r}d\mathbf{r}'d\mathbf{r}'' \\
&= E^0[\rho_0(\mathbf{r})] + E^1[\rho_0(\mathbf{r}),\delta\rho(\mathbf{r})] + E^2\left[\rho_0(\mathbf{r}),(\delta\rho(\mathbf{r}))^2\right] + E^3\left[\rho_0(\mathbf{r}),(\delta\rho(\mathbf{r}))^3\right]
\end{aligned} \tag{3.38}$$

Depending on where the Taylor series is truncated, different DFTB methods are obtained. Inclusion of the zeroth- and first-order term yields the original DFTB method (DFTB1).[50,51] Further inclusion of the second-order term leads to self-consistent-charge DFTB (DFTB2)[52] and the inclusion of the third-order term to DFTB3.[53,54]

DFTB1

The four contributions of $E^0[\rho_0(\mathbf{r})]$ depend only on the reference density and therefore are independent from the chemical environment. Hence, the zeroth-order term can be considered as the core-core repulsion without any electronic contribution and can be approximated as a sum of two-body potentials. The so-called repulsive energy term:

$$E^0[\rho_0(\mathbf{r})] \approx E^{rep} = \frac{1}{2}\sum_{A,B} V_{AB}^{rep} \tag{3.39}$$

where A and B denote atoms. The repulsive potential only depends on atomic numbers and distances and is usually fitted as spline functions to reproduce DFT or empirical data.

In the DFTB formalism, only the valence electrons are considered explicitly and the Kohn-Sham orbitals are expanded in a minimal basis:

$$\Psi_i = \sum_{\mu} c_{\mu i}\phi_\mu \tag{3.40}$$

with only one radial function for each angular momentum state. The basis functions μ are obtained from DFT calculations with an additional confinement potential which makes the orbitals more compact. The first-order term is the band-structure energy and obtained as:

$$E^1[\rho_0(\mathbf{r}), \delta\rho(\mathbf{r})] = \sum_{i}^{\text{occ}} \left\langle \psi_i \middle| \hat{H}^0 \middle| \psi_i \right\rangle = \sum_{i}^{\text{occ}} \sum_{\mu\in A} \sum_{\nu\in B} c_\mu^i c_\nu^i H_{\mu\nu}^0 \tag{3.41}$$

The diagonal Hamiltonian matrix elements $H_{\mu\mu}^0$ are approximated as the orbital energies ϵ_μ of the free atoms which are obtained from DFT calculations using the PBE functional. For off-diagonal elements a two-center approximation is applied, i.e., three- and four-center integrals are neglected. All matrix elements are precomputed and tabulated for each pair of orbitals and interpolated for a given geometry during a DFTB calculation. DFTB1 performs well for systems with no charge transfer or a complete charge transfer. For systems that are sensitive to charge fluctuations higher order terms have to be included.

DFTB2

The density fluctuations in the second order term are expressed as a sum of spherically symmetric atomic contributions:

$$\delta\rho(\mathbf{r}) = \sum_{A} \delta\rho_A(\mathbf{r}) \tag{3.42}$$

They are represented by atomic Mulliken point charges δq_A which decay exponentially:

$$\delta\rho_A(\mathbf{r}) \approx \Delta q_A \frac{\tau_A^3}{8\pi} e^{-\tau_a|\mathbf{r}-\mathbf{R}_A|} \tag{3.43}$$

With these approximations the second-order term can be expressed as:

$$E^2\left[\rho_0(\mathbf{r}), (\delta\rho(\mathbf{r}))^2\right] = \frac{1}{2}\sum_{A,B}\gamma_{AB}\Delta q_A \Delta q_B \tag{3.44}$$

where γ_{AB} is an analytical function. For large distances between atoms A and B the γ_{AB}-function converges to a Coulombic interaction ($\frac{1}{R_{AB}}$), for short distances ($A = B$ and $R_{AB} \longrightarrow 0$) the second order term E^2 describes the on-site electron-electron interaction of atom A as $\gamma_{AA} = U_A$. The Hubbard parameter U_A is the second derivative of total energy with respect to the charge density and obtained from a DFT calculation of an isolated atom. Moreover, U_A is related to chemical hardness which defines how the energy of an atom changes when an electron is added or removed. The γ-function assumes that the width of the atomic charge density is proportional to the chemical hardness which works well for many elements except hydrogen, hence a modified γ^h-function for hydrogen was introduced.

DFTB3

In DFTB2, the chemical hardness of an atom is independent from its charge state which can lead to errors in highly charged systems. The third-order term introduces the Γ_{AB}-function which includes the Hubbard derivative U_A^d, i.e., the charge derivative of the Hubbard parameter. The final DFTB3 energy expression then reads:

$$E^{DFTB3} = \frac{1}{2}\sum_{A,B} V_{AB}^{rep} + \sum_{i}^{\text{occ}}\sum_{\mu\in A}\sum_{\nu\in B} c_\mu^i c_\nu^i H_{\mu\nu}^0 + \frac{1}{2}\sum_{A,B}\gamma_{AB}\Delta q_A\Delta q_B + \frac{1}{3}\sum_{A,B}(\Delta q_A)^2\Delta q_B\Gamma_{AB} \tag{3.45}$$

The minimum energy is obtained with the variation principle, leading to:

$$\sum_{\nu} c_\nu^i\left(H_{\mu\nu} - \epsilon_i S_{\mu\nu}\right) = 0 \tag{3.46}$$

The charge-dependent Hamiltonian $H_{\mu\nu}$ is given as:

$$H_{\mu\nu} = H_{\mu\nu}^0 + S_{\mu\nu}\cdot \underbrace{\sum_C \Delta q_C\left(\frac{\gamma_{AC}+\gamma_{BC}}{2} + \frac{\Delta q_A\Gamma_{AC} + \Delta q_B\Gamma_{BC}}{3} + \frac{\Delta q_C(\Gamma_{CA}+\Gamma_{CB})}{6}\right)}_{\Omega_{AB}} \tag{3.47}$$

where $S_{\mu\nu}$ is the overlap matrix and Ω_{AB} the Hamiltonian shift due to the induced charges.

Charges

The differential atomic charges Δq_a are obtained by subtracting the number of electrons of a neutral atom q_A^0 of the atomic charges q_A:

$$\delta q_A = q_A - q_A^0 \tag{3.48}$$

which are obtained from the Mulliken analysis:

$$q_A = \sum_{i}^{\text{occ}}\sum_{\mu\in A}\sum_{B}\sum_{\nu\in B} c_\mu^i c_\nu^i S_{\mu\nu} \tag{3.49}$$

Since the Hamiltonian depends on the Mulliken charges which in turn depend on the molecular orbital coefficients c_μ^i, Eq. 3.46 and 3.47 have to be solved iteratively until self-consistency is reached.

Coupled-Perturbed DFTB

A recent extension to DFTB are coupled-perturbed (CP) equations which calculate the derivative of atomic charges with respect to atomic coordinates. They were first derived and implemented for DFTB2 by Witek *et al*[55] and later extended to DFTB3 by Benjamin Hourahine.[56] The CP-DFTB equations must be solved iteratively and self-consistent because they include three sets of quantities that depend on each other.

First, the derivative of the MO coefficients with respect to the atomic coordinates a are expressed a matrices U^a:

$$\frac{\partial c_\mu^i}{\partial a} = \sum_m^{\mathrm{MO}} U_{mi}^{(a)} c_{\mu m} \tag{3.50}$$

where the diagonal elements are obtained as

$$U_{ii}^{(a)} = -\frac{1}{2} \sum_{\mu v}^{\mathrm{AO}} c_{\mu i} c_{Vi} \frac{\partial S_{\mu\nu}}{\partial a} \tag{3.51}$$

and the off-diagonal elements as:

$$U_{ij}^{(a)} = \frac{1}{\varepsilon_j - \varepsilon_i} \sum_{M,N}^{\mathrm{atoms}} \sum_{\mu\in M} \sum_{v\in N} c_{\mu i} c_{vj} \cdot \left(\frac{\partial H_{\mu\nu}^0}{\partial a} + \frac{\partial S_{\mu\nu}}{\partial a} \left(\Omega_{MN} - \varepsilon_j \right) + S_{\mu\nu} \frac{\partial \Omega_{MN}}{\partial a} \right) \tag{3.52}$$

Second, the derivative of atomic charges with respect to atomic coordinates is calculated:

$$\frac{\partial \Delta q_A}{\partial a} = \sum_i^{MO} n_i \sum_{\mu\in A} \sum_v^{\mathrm{AO}} \left(c_{\mu i} c_{vi} \frac{\partial S_{\mu\nu}}{\partial a} + \sum_m^{\mathrm{MO}} \left(U_{mi}^{(a)} \left(c_{\mu m} c_{vi} + c_{\mu i} c_{vm} \right) S_{\mu\nu} \right) \right) \tag{3.53}$$

Third, the derivative of the atomic shift with respect to atomic coordinates:

$$\begin{aligned} \frac{\partial \Omega_{AB}}{\partial a} = & \frac{1}{2} \sum_C^{\mathrm{QM\ atoms}} \left(\left(\frac{\partial \gamma_{AC}}{\partial a} + \frac{\partial \gamma_{BC}}{\partial a} \right) \Delta q_C + \left(\gamma_{AC} + \gamma_{BC} \right) \frac{\partial \Delta q_C}{\partial a} \right) \\ & + \frac{1}{3} \sum_C^{\mathrm{QM\ atoms}} \left(\Delta q_A \Gamma_{AC} + \Delta q_B \Gamma_{BC} + \Delta q_C \frac{\Gamma_{CA} + \Gamma_{CB}}{2} \right) \frac{\partial \Delta q_C}{\partial a} \\ & + \frac{1}{3} \sum_C^{\mathrm{QM\ atoms}} \left(\frac{\partial \Delta q_A}{\partial a} \Gamma_{AC} + \frac{\partial \Delta q_B}{\partial a} \Gamma_{BC} + \Delta q_A \frac{\partial \Gamma_{AC}}{\partial a} + \Delta q_B \frac{\partial \Gamma_{BC}}{\partial a} \right. \\ & \left. + \frac{\partial \Delta q_C}{\partial a} \left(\Gamma_{CA} + \Gamma_{CB} \right) + \frac{1}{2} \Delta q_C \left(\frac{\partial \Gamma_{CA}}{\partial a} + \frac{\partial \Gamma_{CB}}{\partial a} \right) \right) \Delta q_C \end{aligned} \tag{3.54}$$

4. Molecular Mechanics and Molecular Dynamics

With semi-empirical methods such as DFTB, systems with several hundreds of atoms can be calculated efficiently. To describe larger systems, further approximations are required. In Molecular Mechanics (MM), quantum mechanical effects are neglected and the molecules are treated as classical particles with the so called “ball and spring” model. Each nucleus and its surrounding electrons is treated as one spherical particle with a net charge, that is connected with other particles via “springs”. The potential energy is obtained as a parametric function of the nuclear coordinates, referred to as force field, compare Fig. 4.1.

4.1. Force Field Energy

The general force field equation consists of different energy terms that describe bonded and non-bonded interactions:

$$E_{\text{total}} = \underbrace{E_{\text{str}} + E_{\text{bend}} + E_{\text{tors}}}_{\text{bonded}} + \underbrace{E_{\text{LJ}} + E_{\text{el}}}_{\text{non-bonded}} \tag{4.1}$$

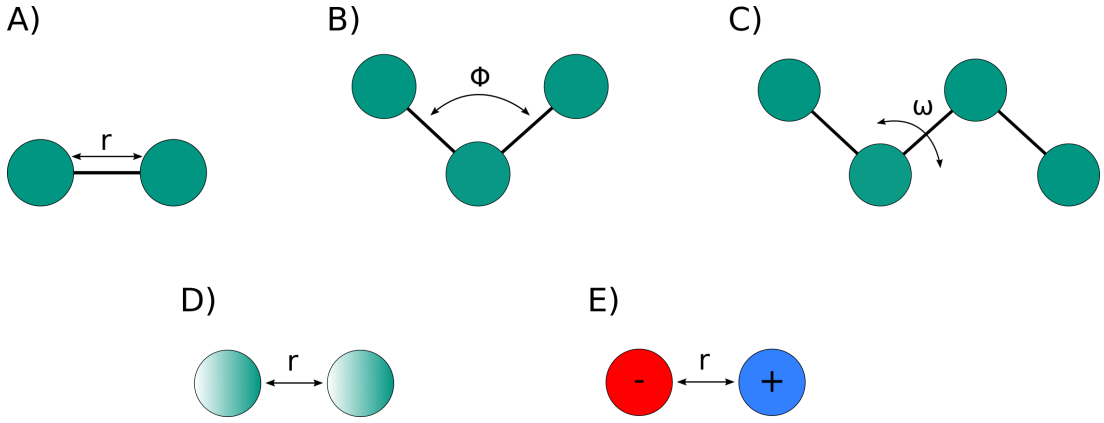

Figure 4.1.: Illustration of force field contributions. A: Bond stretching, B: Angle bending, C: Torsional bending, D: Van-der-Waals interaction, E: Electrostatic interaction.

Depending on their chemical environment, atoms of the same element can have different sets of parameters For example, a sp^3-hybridized carbon has different parameters than a sp^2-hybridized or sp–hybridized carbon. All force field parameters are either obtained from quantum mechanical calculations or fitted to experimental data.

E_{str} – Stretch Energy
The energy that is required to stretch a bond that connects two atom types A and B is calculated with a harmonic potential:

$$E_{\text{str}} = \frac{1}{2} \sum^{bonds} k_i \left(r_i - r_i^0\right)^2 \tag{4.2}$$

where k_i is the force constant for the bond, r_i the bond length and r_i^0 the equilibrium bond length. Due to the harmonic approximation, bonds cannot be broken or formed with molecular mechanics.

E_{bend} – Bending Energy
The energy for bending an angle formed by three atoms A–B–C is also approximated by a harmonic potential:

$$E_{\text{bend}} = \frac{1}{2} \sum^{angles} k_j \left(\phi_j - \phi_i^0\right)^2 \tag{4.3}$$

where k_j is the force constant, ϕ_j the angle between the bonds and ϕ_j^0 the equilibrium angle.

E_{tors} – Torsional Energy
Four sequentially bonded atoms A–B–C–D form a torsional angle ω. Since ω is periodic, the torsion energy is modeled as a cosine function:

$$E_{\text{tors}} = \frac{1}{2} \sum^{torsions} V_n \cdot cos(n\omega) \tag{4.4}$$

where V_n is the amplitude and n the periodicity.

E_{vdw} – Van der Waals Energy
The Lennard-Jones energy expression consists of a repulsive (r^6) and an attractive (r^{12}) part:

$$E_{\text{LJ}} = \sum^{pairs} \epsilon_i \left[\left(\frac{r_i^0}{r_i} \right)^{12} - 2 \left(\frac{r_i^0}{r_i} \right)^{6} \right] \tag{4.5}$$

At small distances the potential is positive, has a slightly negative minimum of depth ϵ at r_0, and approaches zero at large distances.

E_{el} – Electrostatic Energy
The electrostatic energy between two atoms with point charges q_i and q_j is described with the Coulomb potential:

$$E_{\text{el}} = \sum^{pairs} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \tag{4.6}$$

where ϵ_0 is the dielectric constant and r_{ij} the distance between the two atoms.

Note that non-bonded interactions are only calculated between atom pairs that are separated by more than three covalent bonds. Interactions between neighboring atoms

and atoms that are separated by two and three bonds are already covered by the bonded energy terms.

4.2. Hybrid QM/MM

Force field methods make it possible to simulate large systems with several hundreds of thousands of atoms in an efficient way. However, a major drawback is the incapability to describe chemical reactions involving bond breaking and formation or charge transfer processes. This limitation can be overcome by combining force fields with QM methods. In so-called hybrid QM/MM schemes the part of the system in which the reaction occurs (for example a substrate and the active site in an enzyme) is described with QM and the rest of the system with MM. Hence, the system is partitioned into two regions, a QM region and a MM region.

The Hamiltonian of the system is obtained as the sum of the Hamiltonians of the MM region, the QM region and their interactions (additive scheme):

$$\hat{H} = \hat{H}^{MM} + \hat{H}^{QM} + \hat{H}^{QM/MM} \tag{4.7}$$

While the first two contributions can be calculated directly with the force field and the QM method, the latter is more demanding.

Bonded and van der Waals interactions
The border between QM and MM region may intersect covalent bonds which then have to be treated with a link atom scheme. In proteins, only the sidechains are described with QM and a link atom is placed along the C_β and C_α bond at a fixed distance. In this example, the link atom is a hydrogen atom that belongs to the QM region and creates a methyl cap. Following the *divided frontier charge* scheme, the charges of Cα and Hα are set to zero to avoid overpolarization of the atoms at the QM/MM border. The remaining net charge is redistributed among the backbone N and C atoms to achieve charge neutrality. Bending and torsional energies are not calculated by the force field if two out of three or three out of four atoms atoms belong to the QM region, respectively. The van der Waals interaction between the QM and MM atoms is calculated by the force field using the Lennard-Jones potential with standard force field parameters.

Electrostatic interactions
There are several schemes to calculate the electrostatic interactions between MM and QM atoms.

In the *mechanical embedding* scheme, the electrostatic energy is calculated at force field level. The charges of the QM atoms are represented by point charges which are evaluated for an isolated QM region. Consequently, the QM atoms are not polarized by the MM system.

Polarization effects are included in the *electrostatic embedding* scheme which was also used in this work. With DFTB as QM method, the electrostatic potential induced by all MM atoms enters the DFTB3 Hamiltonian matrix elements and affects the QM charge distribution.

With *polarization embedding* both subsystems polarize each other. However, this requires a polarizable force field which is computational costly and therefore not commonly used.

4.3. Molecular Dynamics

Ensemble
The calculations of many properties require a statistical ensemble, a collection of configurations where each configuration is in a different state. Alternatively, one can make use of the *ergodic hypothesis*:

$$\langle A \rangle_{time} = \langle A \rangle_{ensemble} \tag{4.8}$$

which states that the time average from a time-correlated series of structures (trajectory) is equivalent to the average obtained from a large number of atoms over a short time period, for example from experimental measurements.

Equation of Motion
Molecular Dynamics (MD) generates trajectories by propagating the nuclear degrees of freedom with Newton's second equation of motion $F_i = m_i \cdot a_i$. Considering a MM description of the system, the force F_i on atom i is calculated as the negative derivative of the force field energy E with respect to position $\mathbf{r}_i$:

$$-\frac{\partial E}{\partial \mathbf{r}_i} = m_i \cdot \frac{\partial^2 \mathbf{r}_i}{\partial t^2} \tag{4.9}$$

The differential equations are solved numerically by algorithms to calculate the new positions of the atoms after a small time step Δt. A commonly used algorithm is the *leap-frog* integrator which conserves the energy and momentum of the system and is computationally efficient. The new velocities are calculated at half time step and the new positions at full time step:

$$\begin{aligned} \mathbf{v}_{\left(t+\frac{1}{2}\Delta t\right)} &= \mathbf{v}_{\left(t-\frac{1}{2}\Delta t\right)} + \underbrace{\frac{\mathbf{F}_{(t)}}{m}}_{\mathbf{a}_{(t)}} \Delta t \\ \mathbf{r}_{(t+\Delta t)} &= \mathbf{r}_{(t)} + \mathbf{v}_{\left(t+\frac{1}{2}\Delta t\right)} \Delta t \end{aligned} \tag{4.10}$$

The stability and accuracy of the trajectory depends on the time step Δt. As a rule of thumb, Δt should be set to 1/10 of the fastest motion in the system. This is typically the vibration of a hydrogen atom bonded to a heavy atom with ~10 fs. Hence, a time step of 1 fs is the default value for MD simulations which can be increased to 2 fs if constraints are applied to keep the bonds at their equilibrium lengths.

Thermostat and Barostat
The described MD scheme generates a microcanonical ensemble, which is also called an NVE ensemble, where the total number of atoms N is constant as wells as the temperature

T and the total energy E of the system. The temperature T and pressure p fluctuate, which does not correspond to experimental reality. A canonical NVT ensemble in which the temperature is conserved can be generated using a thermostat algorithm that couples the system to a "*heat bath*". The commonly used *Nosé-Hoover* thermostat considers the heat bath as an internal part of the system that adds an extra degree of freedom to the Hamiltonian. In a similar approach, the system can be coupled to an external pressure bath with the *Parinello-Rahman* barostat that scales the volume of the system to generate an isothermal-isobaric NPT ensemble.

5. Enhanced Sampling

According to the *ergodic hypothesis* (Eq. 4.8) the ensemble average of a system is equivalent to the time average of the system. Sampling the complete 6N-dimensional phase space yields the Boltzmann probability function:

$$P(\mathbf{r}, \mathbf{p}) \propto \exp\left(\frac{-E(\mathbf{r}, \mathbf{p})}{k_B T}\right) \tag{5.1}$$

which exponentially depends on the energy $E(\mathbf{r}, \mathbf{p})$ of the system, k_B is the Boltzmann constant and T the absolute temperature. Thermodynamic properties can be calculated from the Boltzmann distribution function such as the Helmholtz (F) free energy:

$$F = k_B T \ln \int \mathrm{e}^{E(\mathbf{r},\mathbf{p})/k_B T} P(\mathbf{r}, \mathbf{p}) \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{p} \tag{5.2}$$

However, in standard MD simulations, the system is limited to configurations that are accessible at the given temperature and therefore the system fluctuates around its equilibrium state. For example, if two configurations are separated by a barrier of more than a few $k_B T$ at room temperature, the time required for the transition can be up to milliseconds which exceeds the time scale of MD simulations. In order to accelerate such processes, different enhanced sampling methods have been developed. They can be divided into two categories: (i) collective variables based methods and (ii) collective variable free methods.

Collective Variable
Collective variables (CVs) are predefined reaction coordinates that are used to describe a reaction path. They can be any function $\xi(\mathbf{r})$ of atomic coordinates such as a distance between two atoms, an angle between three atoms or a dihedral angle between four atoms. Even more complex CVs can be designed which include many or even all atoms, for example, a normal mode from a harmonic vibrational analysis or a RMSD to a reference structure.

Potential of Mean Force
The free energy surface along the selected reaction coordinate ξ is referred to as the potential of mean force (PMF). The free energy difference ΔF between two states ξ_A and ξ_B is obtained as:

$$\Delta F = F(\xi_B) - F(\xi_A) = -k_B T \ln \frac{P(\xi_B)}{P(\xi_A)} \tag{5.3}$$

where $P(\xi_B)$ is the probability of finding the system in state ξ_B and $P(\xi_A)$ the probability of finding the system in state ξ_A. If the correct probabilities of finding the system in a certain state are known, the PMF force along the reaction coordinate can be calculated.

Assume a system with a PMF as shown in Fig. 5.1. There are two minima which are separated by an energy barrier. The system starts in the left minimum and fluctuates around its equilibrium state. When the MD simulation is performed long enough and/or the energy barrier is not too high the system will eventually cross the barrier. Thus, the second minimum is also sampled and from the obtained probabilities the PMF can be calculated with Eq. 5.3. However, if the simulation is too short or the barrier too high the system will remain in the left minimum and an inaccurate PMF will be obtained. To ensure that the whole phase space of the reaction coordinate is sampled accurately, several methods introduce bias potentials, such as Umbrella Sampling or metadynamics.

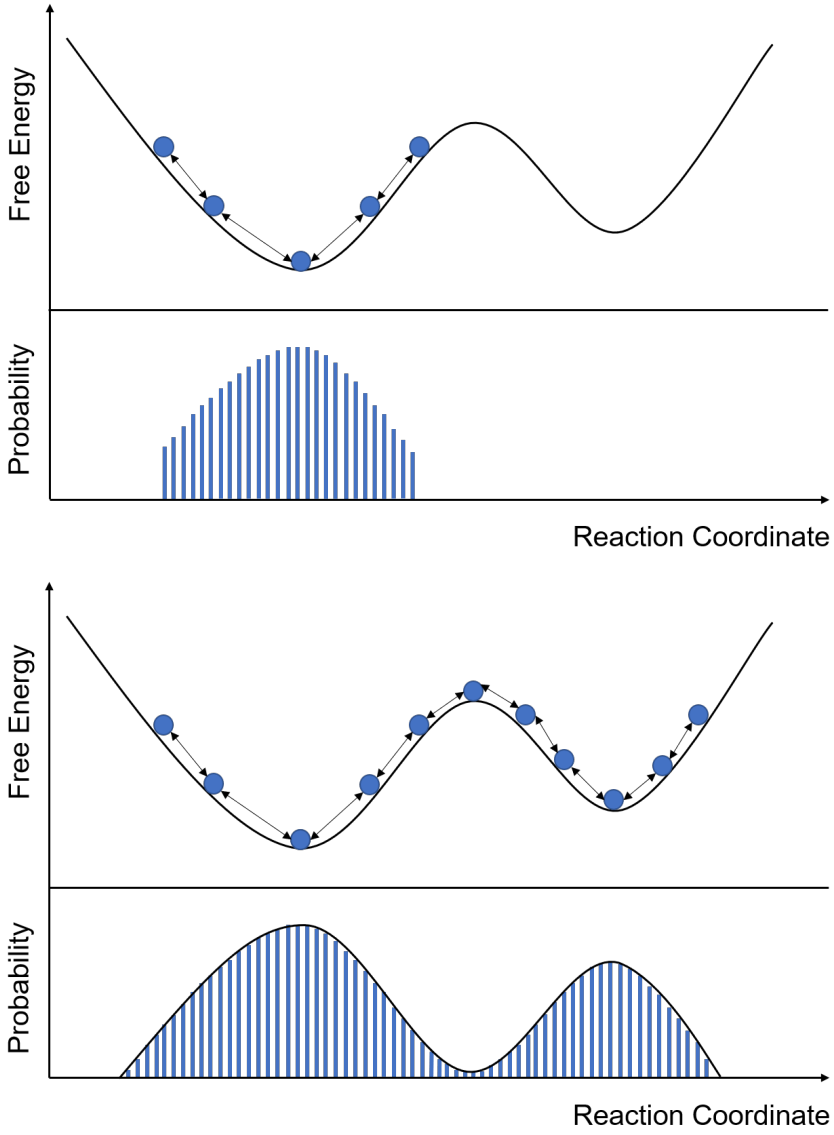

Figure 5.1.: Example of a PMF with two minima that are separated by an energy barrier. The system may overcome the energy barrier when sampled long enough. From the obtained probabilities of finding the system at a certain reaction coordinate value the PMF can be calculated.

5.1. Umbrella Sampling

In the Umbrella Sampling[57] approach the reaction path is split into *windows*, compare Fig. 5.2. In each window bias potential $V(\xi)$ is applied to restrain the system to a region that would otherwise remain undersampled. The bias potential only depends on the reaction coordinate and usually takes the form of a harmonic potential:

$$V(\xi) = \frac{1}{2}k(\xi - \xi_0)^2 \tag{5.4}$$

where k is the force constant and ξ_0 the CV value at which the potential is centered. For each window i one MD simulation is performed from which the biased probability $\mathcal{P}_i(\xi)$ is obtained. The unbiased free energy $F_i(\xi)$ of the ith window is obtained as:

$$F_i(\xi) = -k_B T \ln \mathcal{P}_i(\xi) - V_i(\xi) + C_i \tag{5.5}$$

where C_i is an unknown constant. Hence, $F_i(\xi)$ and $F_{i+1}(\xi)$ are shifted by $C_{i+1} - C_i$. To reconstruct the complete free energy surface, the constants must be determined such that all $F_i(\xi)$ overlap sufficiently, for example with the *weighted histogram analysis method* (WHAM).[58] However, this only leads to meaningful results if the biased probability distribution of each window sufficiently overlaps with the probability distribution of its neighboring windows.

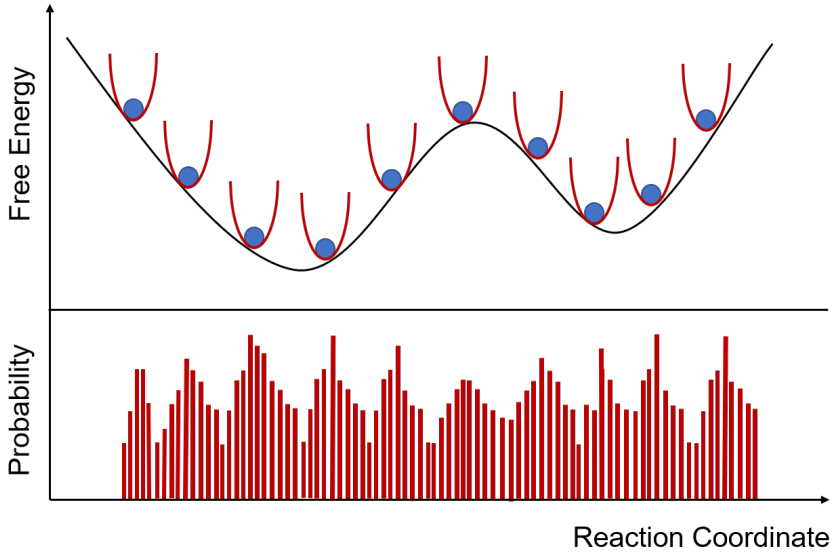

Figure 5.2.: Example of Umbrella sampling. The free energy surface along the reaction coordinate ξ is divided into windows. In each window the system is restrained to a ξ_0 value by a harmonic potential. The complete free energy surface can be reconstructed from the biased probabilities $\mathcal{P}_i(\xi)$ if they overlap sufficiently with $\mathcal{P}_{i-1}(\xi)$ and $\mathcal{P}_{i+1}(\xi)$.

5.2. Metadynamics

In metadynamics, a time dependent repulsive bias potential $V(\xi, \tau)$ is added to the configurational space of the reaction coordinates.[59]. The deposited potential is constructed as a sum of Gaussian potentials:

$$V(\xi, t) = \sum_{k\tau<t} W(k\tau) \exp\left(-\sum_{i=1}^{d} \frac{(\xi_i - \xi_i(k\tau))^2}{2\sigma_i^2}\right) \tag{5.6}$$

with predefined values for the deposition stride τ, the Gaussian height $W(k\tau)$ and the width σ_i of the Gaussian for the i-th CV. Hence, the free energy surface is "filled with computational sand" that elevates the system energetically. This allows the system to explore high energy regions which otherwise would not be accessible on the timescale of MD simulations, compare Fig. 5.3. The underlying free energy profile is calculated as:

$$V(\xi, t \to \infty) = -F(\xi) + C \tag{5.7}$$

where $-F(\xi)$ is the free energy surface estimate and C an additive constant.

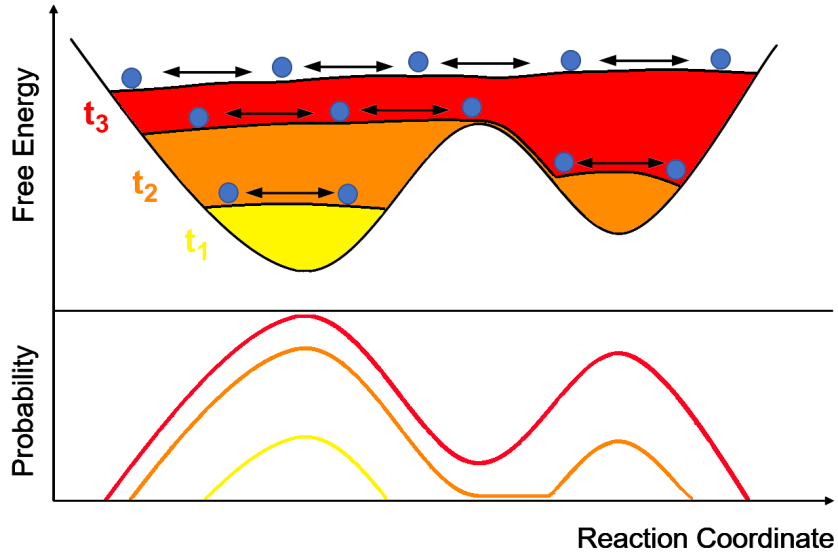

Figure 5.3.: Example of metadynamics. Repulsive gaussian potentials are deposited along the reaction coordinate during the MD simulation that "fill up" the free energy surface. hus, the system is energetically elevated and regions of high energy are sampled. The complete free energy profile is obtained as the sum of the deposited gaussian potentials.

Well-Tempered Metadynamics:
In standard metadynamics the Gaussian height is fixed and remains constant during the simulation. As a consequence, the estimate of the free energy surface oscillates around the "true" free energy surface. One way to minimize the error and smoothly converge

the bias potential $V(\xi, t \rightarrow \infty)$ is well-tempered metadynamics.[60] With this approach the Gaussian height decreases over time according to:

$$W(k\tau) = W_0 \exp\left(-\frac{V(\xi, k\tau)}{k_B \Delta T}\right) \tag{5.8}$$

again with predefined values for the initial height W_0, the deposition rate τ and a parameter ΔT. The free energy profile is obtained as:

$$V(\xi, t \rightarrow \infty) = -\frac{\Delta T}{T + \Delta T} F(\xi) + C \tag{5.9}$$

with the temperature T of the system. The parameter ΔT determines how fast the bias decreases over time and is determined by the bias factor γ:

$$\gamma = \frac{T + \Delta T}{T} \tag{5.10}$$

A bias factor of $\gamma = 1$ ($\Delta T = 0$) corresponds to an unbiased MD simulation whereas a bias factor $\gamma \rightarrow \infty$ is equal to standard metadynamics.

5.3. Simulated Annealing

CVs should be able to distinguish between different states and capture the slowest degrees of freedom during the transition of interest. Therefore, the selection of appropriate CVs is a difficult task and often requires prior knowledge of the studied system. Such information can be obtained using collective variable free enhanced sampling techniques. The easiest way to accelerate the sampling of rare events is to increase the temperature, compare Eq. 5.1. In the simulated annealing approach, the system is heated linearly to a target temperature during the simulation. Consequently, the system might overcome high energy barriers and explore other parts of the free energy surface. By linearly cooling down the temperature again, the system might get trapped in a different energy minimum that would not be accessible during the time-scale of a MD simulation at ambient temperatures. A disadvantage of the method is that barrier heights and generally energy differences between states cannot be determined. In addition, heating the system to high temperatures can lead to conformational changes that would not occur at the reference temperature, e.g., a protein could denature.

5.4. Replica Exchange Molecular Dynamics

Another collective variable free enhanced sampling method is Replica Exchange Molecular Dynamics (REMD).[61] In this approach, several replicas of the system are simulated simultaneously at different temperatures. Each simulation runs independently and the coordinates and velocities of the replicas may exchange between two simulations at different temperatures $T_1 < T_2$. The exchange probability is determined in regular intervals and depends on the instantaneous potential energies U_1 and U_2 according to the Metropolis criterion:

$$\mathcal{P}(1 \leftrightarrow 2) = \begin{cases} 1 & \text{if } U_2 < U_1, \\ \exp\left[\left(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}\right) \cdot (U_2 - U_1)\right] & \text{otherwise.} \end{cases} \tag{5.11}$$

REMD generates a correct ensemble at all temperatures and free energy surfaces can be obtained according to Eq. 5.3. However, the efficiency of REMD scales with the number of degrees of freedom and is therefore not suitable for large systems. Moreover, not all systems are stable at high temperatures as mentioned earlier.

As an alternative, Hamiltonian Replica Exchange[62,63] (HREX) can be used, where the Hamiltonian of the system is scaled by a factor λ. The factor can take values between 0 and 1, which is equivalent to an increase in temperature. For example, scaling the force field energy in Eq. 5.1 by a factor of 2 is equivalent to a doubling of the temperature without actually heating the system. Analogously to REMD, several replicas of the system are simulated with different λ values and at predefined time intervals the exchange probability with neighboring replicas is calculated. Free energy surfaces are obtained from the time-series of the unscaled reference system ($\lambda = 1$).

6. Machine Learning

Machine Learning (ML) learning algorithms estimate functional relationships between a set of input and corresponding output data or from a set of input data alone. Depending on the learning problem, ML algorithms can be divided into two classes: *supervised learning* and *unsupervised learning*.

Supervised Learning
Supervised learning is used when there is a clear input-output relationship in the data. The task of the ML *model* is to find a function that maps the inputs X to the *labelled* outputs Y:

$$\hat{f} : X \xrightarrow{\text{ML}} Y \tag{6.1}$$

for a set of reference data (*training set*). The best fit to the data is determined by minimizing a so-called *cost* or *loss* function that measures the error between the reference value Y_i and the predicted value $\hat{Y}_i$ by the model for all data points N, such as the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2} \tag{6.2}$$

The final model can then be used to predict the outcome for unseen input data , i.e., data points which were not part of the training set on which the model learned.

Supervised learning models can be trained to solve either regression or classification problems. In a regression problem, the outcome is quantitative, for example, the pK_avalue or the energy of a molecule. In this case linear regression or neural network algorithms can be used as learning method. In a classification problem, the outcome is categorical, such as deciding whether a molecule is acidic or basic. Methods such as logistic regression or k-nearest neighbors can be used for this purpose.

Unsupervised Learning
Unsupervised learning algorithms are used when there is only input data without corresponding output data, i.e., the data is not labelled. The goal is to find patterns in the data or to structure the data in a meaningful way. For example, with a *clustering* algorithm different conformers of a molecule from an MD trajectory can be identified. Another application is the reduction of the dimensionality of the data with a *projection* technique. For example, principal component analysis (PCA) can be used to identify the essential dynamics of a molecule by reducing the representation to the most important degrees of freedom.

In this work, only supervised learning with a neural network algorithm was used to learn the energy difference between CC and DFTB level of theory for a small molecule. Thus, only neural networks and the used descriptor (i.e., how the molecule was encoded) are described in the following sections.

6.1. Artificial Neural Networks

Artificial neural networks (ANNs) are inspired on the neural network model of the biological brain, in which neurons are connected by synapses that pass signals from one neuron to other neurons until the target cell is reached. Similarly, the output in an ANN is simply a function of the inputs. The neurons, which are real numbers, are organized in multiple layers: an input layer, an optional number of hidden layers, and an output layer, see Fig. 6.1.

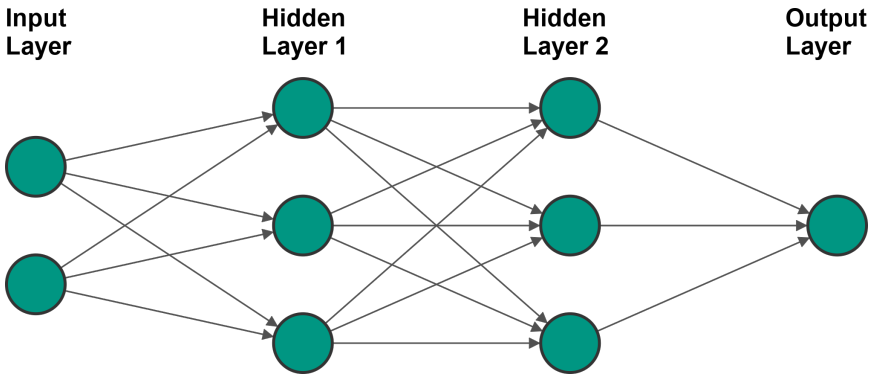

Figure 6.1.: Exemplary structure of an Artificial Neural Network (ANN). The shown ANN consists on an input layer, two hidden layers and an output layer. The output is a linear combination of the input neurons, represented by the connecting arrows. Each connection is associated with a weight parameter which is optimized during the training process.

Each neuron in the hidden layers and the output layer are a linear combination of the neurons in the respective previous layer transformed by an activation function. The value of the j^{th} neuron in the l^{th} layer is obtained as:

$$a_j^l = f\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \tag{6.3}$$

where f is a differentiable, nonlinear *activation function*, w_{jk}^l the *weight* connecting the k^{th} neuron in the $(l-1)^{\text{th}}$ layer to the j^{th} neuron in the l^{th} layer, and b_j^l an additional constant called *bias*. Frequently used activation functions are the sigmoid function or the tangens hyperbolicus (tanh).

Parameter optimization
During the training process, the task is to find the weights that minimize the cost function (e.g., the RMSE in Eq. 6.2). For this, a gradient descent algorithm is used, called *backpropagation* in this context. Starting from randomly assigned weights w^l_{jk}, the partial derivative $\partial C/\partial w_{ij}$ of the cost function C w.r.t. any weight w_{ij} is calculated. The new weights are obtained by following the gradient a small step η in direction of the minimum:

$$w_{ij}^{new} = w_{ij} - \eta \frac{\partial C}{\partial w_{ij}} \tag{6.4}$$

This process is repeated until the minimum is reached. The step size η is called the learning rate and must be chosen carefully. If the learning rate is too high, the minimum may be overshot. However, if the learning rate is too low, the minimization may take too long to converge or can get stuck in a small local minimum. Thus, one should start with a high learning rate at the beginning which then is reduced during the minimization process.

Not only the weights have to be optimized but also the so-called *hyperparameters*. In the context of ANNs, hyperparameters can be the number of neurons and layers or the type of activation and cost functions. Thus, repeated training with the same data is required to find the hyperparameters that give the best results for the task at hand.

Overfitting
One problem that can occur is that the model fits too closely to the training data, called *overfitting*. As a consequence, the model is no longer universal enough to accurately predict the outcome of unseen data. Hence, all ML models should be tested on a subset of the data (*validation* set) that was not part of the training data. Models are usually overfitted if they exhibit small errors in the training set but high errors in the validation set. Overfitting can be avoided by periodically pausing the training process (*early stopping*) and evaluating the performance on the validation set. Alternatively, a "penalty" to input parameters with larger weights can be applied (*regularization*) or more training data can be generated and included in the learning.

6.2. Representation of Molecules

In chemistry, molecules are typically represented by Cartesian coordinates. Cartesian coordinates, however, cannot be used as input in ML because they are not rotationally and translationally invariant. For example, moving or rotating a molecule in vacuum creates a new set of Cartesian coordinates, even though the relative positions of the atoms and the properties of the molecules have not changed. In addition, the representation must be invariant to permutations, i.e., swapping the positions of two atoms of the same element must yield the same result.

Descriptors
Many different *descriptors* have been developed to encode molecules into numerical values that meet the above requirements. They can be classified into *global* and *atomic* descriptors

Global descriptors are typically based on the distance between atoms, such as the Coulomb Matrix[64]:

$$c_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{\left\| R_i - R_j \right\|} & \text{for } i \neq j \end{cases}, \tag{6.5}$$

the Sine matrix[65] or the Bag of Bonds[66].

In contrast, atomic descriptors describe chemical systems as a set of atomic environments. The first atomic descriptor was developed by Behler and Parinello in 2007, the so-called *atom-centered symmetry functions* (ACSF).[67] Since then, many other atomic descriptors have been developed, such as the smooth overlap of atomic positions (SOAP)[68] or the Faber-Christensen-Huang-Lilienfeld (FCHL)[69,70] descriptor. In this work, ACSFs were used and are therefore described in more detail below.

Atom-Centered Symmetry Functions

Only atoms that are within a cutoff radius R_c of a given central atom i are considered as environment in ACSFs. This approach reduces the dimensionality and consequently the computational effort. However, a too small radius will lead to inaccurate fits. Thus, R_C is a hyperparameter that must be optimized during the learning process. For a smooth decay to 0 beyond R_C, a cutoff function $f_c(R_{ij})$ is used, for example:

$$f_c\left(R_{ij}\right) = \begin{cases} 0.5\left[\cos\left(\pi \frac{R_{ij}}{R_c}\right) + 1\right] & \text{for } R_{ij} \leq R_c \\ 0 & \text{for } R_{ij} > R_c \end{cases} \tag{6.6}$$

where R_{ij} is the distance between the central atom i and a neighboring atom j.

Inside the cutoff sphere two types of symmetry functions are used: *radial* and *angular* ACSFs. Radial ACSFs include two-body terms and consists of a element-based sum of Gaussians and cutoff function products:

$$G_{i,Z}^{\text{Rad}} = \sum_{j \neq i}^{N_{\text{atom}}^{Z}} e^{-\eta\left(R_{ij} - R_s\right)^2} f_c\left(R_{ij}\right) \tag{6.7}$$

where R_{ij} the distance between the central atom i, for which the symmetry function is calculated, and a neighboring atom j with atomic number Z. This means that for each atom i one symmetry function $G_{i,Z}$ is calculated per element in the system:

$$N_{\text{sym}} = N_{\text{elem}} \tag{6.8}$$

The η and R_s are hyperparameters that can be modified in order to generate more than one set of symmetry functions. Typically, six radial functions per element are used by modifying η and R_s.

Angular ACSFs include three-body terms via the angle θ_{ijk} between a given central atom i and two neighboring atoms j and k within R_C. A commonly used function is:

$$G_{i,Z_1,Z_2}^{\text{Ang}} = \sum_{j \neq i}^{N_{\text{atom}}^{Z_1}} \sum_{k \neq i,j}^{N_{\text{atom}}^{Z_2}} \frac{2^{1-\zeta}\left(1 + \lambda \cos\theta_{ijk}\right)^{\zeta}}{e^{\eta\left(R_{ij}^2 + R_{ik}^2 + R_{jk}^2\right)^2}} f_c\left(R_{ij}\right) f_c\left(R_{ik}\right) f_c\left(R_{jk}\right) \tag{6.9}$$

which calculates one symmetry function for each combination of elements with atomic numbers Z_1 and Z_2. Including all permutations the number of symmetry functions is:

$$N_{\text{sym}} = \frac{N_{\text{elem}}(N_{\text{elem}} + 1)}{2} \tag{6.10}$$

ζ and λ are hyperparameters which can be modified in order to obtain more than one set of symmetry functions. Typically, λ is 1 and -1, and $\zeta = 1, 2, 4$, and 16.

Example

To clarify the calculation of ACSFs, consider a methylthiolate (CH_3S^-) in vacuum (Fig. 6.2). The system consists of five atoms from three elements: C, H and S. Hence, the environment of each atom is described by three radial symmetry functions (for a fixed η and R_s value) and six angular symmetry functions (for a fixed ζ and λ value). The cutoff radius R_C must be chosen such that all "important" atoms are included for the problem at hand. Here, all atoms "see" each other. Fig. 6.2 illustrates how the symmetry functions are calculated for the carbon atom with index 0.

The first radial function $G_{0,C}$ considers all carbons in the system around the reference atom, which is the only carbon in the system, therefore $G_{0,C} = 0$. The second radial function $G_{0,H}$ considers all hydrogens and is a sum over all Gaussians that include hydrogens. The third function $G_{0,S}$ considers all sulfur atoms. According to the same principle, the angular ACSFs are calculated.

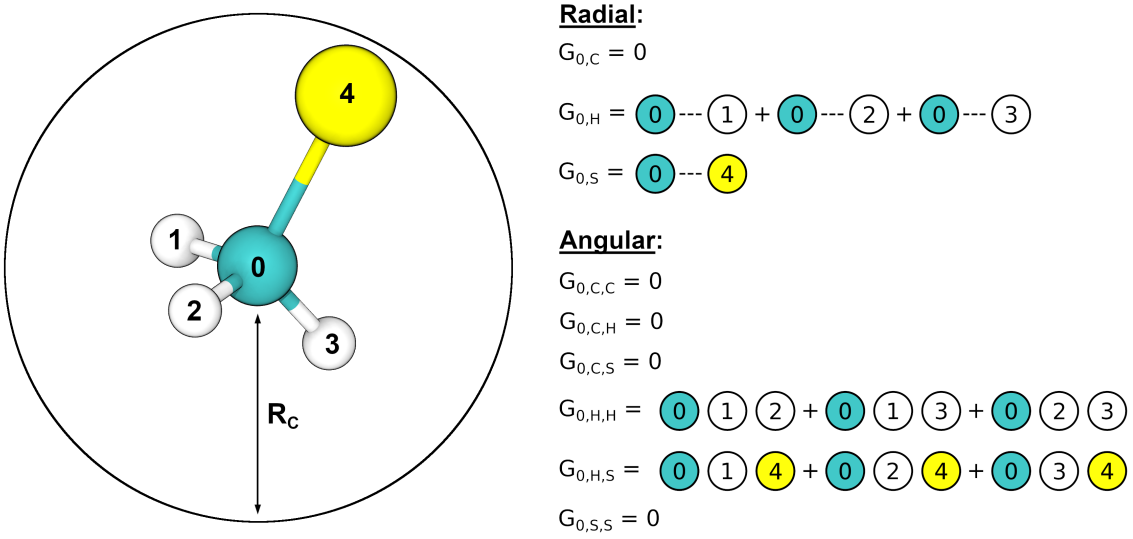

Figure 6.2.: Illustration of the contributions of radial and angular atom-centered symmetry functions for the carbon atom of a methylthiolate in vacuum.

Part III.

Contributions

7. O to bR transition in bacteriorhodopsin occurs through a proton hole mechanism

Chapter 7 is reproduced in parts from Ref. [71]:

- Maag, D.; Mast, T.; Elstner, M.; Cui, Q.; Kubař, T., O to bR transition in bacteriorhodopsin occurs through a proton hole mechanism., Proc. Natl. Acad. Sci. USA 2021, 118 (39).

Author Contributions:
This work was done in cooperation with Thilo Mast, who already published parts of the results in his doctoral thesis (Ref. [72]).

Thilo Mast built the system, i.e., embedded the bR molecule in a POPC lipid bilayer sourrounded by aquoeus solution, and developed structural models of the different protonation states (**bR**, **O** and **O*** state) with classical HREX extended-sampling, cf. Sec. 7.2.1. Moreover, he performed QM/MM molecular dynamics and QM/MM metadynamics to investigate the PT in the **O**→**O*** transition.

Denis Maag carried out additional analyses of the HREX trajectories of the **bR**, **O** and **O*** state and reproduced the QM/MM metadynamics simulations of the PT in the **O**→**O*** transition with different settings, cf. Sec. 7.2.2. Moreover, Denis Maag performed the long-range PT of the **O**→**bR** transition with QM/MM metadynamics and analyzed the obtained trajectories (cf. Sec. 7.2.3) and compared different DFTB3 QM/MM approaches (cf. Sec. 7.2.4).

7.1. Introduction

The generally accepted sequence of the photocyle is shown in Fig. 7.1 together with a bR monomer and two relevant retinal configurations. The retinal is centrally located in the bR monomer and bound to Lys216 via a protonated Schiff base. The intermediate states are characterized by their UV–vis absorption maxima.

bR→K

In the ground state (**bR**), three protons which are involved in the proton transfer steps are stored in the protein: (i) one at D96 at the proton uptake site, (ii) one at the protonated Schiff base (SB) in the center, (iii) one at a so-called proton release group (PRG) near the

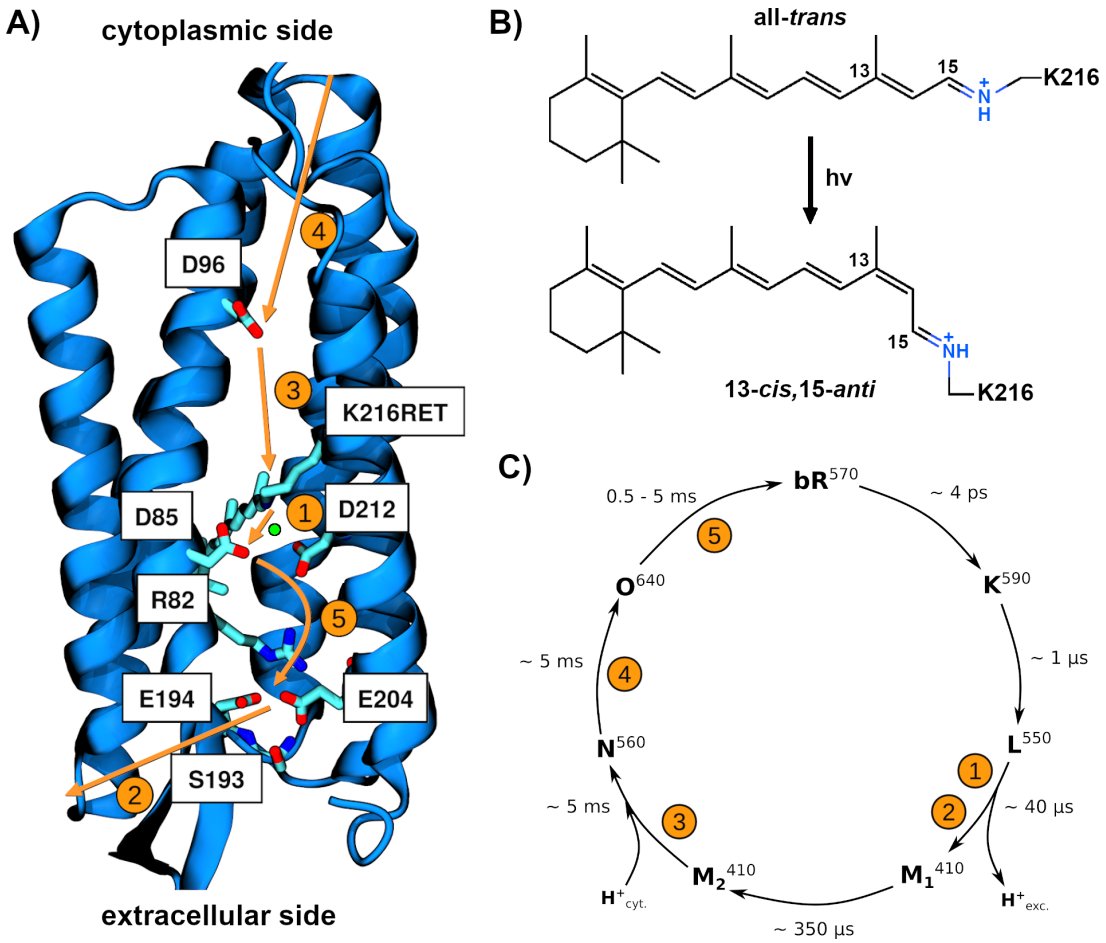

Figure 7.1.: A: Side view of the bR monomer in PDB ID 5B6V.[73] Functionally important amino acids are shown and labeled. Key proton transfer steps are labeled 1 through 5. B: Retinal linked to Lys216 via protonated Schiff base. Absorption of a photon leads to the isomerization from all-*trans* to 13-*cis*,15-*anti* configuration. C: Scheme of the bR photocycle with the important intermediates and their lifetimes.

extracellular side, which is either delocalized between the two glutamates E194 and E204 (or possibly stored on a protonated water cluster).[74,75]

The retinal is in an all-*trans* configuration and the NH group of the Schiff base (SB) is facing towards the extracellular side. The proton of the SB is stabilized by a pentameric cyclic hydrogen bonding network formed between the SB, D85, D212 and three water molecules w401, w402, and w406.[76,77] Additional hydrogen bonds are formed with other nearby polar residues, such as R82. The photocycle is initiated when the retinal absorbs a photon with a wavelength of 570 nm, which corresponds to an energy of ca. 50 kcal/mol. During this process (~4 ps), the retinal isomerizes from all-*trans* to the 13-*cis*,15-*anti* configuration and the **K** state is formed.[78,79]

K→L

Compared to the ground state, the **K** intermediate is red-shifted and the retinal adopts different conformers depending on the temperature of the experiments. At low temperatures around 100 K it is highly twisted and orientated towards D212,[79,80] storing 11-14 kcal/mol of the photon energy due to the high strain. At temperatures >150 K, the retinal chain either relaxes and decays into the blue-shifted **L** state over a complex relaxation path and a timescale of ~1 µs or converts back to the ground state.[81,82] However, the detailed mechanism remains largely unknown, as the outcome of the experiments strongly depends on the experimental conditions ans consequently the **K** and **L** structures are not well defined. Nevertheless, the **K**→**L** transition is known to involve structural changes in the active site which breaks the pentameric HBN and therefore enables the first PT in the next step of the photocycle.[73]

L→M_1

In the **L**→$\mathbf{M_1}$ transition the proton of the SB is transferred to D85 on a timescale of ~40 µs. Again, there is no consensus about the detailed mechanism and several reaction pathways have been proposed involving indirect PTs via intermediate proton carriers, such as T89, D212 including or not including a water molecule.[83,84]

M_1→M_2

As a results of the change in the protonation pattern, conformational changes occur. The deprotonated SB reorients in direction of the cytoplasmatic side and R82 moves closer to E194 and E204, which are located at the extracellular side of the protein.[76,85,86] Due to the change in the electrostatics, the pK_a of the PRG is lowered to 5.7 and the proton released at neutral pH, leading to the blue-shifted $\mathbf{M_2}$ state over a timescale of ~350 µs.[87–89] At acidic pH below the pK_a of the PRG, the proton release is delayed until the end of the photocycle.

M_2→N

Subsequently, the SB is reprotonated by the 10 Å distant D96 located near the cytoplasmatic side of bR. Water enters the proton uptake side and forms a water wire connecting D96 and the SB.[90,91] As a result, the pK_a of D96 is lowered to ~7.5 and the proton transferred to the SB which has a pK_a of ~8.[92,93] The process takes place on a timescale of ~5 ms and results in the red-shifted **N** state.

N→O

Over the following ~5 ms, the **O** state is formed: D96 gets reprotonated by cytoplasmatic water and the retinal reisomerizes to the all-*trans* form which is assumed to adopt a twisted conformation.[94,95] Moreover, the extracellular side of the protein is supposed to be in a more open conformation state due to structural rearrangements of the helices. However, there is only little structural information available for the red-shifted **O** state; only four **O**-like crystal structures are available in the protein database which are based on bR mutants.

O→bR

The final PT occurs during the transition from the **O** state back to the ground state **bR** at a pH >6.[96] In this last step, the proton is transferred from the centrally located D85 to the ~15 Å distant PRG at the extracellular side of the protein on a timescale of 0.5–5 ms.[96–99] However, due to the sparse structural information of the **O** state, the detailed mechanism of this remarkably long-ranged proton exchange remains largely unclear. Time-resolved Fourier transform infrared (TR-FTIR) spectra indicate that D212 is transiently protonated during the transition from **O** to **bR**, constituting a new intermediate termed **O***.[100–102] In addition, the side chain of R82 is expected to move towards D85 and D212 during long-range PT.[103]

Several pathways have been proposed for the long-range PT, e.g., a Grotthus-like mechanism via a continuous water wire connecting D85 and the PRG (including E194 and E204). However, such a water wire may be disrupted by the side chain of R82, which is located between the two motifs. Previous QM/MM minimum energy path (MEP) calculations of the Grotthus mechanism showed a very high reaction barrier of 22 kcal/mol and a strong endothermic product (i.e., the PRG being protonated)[104], which contradicts the experimentally estimated kinetics and exergonicity of the reaction based on pK_a differences of D85 and the PRG.[105,106] Therefore, it was proposed that the PT is mediated by the R82 side chain, i.e., that R82 acts as a temporary proton acceptor/donor[103], for which, however, QM/MM-MEP calculations revealed an even higher barrier of 36 kcal/mol.[107] Moreover, bR mutants in which R82 has been replaced by an alanine or glutamine are still capable of pumping protons, albeit less efficiently.[108,109]

In this study, by combining extensive classical and QM/MM molecular dynamics (MD) simulations, we are able to gain new insights into the structural features of the **O** state and its transition back to the **bR** state. In particular, we show that the hydration level of the protein cavity increases substantially in the **O** state to enable the long-range proton exchange between D85 and PRG. Moreover, by computing three-dimensional free energy surfaces with ~0.1 μs multi-walker metadynamics, we show that the proton exchange in fact involves the transfer of a proton hole[110,111], which is stabilized by the intervening R82; the unique simulations were made possible by the use of an efficient, semiempirical QM/MM potential and a collective coordinate for describing PTs without a priori assumption of the transfer mechanism.[112] The computed free energy barrier and exergonicity are consistent with experimental estimates, and the mechanism provides an explanation for the Cl^- pumping activity of the D85S mutant.

7.2. Computational Details

7.2.1. Models for the bR, O and O* states – Hamiltonian Replica Exchange simulations

For developing the structural models for **O** and **O*** state, we started with a recently released structure of the bR ground state obtained by time-resolved serial femtosecond crystallography, PDB ID 5B6V.[73] The retinal moiety was in the all-*trans* state. Water molecules resolved in the crystal structure (17) were included in the simulation because most of these water molecules are located in the active site cavity close to D85, D212, R82, E204 and E194, and this region contains structurally important water molecules that constitute a hydrogen bond network (HBN). This decision is supported by the relatively high crystallographic resolution of 2.0 Å and the very low, thus reliable, B-factors (38.9–72.7 Å^2) of the water oxygens.

The CHARMM36 force field[113] was used for the description of the whole system comprising protein, lipid phase, water phase and ions. Parameters for the retinal molecule were taken from Refs. [114–118]. Gromacs 5.0[119–122] in combination with Plumed 2.1.1[123] was used for MD simulations. All titratable amino acids were kept in their physiological protonation states, identical to the bR state, during topology creation with *pdb2gmx*; Asp96 was protonated in compliance with both O and bR states. A POPC bilayer pre-equilibrated for 500 ns was considered as a lipid environment.

The InflateGRO methodology[124] was used to embed the bR protein into the POPC bilayer. The bR protein was embedded into the POPC bilayer with InflateGRO. The protein was centered in the lipid bilayer during this procedure. In the next step, overlapping lipids within a cutoff radius of 14 Å were removed, and the bilayer was artificially inflated in its plane. Then, several shrinking steps followed which compressed the bilayer and packed the lipids around the protein. The shrinking procedure was terminated when the area per lipid value of the POPC bilayer approached the experimental value of 65.8 Å^2.[125] The protein/membrane-complex was solvated by using the *gmx solvate* routine with the CHARMM-TIP3P water model. Small gaps between the lipid acyl chains also got filled with water; these misplaced water molecules were removed. To achieve electroneutrality, a chloride counterion was added with the *gmx genion* routine. The whole system comprised a ground-state bR molecule truncated to 230 amino acid residues, 284 POPC lipids, 16381 TIP3P waters and one chloride counter ion.

Energy minimization was conducted using the steepest descent minimization algorithm with a maximum force threshold of 1000 kJ/(mol·nm). The system was equilibrated in the NVT ensemble for 200 ps at 300 K using the Bussi thermostat.[126] Furthermore, Cartesian position restraints were placed on the heavy atoms of protein and lipids. An unrestrained NPT simulation over 100 ns followed at 300 K and 1 bar, controlled with the Nosé–Hoover thermostat and the Parrinello–Rahman barostat, respectively. All classical simulations were performed with the leap-frog integrator with a time step of 2 fs. Since MM force fields are incapable of preserving the pentagonal HBN[104], an additional bias potential was applied to keep the HBN intact. To this end, distance-dependent, harmonic restraints were applied between the HBN water molecules and the sidechains of D85, D212 and K216RET. The restraints with a force constant of 8000 kJ/(mol·nm^2) were activated whenever the

distance between the water oxygen atom and the respective amino acid atom exceeded 2.8 Å. Following 100 ns of equilibration of the ground **bR** state, the protonation states of

Table 7.1.: Amino acids in the HREX hot region and reasons for their selection. Reproduced from Ref. [72].

Amino acid	Function
D85	Proton donor + member of the pentagonal HBN
D212	Proton acceptor/donor + member of the pentagonal HBN
K216RET	Member of the pentagonal HBN
R82	Up-/Downswing movement during photocycle
E194	Part of the PRG region
E204	Part of the PRG region
Y185	Hydrogen-bonded to D212
W86	Hydrogen-bonded to D212
E9	Polar amino acid near to the PRG
Y83	Polar amino acid near to the PRG
S193	Hydrogen-bonded to the E204
P77	Amino acid near to the PRG + hydrogen-bonded to S193
Y57	Hydrogen-bonded to D212
F208	Serves together with R82 as hydrophobic plug
Y79	Polar amino acid near to the PRG + hydrogen-bonded to E9
P200	Hydrogen-bonded to E204
W189	Amino acid near to the PRG + hydrogen-bonded to Y83

the key amino acids were adjusted: E204 was deprotonated and D85 was protonated for the **O** state, while D212 was protonated for the **O*** state. The resulting models were then prepared for enhanced sampling simulations employing Hamiltonian Replica Exchange (HREX).[127] Since it is known that the NPT ensemble disrupts the HREX exchange rates, these simulations were conducted in the NVT ensemble with the Bussi thermostat.[126] The actual HREX simulations had the same setup for all states: 16 replicas were used which spanned a range of scaling with λ between 1–0.16, corresponding to an effective temperature range of 300–1900 K, and each replica was simulated for 10 ns. The frequency for exchange attempts was set to 4 ps. The so-called hot region (i.e., the part of the molecular system in which the scaling is applied) consisted of the active site amino acids residues D85, D212, K216RET, R82, E194 and E204, and additional residues surrounding these, as detailed in Tab. 7.1. The finished simulations were then checked for sufficient exchange rates between pairs of replicas. The trajectory of each respective unperturbed replica ($\lambda = 1$) served as basis for the structural analysis.

7.2.2. O→O*: Direct proton transfer — 3D metadynamics.

Initial structures of the **O** and **O*** state were taken from the HREX simulations of the **O** state. Since the PT product (**O*** state) is very similar to the **O** state, the same starting structure was considered with only the protonation state being changed (deprotonated D85 and protonated D212).

The QM/MM simulations and the equilibration were set up analogously for both **O** and **O*** states: The QM region comprised the amino acid sidechains of D85, D212 and K216RET from the Cβ atoms on. Furthermore, the QM region also included four water molecules located in the proximity of the three mentioned amino acid residues. In total, the QM zone comprised 90 atoms, and was treated using DFTB3 with the 3OBw parameters[128]; the purpose of the 3OBw reparametrization is to improve the description of liquid water, and it was chosen in this study in order to describe the water wire well. The BJ implementation of D3 correction[129] was employed to compensate for the missing description of dispersion interaction in DFTB. The good performance of DFTB3/3OBw for the simulation of PT reaction involving aspartate sidechains and water molecules was confirmed in a benchmark, performed by Thilo Mast, that used B3LYP calculations as a reference. For a more detailed discussion of the benchmark see Ref. [71, 72].

The rest of the system was described with the CHARMM36 force field. The leap-frog algorithm was used to solve Newton's equations of motion with a time step of 0.5 fs. For the temperature and pressure coupling, the Nosé–Hoover thermostat and the Parrinello–Rahman barostat were used, respectively. An NPT equilibration was conducted over 2 ps at 300 K and 1 bar followed by a production simulation over 200 ps. The QM/MM simulations were conducted with a local version of Gromacs 5.0 featuring an implementation of DFTB3[130] combined with Plumed 2.1.1.

The systems were equilibrated with the simulated annealing technique. The system was heated from 300 K to 340 K over the first 30 ps. The temperature stayed at 340 K for the next 40 ps and was then decreased back to 300 K within 30 ps. The last 100 ps were conducted at 300 K. Moreover, to prevent the leakage of the QM-treated water molecules out of the QM region or the influx of MM water molecules into the QM region, spherical harmonic position restraints were applied to the oxygen atoms of the four water molecules. Thereby, the oxygen atom of each of the water molecules was allowed to move freely within a sphere of a radius of 3 Å centered in its initial position. Whenever the distance from the initial position increased beyond 3 Å, a harmonic restraint with a spring constant of 2000 kJ/(mol·nm^2) set in, pulling the water back towards its initial position.

From the last 100 ps of the corresponding trajectories starting structures were selected to generate he free energy profile of the **O**→**O*** with multiple walker (MW) metadynamics.[131] The MW metadynamics was conducted with an in-house version of Gromacs 2020 interfaced with DFTB+ 19.1[132,133] and Plumed 2.5.1.[123,134]

A set of three reaction coordinates was used to describe the **O**→**O*** transition. The first coordinate describes the long-range PT reaction, and is based on the modified center of excess charge (mCEC) coordinate introduced by König et al.[112],

$$\xi = \sum_{i=1}^{N_\mathrm{H}} \mathbf{r}^{\mathrm{H}_i} - \sum_{j=1}^{N_\mathrm{X}} w^{\mathrm{X}_j} \mathbf{r}^{\mathrm{X}_j} - \sum_{i=1}^{N_\mathrm{H}} \sum_{j=1}^{N_\mathrm{X}} f_\mathrm{sw}(d^{\mathrm{H}_i,\mathrm{X}_j}) \left(\mathbf{r}^{\mathrm{H}_i} - \mathbf{r}^{\mathrm{X}_j} \right). \tag{7.1}$$

The first term in Eq. 7.1 is a sum of all hydrogen coordinates $\mathbf{r}^{\mathrm{H}_i}$, and the second term is a weighted sum of positions $\mathbf{r}^{\mathrm{X}_j}$ of all hydrogen-coordinating oxygens taking part in the PT process (with w^{X_j} being the number of hydrogens coordinated to X_j in the least protonated configuration). The last term can be seen as a correction running over all distances between hydrogens and coordinating atoms, in order to decide – on basis of a switching function $f_{\mathrm{sw}}(d^{\mathrm{H}_i,\mathrm{X}_j})$ – which H_i and X_j atoms are connected by bonds,

$$f_{\mathrm{sw}}(d^{\mathrm{H}_i,\mathrm{X}_j}) = \frac{1}{1+\exp\left[\left(d^{\mathrm{H}_i,\mathrm{X}_j} - r_{\mathrm{sw}}\right)/d_{\mathrm{sw}}\right]}. \tag{7.2}$$

Here, $r_{\mathrm{sw}} = 0.125$ nm and $d_{\mathrm{sw}} = 0.006$ nm are empirical parameters that control the steepness and centering of the switching function on the bond length scale. The vector variable $\boldsymbol{\xi}$ represents the spatial location of the excess proton being transferred. $\boldsymbol{\xi}$ was converted to a scalar quantity ζ by considering the distances $d_{\xi,D}$ between mCEC and the initial proton donor D as well as the distances $d_{\xi,A}$ between mCEC and the final proton acceptor A,

$$\zeta = \frac{d_{\xi,D} - d_{\xi,A}}{d_{\xi,D} + d_{\xi,A}}. \tag{7.3}$$

Then, the transfer of the proton is defined in a range from $\zeta = -1$ for the proton located on the initial donor, to $\zeta = 1$ for the proton located on the final acceptor. The mCEC coordinate was used to describe the PT reaction from D85 to D212 and included all four water molecules of the QM/MM region and the D85/D212 oxygen atoms as possible proton coordination sites. The transfer of the excess charge was mapped relative to the D85Cγ and D212Cγ atom using the quantity ζ from Eq. 7.3. The width of the Gaussian hills along this coordinate was 0.05.

The PT may occur either directly or water-mediated. Therefore, a second reaction coordinate was introduced to cover the effect of structure of water present in the active site on the energetics of PT. It appears difficult to drive these processes with a single coordinate, thus it was instead described indirectly, by means of the distance D85Cγ–D212Cγ, which is assumed to correlate with the formation and destruction of the HBN. The width of the Gaussian hills along this coordinate was 0.025 nm, and the sampling space was restricted to the distances D85Cγ–D212Cγ in an interval of 0.35–0.66 nm. Outside of this range, a harmonic restraint with a spring constant of 200,000 kJ/(mol·nm^2) set in.

Water-mediated PT can proceed via a hydronium or a hydroxide ion intermediate. Therefore, a third reaction coordinate, based on the hydrogen coordination number of the water oxygen atoms, was introduced to distinguish between these two pathways. For a hydronium ion, the oxygen coordination number equals three, whereas it equals one for a hydroxide ion. In the case of a direct PT, the coordination number is two. Since the QM region comprises multiple water molecules and each of them could occur in the PT as hydronium or hydroxide, the average water oxygen coordination number was used. The oxygen coordination number was introduced as

$$s_{\mathrm{O}}(r_{ij}) = \sum_{i\in\{\mathrm{O}\}} \sum_{j\in\{\mathrm{H}\}} \frac{1-\left(\frac{r_{ij}}{r_0}\right)^n}{1-\left(\frac{r_{ij}}{r_0}\right)^m} \tag{7.4}$$

with the parameters taking values of $r_0 = 1.2$ Å, $n = 45$ and $m = 90$. $\overline{s_O(r_{ij})}$ was defined as the average water oxygen coordination number of all four water molecules present in the QM region, and the Gaussian width for $\overline{s_O(r_{ij})}$ was set to 0.025. For a more detailed discussion of the coordination number as reaction coordinate see Ref. [71, 72]. For the purpose of presentation, this quantity was converted to net charge on the QM water molecules,

$$q_{\text{net}} = 4 \cdot \left(\overline{s_O(r_{ij})} - 2\right), \tag{7.5}$$

which takes values of −1, 0 and +1 for hydroxide anion, all-neutral QM water and hydronium cation, respectively.

The MW metadynamics simulation was conducted in two steps. First, standard metadynamics was performed over 10 ns with a Gaussian hill height of 0.6 kJ/mol and a deposition period of 0.5 ps. In the second step, the simulations were extended for additional 5 ns with well-tempered metadynamics and a bias factor of 20 resulting in a total simulation time of 15 ns. The free energy surfaces were created with the Plumed program *sum_hills*.

7.2.3. O→bR: Long-range proton transfer — 3D metadynamics.

The **O→bR** proton transfer was also simulated with MW-metadynamics and a QM/MM setup. The simulation employed 20 walkers, and the initial structures were taken from an NPT equilibrated QM/MM simulation of bR in the ground state. The QM region comprised the sidechains of R82, D85, D212, E194 and E204 as well as 12 nearby water molecules. In total, 90 QM atoms were treated with DFTB3 with 3OBw parameters, with additionally modified N–H parameters for 3OB[135], and the DFT-D3 dispersion correction with BJ damping. The rest of the system was described with CHARMM36 forcefield parameters and TIP3P water, as in the previous simulations. Electrostatic interactions between the QM and MM region were scaled down by a factor of 0.75, corresponding to the inverse square root of the optical dielectric constant, to compensate for possibly overestimated interactions between the charged QM molecules and the (electronically non-polarizable) MM environment. The simulation was performed with an in-house version of Gromacs 2020 interfaced with DFTB+ 19.1 and Plumed 2.5.1.

A set of three reaction coordinates, or collective variables (CV), was selected to describe the O→bR transition. First, the modified center of excess charge (mCEC) coordinate, transformed to the scalar quantity ζ, described the proton transfer from D85 to the PRG, cf. Sect. 7.2.2. The oxygen atoms of D85, D212, E194 and E204, and those of all 12 QM water molecules were included in the mCEC coordinate as possible proton donors, relays or acceptors. D85-Cγ was considered as the initial proton donor, and E204-Cδ was the final proton acceptor. The parameters of the switching function were set to $r_{\text{sw}} = 0.125$ nm and $d_{\text{sw}} = 0.006$ nm. As a second CV, the distance between the rigidly placed nitrogen atom of A44 and the Cζ atom of R82 (dR82) represented the orientation of R82 towards the extracellular or cytoplasmic side. The third CV is the distance between the Cδ atoms of E194 and E204 (dPRG) to describe the opening and closing of the PRG.

In pilot simulations, deprotonation of the guanidinium group of protonated R82 sidechain, $R82H^+$, was observed occasionally when a hydroxide ion (proton hole carrier) was in close proximity. Since the description of proton affinity of nitrogen bases with DFTB is poten-

tially difficult, it is a priori unclear if this is a genuine behavior of the system. In fact, previous work showed deprotonated R82 sidechain as an intermediate to lie excessively high in energy[104], and we also show other related protonation patterns in the lower reaction site to be unlikely in Sect. 7.3.5 below. Therefore, it seems unnecessary to consider the case of deprotonated R82° in the description of the **O→bR** transition. Further, we note that several different protonation patterns may correspond to a single ζ value, as explained in Sect. 7.3.4, and including R82 in the PT pathway would make the situation even more complex. Distinguishing between the possible protonation patterns would only be possible if the metadynamics setup is extended with a fourth CV, e.g., the mean coordination number of QM water molecules with all hydrogens like in Sect. 7.2.2. This is not computationally feasible; instead, as a solution to this overwhelming complexity, we decided to forcibly exclude the R82 sidechain from the PT pathway by means of keeping its guanidinium group in the protonated state. Technically, this was performed by restraining the coordination numbers of Nμ1 and Nμ2 with their bonded hydrogen atoms to 2.

Additional restraint potentials were applied to all CVs and QM water to improve the efficiency and stability of the simulations. ζ was restricted to values between −0.84 and 0.84, which corresponds to one of the oxygens of D85 or E204 protonated, respectively. Values beyond these limits lead to pilot simulations crashing because the proton was pushed closer to the carboxyl atoms of D85/E194. The position of R82 was kept in the interval of dR82 of 2.3–2.9 nm to reduce the sampled configurational space. In order to ensure the proper up- and down-swing movement, the side-chain dihedral angles of R82 were restrained to the following ranges: $29° < \chi_1' < 92°$ (C–Cα–Cβ–Cγ), $46° < \chi_2 < 194°$ (Cα–Cβ–Cγ–Cζ), and $43° < \chi_3' < 260°$ (C–Cα–Cγ–Cζ). Pilot simulations performed without these restraints resulted in a 'coiled'-like structure of R82 during metadynamics. The opened PRG conformation was limited to dPRG values smaller than 0.65 nm in order to reduce the configurational space and to avoid influx of MM water into the QM region. To prevent QM water from leaving the reaction center, the O–N distance between QM water molecules and F208 was restrained below 1.2 nm.

Each walker was simulated at 300 K and 1 bar for 5 ns, consisting of 3 ns standard metadynamics and subsequent 2 ns well-tempered metadynamics, resulting in a total simulation time of 100 ns. Gaussian biasing potentials with a height of 0.6 kJ/mol, and a width of 0.05, 0.05 nm and 0.05 nm for ζ, dR82 and dPRG, respectively, were deposited every 500 fs in each walker with an exchange period of 1 ps. A bias factor of 60 was set for well-tempered runs. After every ns, the simulations were checked for MM water in the QM region. Whenever an MM water molecule entered the protein pocket, the initial or a comparable structure of the respective walker was further simulated instead.

The **O→bR** transition might proceed via proton hole or hydronium ion, hence the mean coordination number of the QM water molecules was monitored. If one or two water molecules form a strong (and therefore short) hydrogen bond with a hydroxide ion, the mean coordination number may exceed 2 with standard coordination functions, which falsely indicates a hydronium ion. To ensure that every hydrogen is only accounted for

once, the coordination numbers of all 12 QM water oxygens with each of the 25 protons (24 of QM water and one located initially on D85) were calculated separately,

$$\forall i \in \{1, 2, \ldots, 25\} : s_{\text{coord}}(r_{ij}) = \sum_{j \in \{\text{O}\}} \frac{1 - \left(\frac{r_{ij}}{r_0}\right)^n}{1 - \left(\frac{r_{ij}}{r_0}\right)^m}, \tag{7.6}$$

with $r_0 = 1.22$ Å, $n = 45$ and $m = 90$. Each of the 25 coordination numbers $s(r_{ij})$ is then passed to a hyperbolic tangent function,

$$\forall i \in \{1, 2, \ldots, 25\} : s_{\text{tanh}}(i) = \tanh\left[3 s^2_{\text{coord}}(r_{ij})\right]. \tag{7.7}$$

The net charge of the set of QM water molecules is then obtained as the sum of all hyperbolic tangent values from which the number of hydrogen atoms in the neutral state of the water molecules is subtracted (here, 24),

$$q_{\text{net}} = \sum_{i=1}^{25} s_{\text{tanh}}(i) - 24. \tag{7.8}$$

This results in $q_{\text{net}} = 0$ for all-neutral QM water, $q_{\text{net}} = -1$ for a proton hole (OH^-) and $q_{\text{net}} = +1$ for a hydronium ion (H_3O^+). To assign a single mean net charge value to each relevant combinations of CV values, the mean values of q_{net} were calculated on a 3D grid for every triplet of ζ, dR82 and dPRG values occurring in the simulations. Specifically, each snapshot of the MW metadynamics trajectories was used with a bin width of 0.02 for ζ and 0.025 nm for both dR82 and dPRG.

7.2.4. Comparison of different DFTB3 QM/MM approaches

Several different combinations of the following settings were used in the current work:

- the general-use DFTB3 parametrization 3OB, or the modified parameter set 3OBw designed to reproduce the structure of liquid water
- the electrostatic interaction between the QM and MM was optionally scaled down by the factor of 0.75, corresponding to the inverse square root of the optical dielectric constant; the purpose of this treatment is to compensate for the missing electronic polarization of the MM region (which, if not compensated, may lead to overestimated QM–MM interaction)

It is somewhat difficult to judge which of the four possible combinations of these choices is the most appropriate. Therefore, pilot QM/MM DFTB simulations were performed using all of them, investigating the proton transfer reaction between the E194 and E204 sidechains constituting the proton release group (PRG).

Multiple walker well-tempered QM/MM metadynamics simulations of the PT between the PRG members, E194 and E204 were performed. The QM region comprised the sidechain of E204, the complete residues E194 and S193, the backbone carbonyl group of G192, as

well as 7 water molecules in the vicinity of the PRG; link atoms were placed along the bonds Cα–Cβ in E204, Cα–C in E194, and C–Cα in G192. The leap-frog integrator was used with a time step of 0.5 fs, and the temperature of 300 K and the pressure of 1 bar were maintained by the Nosé–Hoover thermostat and the Parrinello-Rahman barostat, respectively. The mCEC coordinate was used to drive the reaction with the Cδ atom of E194 considered as the initial donor, and Cδ of E194 as the final acceptor. Thus, the values of $\zeta = -0.46$ and $\zeta = +0.46$ correspond the to protonated carboxyl group of E194 and E204, respectively. To prevent the proton from being driven closer towards the respective Cδ atoms, restrains were applied to $\zeta < -0.5$ and $\zeta > +0.5$. The distance between the Cδ atoms was also restrained to values below 6.5 Å. Water molecules were kept near the PRG by means of spherical restraints to their initial positions with a diameter of 2 Å. The simulation consisted of 16 walkers; each was simulated for 175 ps resulting in a total sampling of 2.8 ns. Gaussian potentials with an initial height of 0.6 kJ/mol were deposited every 0.5 ps, and a bias factor of 10 was applied.

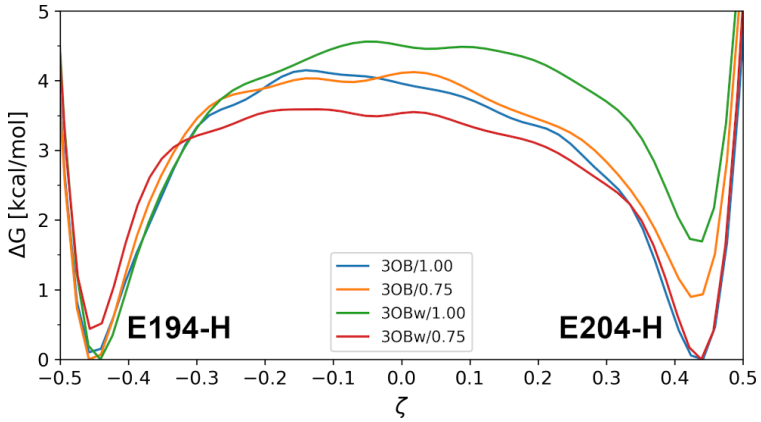

Figure 7.2.: Free energies of PT reaction in the proton release group obtained with the different DFTB3/3OB parametrizations and optional scaling of QM–MM electrostatics.

The resulting free energy curves are presented in Fig. 7.2. Notably, classical equations of motion were integrated in the MD simulations, therefore no quantum effects were considered, not even zero-point effects, which are quite massive for proton transfer. That is why there are energy minima corresponding to the proton located on E194 and on E204, rather than a broad minimum corresponding to the proton shared by the glutamates, which would be favored by the zero-point effects. Notwithstanding, rather than investigating the PRG itself, the purpose of these simulations is to compare the results obtained with the different setups.

All of the four setups applied provide a free energy profile with two minima corresponding to each of the glutamates being protonated separately from the other. With the exception of the DFTB3/3OBw simulation without QM–MM scaling, which favors the protonation of E194, the reaction is nearly isoenergetic. The height of the barrier opposing the proton transfer is within a rather narrow range of 4.0 ± 0.5 kcal/mol. Considering the polarized character of the lower active site of bR, a scaling of QM–MM electrostatics might be expected to induce great changes. This is not the case, however, and the free energy profiles obtained with the different QM–MM setups are qualitatively the same.

7.3. Results

7.3.1. O and O* states feature elevated internal hydration levels relative to the ground state

The appearance of a new C=O stretch band in time-resolved Fourier transform infrared (TR-FTIR) spectra hinted at the transient protonation of D212 during the **O**→**bR** reaction,[101,102] constituting an intermediate **O***. Starting with the ground state crystal structure, we modify the protonation states of D85, D212 and the PRG to construct models for the **O** and **O*** states; extensive classical MD simulations using Hamiltonian replica exchange are then used to establish reliable structural models for the **O** and **O*** states in an explicit lipid membrane environment.

D85 and D212 are deprotonated in the ground state (Fig. 7.3A); to help stabilize the negative charges in the protein interior, R82 prefers an upward orientation, and is aided by the Schiff base and a cluster of water molecules resolved in crystal structures.[73] Moreover, the PRG features an excess proton stored via a strong hydrogen bond between E194 and E204. A small number of water molecules separate the PRG and R82, whose upward orientation prevents the formation of a continuous water wire between the PRG and D85/D212, an essential feature that helps prevent the backflow of the proton from the PRG to D85.[136]

In the **O** and **O*** states, the PRG is deprotonated, while D85 and D212 is protonated, respectively. These changes in the protonation pattern are coupled with several major structural rearrangements compared to the ground state. First, the negatively charged E194 and E204 sidechains become separated (see Fig. 7.3C for distance histograms for the Cδ atoms of E194 and E204); in particular, E194 rotates away to form hydrogen-bonding interaction with Y83. Moreover, R82 rotates to a downward orientation to better stabilize the negatively charged PRG through a salt-bridge interaction with E204; the shift in the preferred R82 orientation is supported by distance histograms from unbiased MD simulations (Fig. 7.3C) as well as explicit umbrella sampling simulations using the A44N–R82Cζ distance as the collective variable. Interaction of the R134 sidechain with E194 was also analyzed because of its previously proposed effect of R134 on the acidity of E194[137] cf. section A in Appendix; no significant coupling with PT processes was found.

Separation of the E194/E204 sidechains in the **O** and **O*** states leads to elevated hydration level in the protein cavity (Fig. 7.3A); as shown in Fig. 7.3C, the number of water molecules in the central part of the cavity increased from ∼9 in the ground state to ∼12–13 in the **O** and **O*** states. Moreover, the downward rotation of the R82 sidechain leaves enough space in the cavity to form a continuous water network that spans from the PRG to D85/D212, setting the stage for proton exchange between these distant regions.

7.3.2. Competing proton transfer pathways for the O to O* transition

Since the **O*** state was captured in TR-FTIR studies[101,102], we first analyze the **O**→**O*** transition, which involves PT from D85 to D212. Upon QM/MM equilibration of the **O** and **O*** states, there are minor changes in the configurations of water molecules near D85/D212

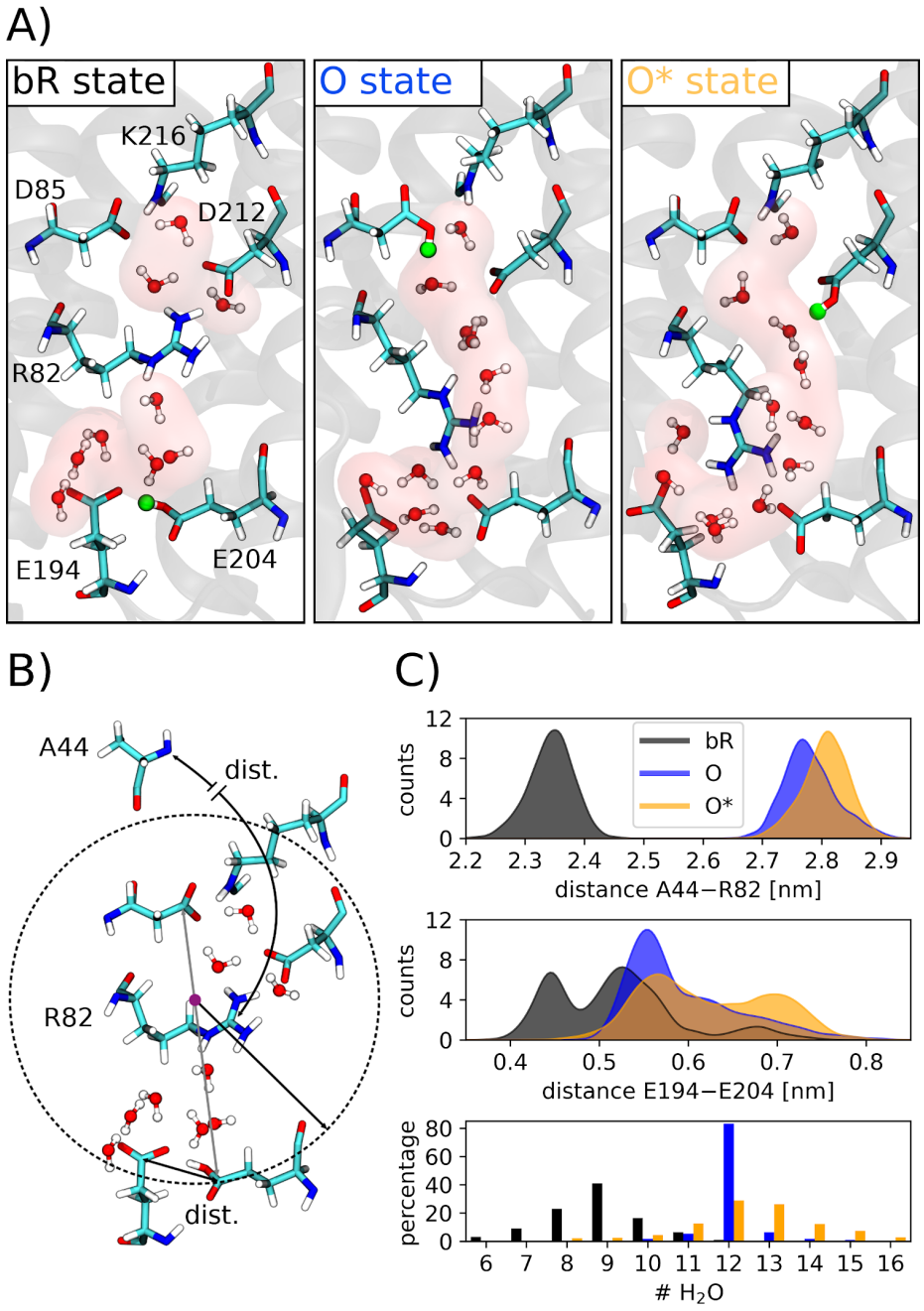

Figure 7.3.: A: Structures of the active site for the states **bR**, **O** and **O*** obtained from HREX simulations; green spheres: proton bound to an Asp or Glu; water density is emphasized by pink shading; retinal moiety of K216 is not shown for clarity. B: Illustration of the collective variables used to analyze HREX simulations and to run metadynamics simulations. C: Histograms of the swing movement of the sidechain of R82, of the PRG distance and of the number of water molecules within 10 Å of the center of the protein cavity from the HREX simulations of **bR** (black), **O** (blue) and **O*** (yellow) states. (The point on the line E204–D85, dividing that distance in the ratio of 1:2, is considered as the center of cavity). Histograms of the swing movement and of the PRG distance were generated from the final 9 ns, while the histograms of water distributions originate from the final 7 ns of HREX simulations.

as compared to classical MD simulations. Nevertheless, these water molecules modulate the separation of the D85/D212 sidechains and therefore potentially impact the PT mechanism and energetics. Accordingly, three-dimensional metadynamics simulations are used to probe the transfer mechanism in detail; the collective variables are (cf. subsection 7.2.2): ζ that describes the progress of PT, the D85Cγ–D212Cγ distance, and the average oxygen coordination number $\overline{s_O(r_{ij})}$ of the four QM water molecules.

The free energy surface (FES) in 2D representations is plotted with ζ as one coordinate (horizontal in the plots) and either the distance between D85Cγ and D212Cγ (Fig. 7.4A) or the net charge of the QM waters (computed based on oxygen coordination number $\overline{s_O(r_{ij})}$ of the QM waters, Fig. 7.4B) as the other coordinate (vertical in the plots). Three representative transition state structures are shown in Fig. 7.4C.

In Fig. 7.4A, there are two low-energy basins, one for the reactant state at $\zeta = -0.58$ and the other for the product state at $\zeta = 0.58$. In terms of energetics, the reactant **O** is the global minimum on the FES, and the product lies $\Delta G_{O\rightarrow O^*} = 3.6$ kcal/mol higher. This means the proton is better stabilized on D85 than on D212, and the PT from D85 to D212 and thus the **O**→**O*** transition is endergonic. The minimum free energy path connecting both minima leads through a transition state, which lies at $\Delta G^{\ddagger}_{O\rightarrow O^*} = 7.8$ kcal/mol. A transfer via this path is accompanied by a strong reduction of the D85Cγ–D212Cγ distance, so that the approximate TS ($\zeta = 0.0$) has the aspartates 0.42 nm apart; as illustrated in Fig. 7.4C (black dot), this corresponds to a direct proton exchange between D85 and D212 sidechains without the explicit involvement of any water molecule.

In Fig. 7.4B, plotting the FES along ζ and the net charge of QM water molecules illustrates two alternative PT pathways, which involve a hydronium (Fig. 7.4C, blue dot) and hydroxide (Fig. 7.4C, red dot), respectively, in the transition state; the corresponding barriers are ~13.2 and 12.1 kcal/mol, respectively, thus the direct proton exchange is the dominant mechanism. Apparently, the penalty of rearranging the hydrogen bonding network is more than compensated for by positioning the proton donor/acceptor groups next to each other to lower the intrinsic PT barrier. These observations highlight the importance of allowing water molecules to fully equilibrate during the PT[138]; they also caution against inferring the dominant PT mechanism based solely on the water structure prior to the reaction.[139]

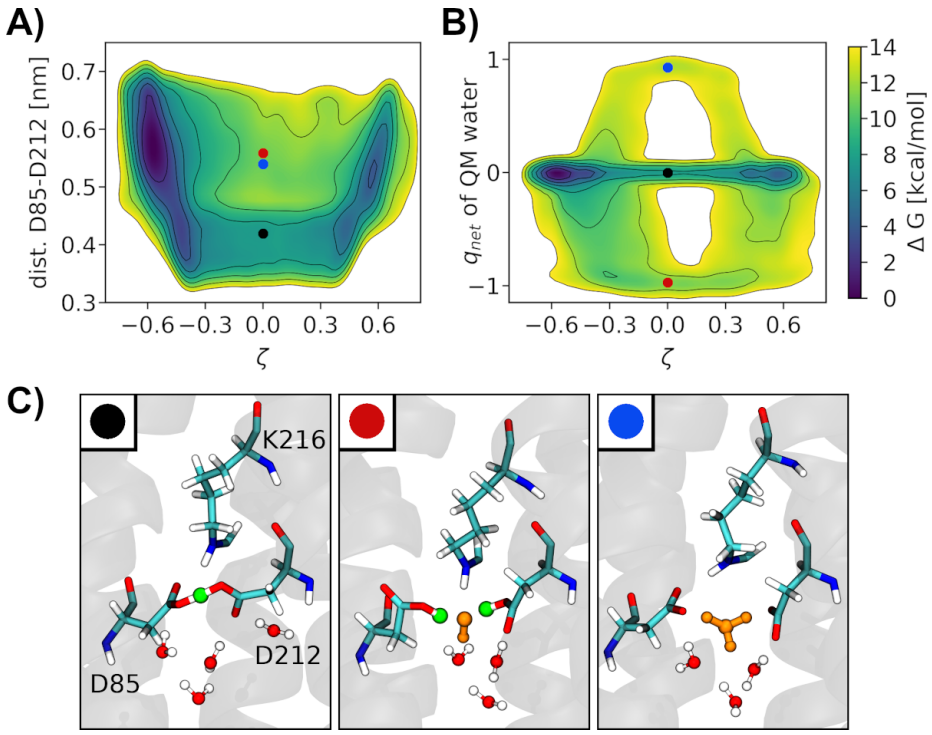

Figure 7.4.: A: Computed free energy surface for the **O**→**O*** transition as a function of the D85Cγ–D212Cγ distance and the PT reaction coordinate ζ. B: The same surface as a function of the net charge on the four QM water molecules q_{net} and ζ. There are three distinct pathways, which differ in the barrier heights. C: The geometries of the upper active site found in the transition structures marked by colored dots. Black – minimum free energy path, direct proton transfer; red – proton hole transfer (OH^-); blue – excess proton transfer (H_3O^+); green spheres: protons bound to an Asp or Glu; orange molecules: charged water species (OH^- or H_3O^+)

7.3.3. O to bR transition occurs through a proton hole mechanism

Since the **O** to **bR** transition involves not only proton exchange between D85 and the PRG, but also reorientation of the R82 sidechain as well as separation of E194/E204 in the PRG, three-dimensional metadynamics simulations are required to fully elucidate the underlying mechanism. The three collective variables are (cf. subsection 7.2.3): ζ that describes the progress of the proton exchange, the A44N–R82Cζ distance (dR82) and the E194δ–E204Cδ distance (dPRG). The resulting 3D FES following ~100 ns of QM/MM metadynamics simulations is shown in Fig. 7.5A. To monitor the mechanism of the proton exchange, we also plot the mean net charge of QM water molecules for snapshots taken from the metadynamics simulations in Fig. 7.5B; a net charge was obtained from the number of hydrogen atoms bonded to each oxygen. Also, care was taken to count each QM hydrogen only once to avoid artifacts associated with water molecules explicitly engaged in proton or proton hole exchange with neighboring groups. Representative snapshots are shown in Fig. 7.5C.

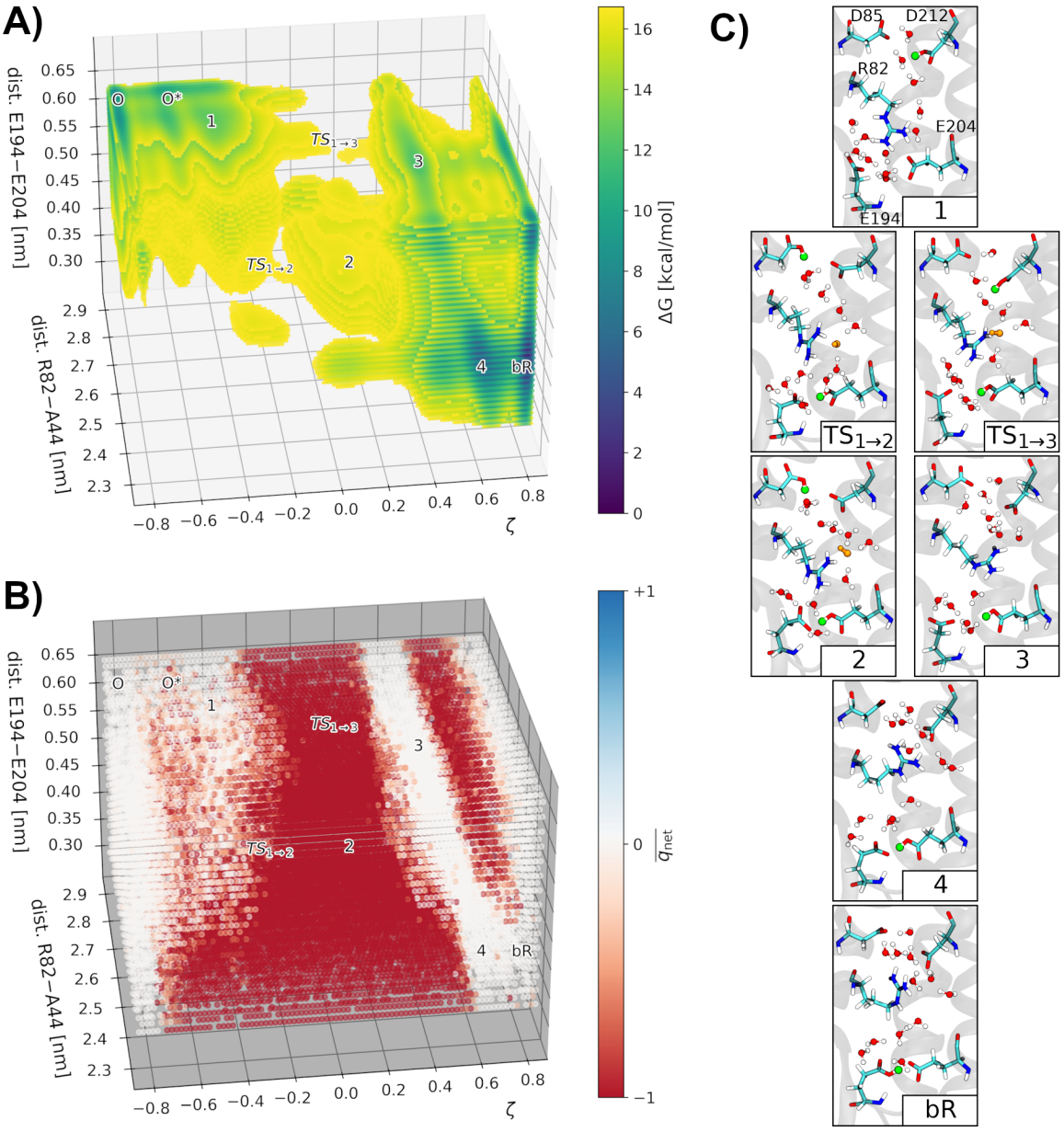

Figure 7.5.: A: The 3D free energy profile of the complete **O**→**bR** transition. The relevant states are labeled. Two pathways with similar barrier heights can be identified: **O** → **O*** → **1** → $\mathbf{TS_{1\to2}}$ → **2** → **4** → **bR** and **O** → **O*** → **1** → $\mathbf{TS_{1\to3}}$ → **3** → **4** → **bR**. B: Net charge of QM water molecules. White: $\overline{q_{net}} = 0$, neutral water; red: $\overline{q_{net}} = -1$, OH^-; blue: $\overline{q_{net}} = +1$, H_3O^+. C: Representative structures corresponding to the states identified with metadynamics. Green: proton bound to an Asp or Glu; orange: charged water species (here, always OH^-).

The local minimum corresponding to the **O** state has a free energy of ΔG = 3.6 kcal/mol relative to the **bR** ground state (product of the proton exchange reaction studied in this subsection). Its value of $\zeta = -0.82$ corresponds to one of the two oxygen atoms of D85 being protonated. With an A44–R82 distance of 2.87 nm, R82 is swung down to stabilize the two negative charges of deprotonated E194 and E204, which maintain a long Cδ–Cδ distance of 0.62 nm, with a water molecule in between.

The first step towards the **bR** ground state is the **O**→**O*** transition, as described in the previous subsection. In this step, the proton is transferred from D85 to D212 in an endergonic process as the minimum corresponding to the **O*** state lies 2.5 kcal/mol higher in free energy than the **O** state. This PT event proceeds over a barrier of $\Delta G^{\ddagger}$ = 5.7 kcal/mol, either directly or via a proton hole mechanism. The barrier height obtained here is 2.1 kcal/mol lower than that found for the **O**→**O*** transition discussed above using a smaller QM region. This modest discrepancy provides an evaluation of sensitivity of our QM/MM free energy simulations to differences in the simulation setup, such as the QM region size and the scaling of QM–MM electrostatics applied or not (cf. subsection 7.2.4).

Subsequently, the R82 sidechain moves slightly upwards to an A44–R82 distance of 2.75 nm while dPRG remains at 0.62 nm. This conformation corresponds to a flat metastable state denoted as **1** with $\zeta = -0.40$, which lies 2.8 kcal/mol higher in energy than **O***, to which it is still highly similar in structure. Starting from state **1**, the subsequent proton exchanges may follow one of two different pathways, which have comparable energetics according to the computed FES.

Path 1 proceeds via $\mathbf{TS_{1\to 2}} \to \mathbf{2} \to \mathbf{4} \to \mathbf{bR}$. Here, E204 is protonated by a water molecule resulting in a proton hole, which is stabilized by the positively charged guanidinium group of R82. Afterwards, the E194–E204 distance decreases, suggesting that E204 and E194 form a hydrogen bond. In addition, D85 is reprotonated by D212, and the R82 sidechain moves towards D85 gradually, accompanying the proton hole apparently. In the transition state $\mathbf{TS_{1\to 2}}$, a barrier of $\Delta G^{\ddagger}$ = 7.7 kcal/mol w.r.t. state **1** is passed, before a broad, flat metastable state **2** is reached leading to a free energy gain of $\Delta G = -2.8$ kcal/mol w.r.t. $\mathbf{TS_{1\to 2}}$. In **2**, the water wire is disrupted, and the proton hole is closer to D212 and stabilized by the guanidinium group of R82. Then, a low barrier of $\Delta G^{\ddagger}$ = 1.0 kcal/mol has to be overcome to move the proton hole further in the direction of D85, while R82 advances towards D212 simultaneously. Eventually, D85 is deprotonated by the proton hole, and the positively charged sidechain of R82 stabilizes the two negative charges of D85 and D212 at dR82 = 2.33 nm. This results in a large free energy gain of $\Delta G = -10.6$ kcal/mol in state **4**. In **4**, E204 is protonated and hydrogen-bonded to E194. Finally, after the final PT from E204 to E194 over a barrier of $\Delta G^{\ddagger}$= 2.9 kcal/mol, the system reaches the global minimum, the **bR** ground state, which we set to $\Delta G = 0$. In **bR**, $\zeta = 0.82$ with a small E194Cδ–E204Cδ distance of 0.41 nm, corresponding to a strong hydrogen bond between E194 and E204.

Path 2 starts with the protonation of E204 from a water molecule as well, and it proceeds through states $\mathbf{TS_{1\to 3}} \to \mathbf{3} \to \mathbf{4} \to \mathbf{bR}$. D85 is not reprotonated in this pathway, and the distance E194Cδ–E204Cδ remains large as the proton hole moves towards the protonated D212. First, the sidechain of R82 swings towards D212, being accompanied by the proton hole. The transition state $\mathbf{TS_{1\to 3}}$ poses a barrier of $\Delta G^{\ddagger}$ = 7.8 kcal/mol w.r.t. state **1**. Note that another, minor protonation pattern may occur in this region of the 3D space spanned

by the applied reaction coordinates – D85 and E204 protonated (but not D212) with an OH^- proton hole located further down, closer to the PRG; this observation is detailed in subsection 7.3.4. Once this barrier is passed, D212 is deprotonated by the proton hole, and R82 approaches D212. This results in the local minimum **3**, which lies −8.9 kcal/mol beneath $\mathbf{TS_{1\rightarrow 3}}$. Next, the protonated E204 and the negatively charged E194 move closer together, while R82 swings closer to D85 and D212, leading to the local minimum **4** and, finally, the global minimum **bR** following the final PT from E204 to E194, as described above.

The current simulation setup does not cover the possibility of deprotonation of the R82 sidechain and thus participation of R82 as a PT relay. Although an arginine might deprotonate in principle, a deprotonation of R82 in bR would lead to largely increased energy barriers opposing the long-range PT; a detailed mechanistic analysis is presented, cf. subsection 7.3.5. Such a pathway is therefore unfavorable. Also, with the current simulation setup, E194 is the final acceptor of the excess proton, and is also hydrogen-bonded to E204. The protonated E194 is favored over the protonated E204, and a proton exchange between them is opposed by a modest barrier of 2.9 kcal/mol. A recent study by Tripathi et al. found the excess proton centered between E204 and E194, resulting in an extremely flat and symmetric free energy minimum along the proton exchange reaction coordinate[75], similar to our previous QM/MM studies.[140,141] Our current simulations focus on resolving the long-range PT mechanism rather than elucidating the precise molecular features of the PRG, which requires more systematic variations of the QM region, the QM level (e.g., DFTB3 vs. DFT), the local level of hydration and explicit treatment of nuclear quantum effects.

Notwithstanding, the overall **O**→**bR** transition is computed to be an exothermic process with a driving force of −3.6 kcal/mol. This value can be compared to the pK_a values of D85 and of the PRG estimated from experimental studies. To this end, Balashov et al. calculated an equilibrium constant of $10^{-2.2}$ for proton exchange between D85 and the PRG in the ground state.[96] Under the assumption that the measured values for the ground state sufficiently resemble the **O**→**bR** transition, the estimated $\Delta G_{\mathrm{O}\rightarrow\mathrm{bR}} = -3.0$ kcal/mol. The value of $\Delta G_{\mathrm{O}\rightarrow\mathrm{bR}} = -3.6$ kcal/mol obtained from our simulations agrees well with this estimate.

The involvement of the proton hole during the **O** to **bR** transition is clearly illustrated by the the mean net charge of QM water molecules shown in Fig. 7.5B. Naturally, $\overline{q_{\mathrm{net}}} = 0$ in the **O** state, which means that all water molecules are in their neutral form. The **O**→**O***→**1** transition and the **O*** state itself exhibit $\overline{q_{\mathrm{net}}} = 0$ mostly, corresponding to a direct PT, with a trace of $\overline{q_{\mathrm{net}}} < 0$ that may be attributed to a minor proton hole pathway; these features are consistent with the study of **O**→**O*** transition using a smaller QM region discussed in the last subsection, further supporting the robustness of our computational methodology. The mechanism of the subsequent **1**→**bR** transition is also clear: both possible pathways passing through $\mathbf{TS_{1\rightarrow 2}}$ and $\mathbf{TS_{1\rightarrow 3}}$ proceed via a proton hole with D85/E204 or D212/E204 protonated, cf. Fig. 7.5C. After the deprotonation of D85 or D212, the proton hole vanishes, and all QM water molecules are present in the charge-neutral form (in the states **3** and **4**). The final transition **4**→**bR** proceeds as a direct proton exchange between E194 and E204, as corroborated by the small E204–E194 distance of ca. 0.41 nm in the relevant region of the FES.

7.3.4. Co-existence of multiple protonation patterns.

It is of interest to monitor how the protonation states of D85, D212, E204 and E194 change with the varying reaction coordinates ζ, dR82 and dPRG. To this end, a 3D grid in the reaction coordinates was created, and the occurrences of all possible combinations of protonation states were counted in every bin on the grid; the results are provided in the attached spreadsheet. As an example, the bin containing the state $\mathbf{TS_{1\to 2}}$ ($-0.1 < \zeta < 0.0$, $2.7 <$ dR82 < 2.8 nm, $0.35 <$ dPRG < 0.45 nm) exhibits the frequencies of the different protonation patterns shown in Tab. 7.2. The protonation of D85 and E204 is favored, and the long-range PT proceeds via a configuration with D85 and E204 protonated.

Table 7.2.: The frequencies of the different protonation states patterns observed in the transition states.

protonated residues	number of occurrences in $\mathbf{TS_{1\to 2}}$	in $\mathbf{TS_{1\to 3}}$
D85	1	0
D85 & E204	171	39
D85 & E194	9	9
D212 & E204	44	45
D212 & E194	2	0

In some combinations of values of reaction coordinates located between the states **1** and **4**, however, there are multiple, equally significant protonation patterns. For instance, the bin containing the state $\mathbf{TS_{1\to 3}}$ ($0.1 < \zeta < 0.2$, $2.6 <$ dR82 < 2.7 nm, $0.55 <$ dPRG $<$ 0.65 nm) represents in fact a superposition of two protonation patterns, one with D85 and D212 protonated and the other with D212 and E204 protonated.

This makes it difficult to state that a certain pathway exhibits a distinctive protonation pattern. The reason for this overlap is that the ζ coordinate cannot distinguish between these patterns, as it only depends on the distances of the center of excess charge (OH^-) from the initial proton donor (D85) and that from the final proton acceptor (E194). If a larger number of protonation patterns are possible, as is the case in the current study, then some ζ values can be realized with several different protonation patterns accompanied by a suitable location of the excess charge. This can be exemplified by looking at $\mathbf{TS_{1\to 2}}$ again: The proton hole is closer to D212 (and therefore further away from E204) when D85 and E204 are protonated compared to structures in which D212 and E204 are protonated.

From chemical intuition, one would assume that the $\mathbf{O^*\to bR}$ transition proceeds solely via the state with protonated D212 and E204, because the proton hole may move towards D212 for reprotonation easily. By contrast, the proton hole has to pass the negatively charged D212 in order to arrive at D85 for reprotonation, and the electrostatic repulsion should render that process unfavorable. However, the bR mutant D212N is still capable of pumping protons (with a reduced activity)[142], even though N212 cannot be reprotonated, and so the proton hole has to pass by this residue, and the configuration with D85 and E204 protonated is the only possible in that mutant. Therefore, it is less of a surprise to observe this protonation pattern in bR.

The co-existence of multiple protonation patterns is probably not easy to avoid, and also it makes the analysis of simulations somewhat less transparent. Still, it does not seem to pose a principal problem of the simulation setup as it does not put the convergence of metadynamics (or the like) into jeopardy. The most serious adverse effect on the free energies may potentially be slightly underestimated energy barriers, as a consequence of contribution of additional protonation patterns to the histogram counts in the transition states.

7.3.5. Why Arg82 is not considered as a proton relay.

There is a question if the side chain of R82 participates in the long-range PT actively as proton relay or not. Judging by the pK_a values of arginine sidechain and of water, which take values of 12.5–13.8 and 14.0, respectively, the state R82°/H_2O would be favored over $R82H^+/OH^-$ in the limit of infinitely separated reactants in pure water environment. This reasoning would lead to the expectation that R82 participates in the PT as soon as a (transition) state involving a hydroxide ion is formed. However, it is expected that a closely interacting guanidinium–hydroxide pair in protein environment behaves differently, and also, de- and re-protonation of R82 may be hindered for other reasons. In fact, several points speak against the participation of R82, as described below.

It was found previously that bR mutants with a non-titratable residue in place of R82 (R82A and R82Q) retain a proton pumping activity.[109] This is only possible as long as R82 itself does not relay protons. The reduced efficiency of proton pumping reported for the mutants was attributed to the missing electrostatic stabilization of the deprotonated state of D85 by the positive charge of the R82 side chain.

Phatak et al. performed a QM/MM simulation study of the PT path from D85 to the PRG through the R82 sidechain.[107] They divided the whole process into several PT steps: (1) from R82 to E204 to form a neutral R82°; (2) between the amino groups of the guanidinium group of R82 via a water molecule; (3) from the primary donor D85 to D212 via water molecules w401 and w406, to form a transient state with D85 deprotonated and D212 protonated (which is not the **O** state because R82 is deprotonated); and finally (4) from D212 to the neutral R82 via a water molecule to form the ground state. The minimum energy pathway and the associated energy profile were obtained. The step (2) gives rise to the rate-limiting energy barrier of 36 kcal/mol, much too high to comply with the **O**→**bR** transition taking place in milliseconds. That step might be unnecessary if there is a pathway that involves de- and re-protonation of the same amino group of the R82 sidechain.[103] Even then, the energy barrier of ca. 30 kcal/mol posed by the remaining processes is still prohibitively high.

Lastly, we demonstrate that pathways involving deprotonated R82° and a hydronium ion between the transition state $\mathbf{TS_{1\to3}}$ and the **bR** ground state are energetically unfavorable. $\mathbf{TS_{1\to3}}$ is the highest-energy state of bR found in our study, see Fig. 4C in the main text for the structure. Let us investigate the free energy along a hypothetic alternative pathway leading from this transition state to the ground state **bR**. Such a proton transfer would pass through the following states:

- In $\mathbf{TS_{1\to3}}$, D212 (or D85) and E204 are protonated, and R82 and OH^- are in the center of the cavity:
  $R82H^+$, OH^-, D212H, E204H.
- Deprotonation of the R82 sidechain by recombination with the proton hole would lead to a hypothetic state $\mathbf{HS_1}$:
  $R82^\circ$, H_2O, D212H, E204H.
- In order to approach the **bR** ground state, it is now necessary to transfer a proton from D212 to R82. This would proceed via the Grotthuss mechanism involving a hydronium ion located between $D212^-$ and $R82^\circ$ in a hypothetic state $\mathbf{HS_2}$:
  $R82^\circ$, H_3O^+, $D212^-$, E204H.
- As soon as the hydronium ion reaches and re-protonates R82, the ground state **bR** would be reached:
  $R82H^+$, H_2O, $D212^-$, E204H.

Let us estimate the free energy changes along this process using the available acidity constants of the participating species; the approximation here is that the acidities in proton environment equal those estimated in fully aqueous environment, see Fig. 7.6.

Deprotonation of $R82H^+$ yields the state $\mathbf{HS_1}$, 2 kcal/mol below $\mathbf{TS_{1\to3}}$. The consequent deprotonation of D212H, however, leads to the state $\mathbf{HS_2}$ that lies more than 3 or 5 kcal/mol above $\mathbf{TS_{1\to3}}$ (the particular value depends on the value of pK_a of arginine used), and thus it now represents the state of highest energy on the entire reaction pathway. This observation makes this pathway by a large margin less favorable than the process involving long-range proton hole transfer avoiding deprotonation of $R82H^+$. Finally, the reprotonation of $R82^\circ$ yields the final product at −14 kcal/mol w.r.t. the transition state $\mathbf{TS_{1\to3}}$. This is a rough estimate based on aqueous-solution pK_a values, in a modest agreement with the value of −17 kcal/mol resulting from our simulations as presented in the main text. In summary, the high energy of the state with deprotonated R82 and D212, featuring a hydronium ion, makes the pathway that involves the R82 sidechain as a relay highly unfavorable.

All of this combined leads to the conclusion that the deprotonation of R82 is a dead end, energetically speaking, and rather shows that there is a strongly interacting ion pair $R82H^+/OH^-$, so that the proton hole is in fact stabilized by the presence of the positively charged R82 sidechain.

7.3.6. Overall rate of the PT reaction from transition state theory

The rate of the formation of the proton hole during the O*→bR transition was estimated with the transition state theory[144] as

$$k = \frac{k_B T}{h} \cdot \exp\left[-\frac{\Delta G^\ddagger}{k_B T}\right], \tag{7.9}$$

where k_B is the Boltzmann constant, T is the temperature, h is Planck's constant, and $\Delta G^\ddagger$ is the free energy barrier. With $\Delta G^\ddagger_{O\to bR}$ of the rate-limiting step at $T = 300$ K, the reaction rate is estimated to $k = 1800\ s^{-1}$.

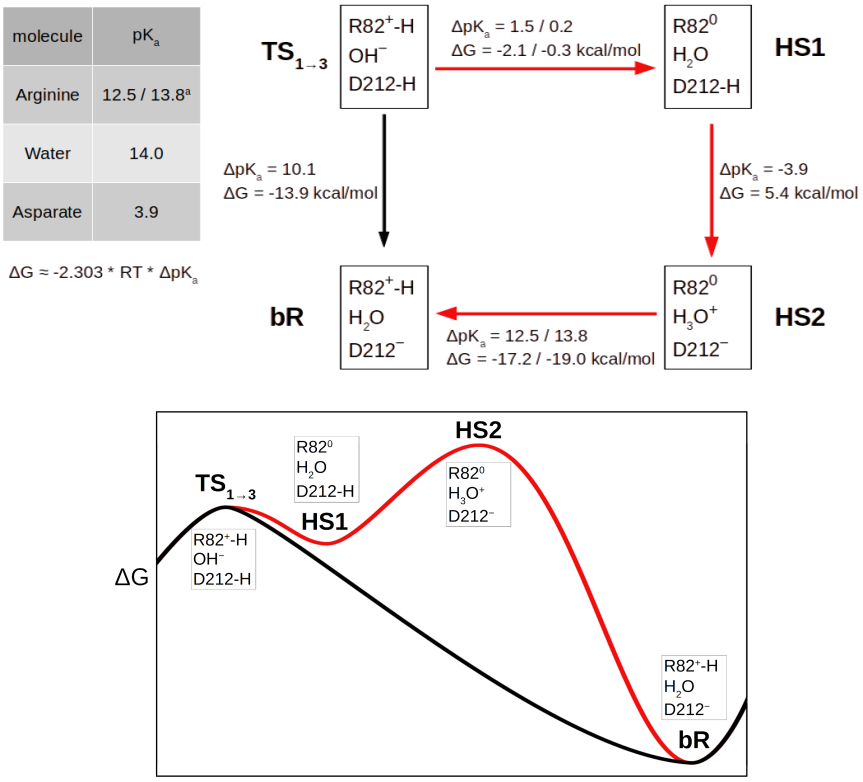

Figure 7.6.: Energetics of a hypothetical transition from the transition state $\mathbf{TS_{1\to3}}$ to **bR** ground state proceeding via intermediates involving a deprotonated R82° sidechain (red) compared to the pathway presented in the main text (black). [a]The value of pK_a =13.8 for arginine is taken from Ref. [143]; the value of 12.5 appears in commonly available textbooks.

While quantum tunneling is not relevant for a process passing over a broad barrier like the PT reaction investigated, the zero-point effects should still be included; a typical vibrational frequency of 3500 cm^{-1} for bonds involving a hydrogen atom leads to a zero-point vibrational energy ZPVE = $\frac{h\nu}{2k_B T} \approx 5$ kcal/mol at $T = 300$ K.

Energy barriers to proton transfer reactions were considered during the benchmarking of the DFTB3/3OB method.[135] For both excess proton transfer and proton hole transfer, the energy barrier is underestimated by ca. 1 kcal/mol at the distance of 2.8 Å between the donor and acceptor oxygen atoms, and this deviation increases with increasing distance between the donor and the acceptor; notably, the performance of DFTB3/3OB was very similar to that of B3LYP. A benchmark of proton transfer barriers confirmed that B3LYP underestimates barriers to proton transfer reactions, whenever the transfer proceeds via bridging water molecules[145]; the error was 0.3, 1.4 and 1.7 kcal/mol for zero, one and two bridging waters, respectively. As DFTB3/3OB closely follows the performance of B3LYP for energy barriers of simple proton transfer reactions, we can assume that the B3LYP trend of increased error is the case with DFTB3/3OB also.

Altogether, DFTB3/3OB underestimates energy barriers to proton transfer reactions; the error amounts to over 1 kcal/mol for intermediate-to-long distances between the donor

and the acceptor, and it further increases by 1–2 kcal/mol if the transfer proceeds via bridging water molecules. In both aspects, the error increases with increasing distance, so that the overall error for a long-range PT reaction like that in the O to bR transition amounts to at least 3 kcal/mol, and is larger most probably.

At the same time, the calculation of reaction rate with Eq. 7.9 misses the ZPVE of 5 kcal/mol. It turns out that such a calculation using the energy barrier obtained with DFTB3/3OB exhibits an almost complete cancellation of errors. In a worst-case scenario, if the error in barrier is $3\,k_{\mathrm{B}}T = 1.8$ kcal/mol smaller than the ZPVE, the calculated reaction rate would be underestimated by a factor of 20. This is the largest expected systematic error in the rate, and is acceptable considering that the purpose of this calculation is merely to make a rough comparison with experimental observations.

In the context of proton transfer rate, it may be of interest to discuss the kinetic isotope effect (KIE) accompanying the O→bR transition, which is rather small with a value of 3.5.[99] That may be rationalized by several generally valid arguments[146,147]: First, the highest (rate-limiting) barrier for this long-range reaction is rather broad, and in such a case, the role of tunneling is not critical, and the KIE is largely due to the ZPVE. Also, KIE is large only in cases where the donor/acceptor atoms are relatively rigid, so that nuclear wavefunction overlap is small and sensitive to the H/D substitution. In our case, the donor/acceptors in the highest barrier region correspond to mobile atoms – oxygen atoms of water – so a large overlap and thus small KIE is expected, consistent with the experimental value.

7.4. Concluding Discussion

Despite more than 45 years of studies, the mechanistic understanding of bacteriorhodopsin remains incomplete due largely to the poor characterization of the **O** state and its conversion back to the ground **bR** state. By combining extensive classical and QM/MM simulations, we are able to fill these important voids in the photocycle of this prototypical proton pump. Multi-dimensional QM/MM metadynamics simulations enabled us to explicitly probe the mechanism and energetics of the proton exchange process that underlies the **O** to **bR** transition. The energetics and kinetics from the free energy simulations are summarized schematically in Fig. 7.7. The rate-limiting barrier corresponds to a rate constant of 1.8×10^3 s^{-1} using transition state theory; this calculation of rate relies on a minor compensation of errors, thus no perfect agreement with reference data can be expected. Still, the resulting time scale of ~0.6 ms for the **O** to **bR** transition lies within the experimentally estimated range of 0.5–5 ms.[96–99] As noted above, the computed exergonicity of −3.6 kcal/mol is also consistent with the experimental estimate based on pK_a differences between D85 and the PRG. Therefore, our free energy calculations provide, for the first time, a mechanism that is consistent with experimental kinetic and thermodynamic data.

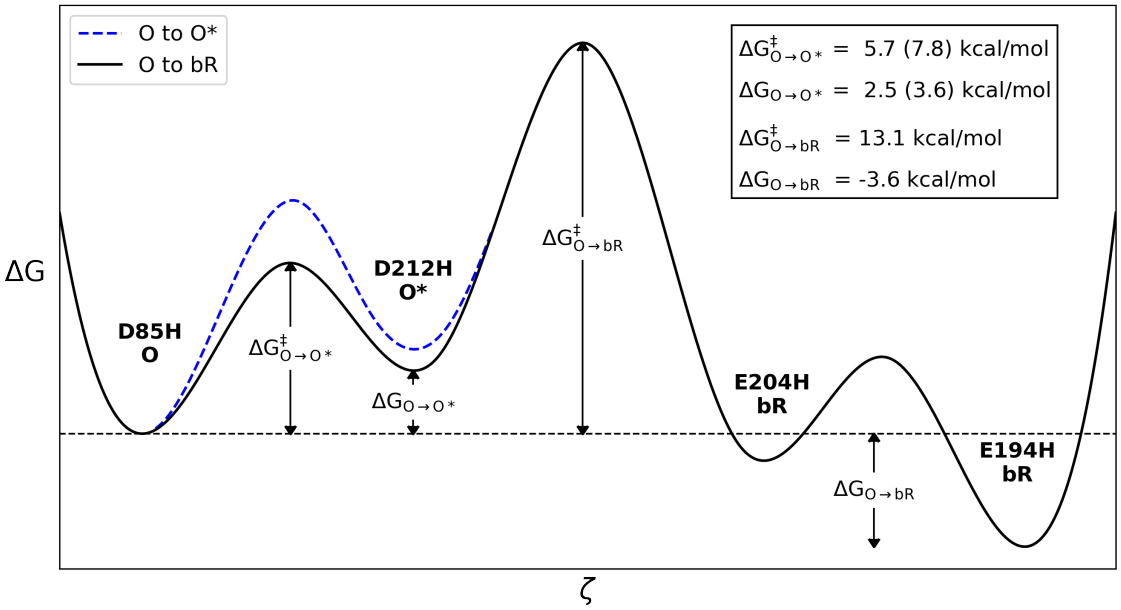

Figure 7.7.: One-dimensional Gibbs free energy profile for the **O**→**bR** transition, showing data obtained from the simulation of the complete **O**→**bR** process (black curve) and those of the first step, **O**→**O*** (blue curve). Free energy values in parentheses are from the simulation of **O**→**O*** with a smaller QM region.

The simulations also provided microscopic insights into the transition mechanism and structural motifs that play key roles during the process. We find that the changes of protonation states of buried amino acids (D85, D212 and the PRG) are coupled with

structural rearrangements of charged internal groups and water molecules. Specifically, the R82 sidechain swings between two distinct orientations to preferably stabilize the negative charges of D85/D212 and PRG in the **bR** and **O** states, respectively. The structural rearrangement of R82 together with that of the PRG also modulate the level of hydration in the protein interior, another feature essential to the long-range proton exchange between D85 and the PRG. In the **bR** state, the strong hydrogen bond between E194–E204 ensures a low level of hydration between D85/D212 and PRG, and the upward orientation of R82 prevents the formation of a continuous water wire; together these features constitute a "double-gate security" that prevents the wasteful proton back flow from the PRG to D85 in the **bR** state. By contrast, in the **O** state, the level of hydration between D85/D212 and PRG is elevated due to the breakup of the strong hydrogen bond E194–E204 upon proton release; moreover, reorientation of the R82 sidechain allows the formation of a continuous water network between the PRG and D85, again highlighting the regulatory role of R82. The positive charge of R82 is also apparently essential to facilitating the generation and transfer of the proton hole from the PRG to D85/D212.

Many of these mechanistic details are not limited to bacteriorhodopsin. For example, the importance of internal hydration level change coupled with protonation events has been discussed in several proton pumping systems such as cytochrome c oxidase[148–150] and Complex I[151], thus the observations here further support the general relevance of this electrostatics-hydration coupling mechanism. The gating function of the arginine sidechain was previously also observed in voltage sensitive ion channels by Armstrong et al.[152], who found that plug-like movement of an arginine sidechain controlled the passage of K^+ ions through the channel.

The observation of a proton hole mechanism in bacteriorhodopsin is somewhat unexpected considering that the system has been regarded as a prototypical proton pump, rather than a hydroxide pump. While our computational methodology has the flexibility of describing the hydronium mechanism as well, the current simulations do not provide an explicit comparison of the hydroxide and hydronium mediated pathways, thus the latter cannot be excluded definitely and deserves to be characterized in detail in future studies.

On the other hand, our observation provides a plausible explanation that the D85S mutant of bacteriorhodopsin was shown to be a Cl^- pump[153], and additionally, corroborates with the similarities of bacteriorhodopsin and halorhodopsin as pointed out previously[154]: Both proteins share a common pumping mechanism that relies on a conserved pattern of electrostatic interactions. Specifically, crucial acidic residues of bacteriorhodopsin are replaced by Cl^- binding sites in halorhodopsin. Also, the transition state observed in our current work involves an OH^- anion and protonated D85/D212, and this configuration is analogous to halorhodopsin as well as to the chloride-pumping mutants of bacteriorhodopsin, D85X (X=S,T,N).[155,156] All of this suggests that the system is poised to stabilize and allow the passage of a negatively charged ion such as hydroxide. Therefore, as discussed by two of us[110,111], it is important to consider all relevant titration states for water in the analysis of PT reactions, especially in highly heterogeneous environments such as the interior of a protein. Along this line, our discussion of competing PT pathways in the **O** to **O*** transition cautions against inferring the dominant PT mechanism based solely on the water structure prior to the reaction.[139]

Finally, from the technical point of view, the current study highlights the value of employing a calibrated semi-empirical QM method like DFTB3 in QM/MM free energy simulations. The balance of accuracy and efficiency makes it feasible to construct meaningful multi-dimensional free energy surfaces with ~100 ns of sampling with minimal *a priori* assumptions regarding the possible reaction mechanism (e.g., Grotthuss vs. proton-hole transfer). Therefore, there continues to be pressing need to further develop effective semi-empirical QM methods for increasingly complex biological applications.[157,158]

8. Histidine Kinase Rhodopsin

Author Contributions:
This work was done in cooperation with Franziska Wolff, who already published parts of this work in her doctoral thesis (Ref. [16]). Franziska Wolff built the homology model and simulated the Rh-Bl to P550 transition by isomerizing the retinal with the CASSCF method and subsequent 2 ns QM/MM MD equilibrations. Denis Maag simulated the P550 to Rh-UV transition by means of QM/MM MD simulations, excitation energy calculations and proton transfer simulations.

8.1. Introduction

Histidine kinase rhodopsin (HKR) exhibits two bistable state, Rh-Bl and Rh-UV, which are photoconvertable. During the photocycle, several intermediates are formed and a PT between the Schiff base and D239 occurs, compare Fig. 8.1

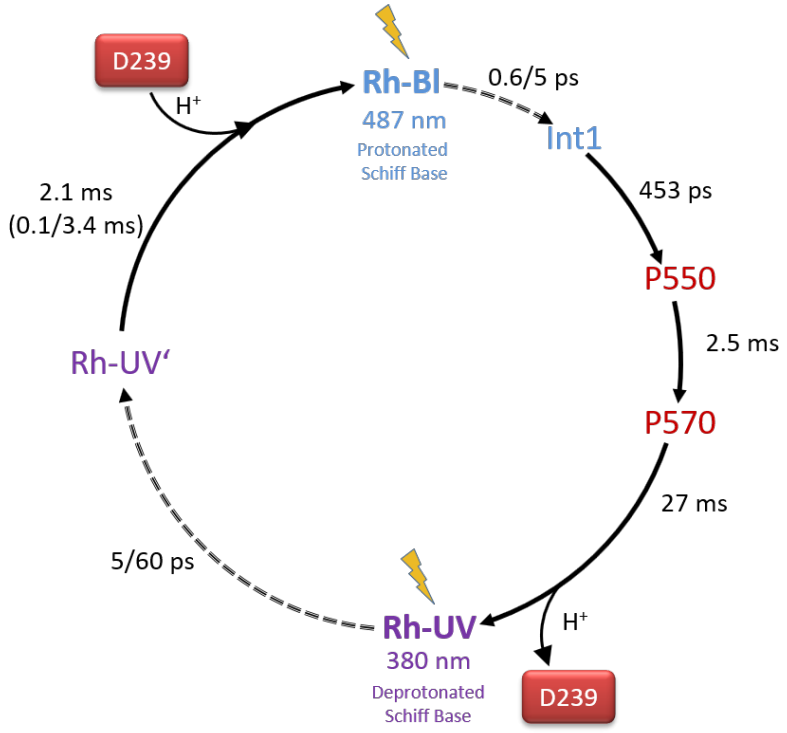

Figure 8.1.: Photocycle of HKR. In the Rh-Bl state, the retinal is in the 13-*cis*,15-*anti* configuration and isomerizes to 13-*cis*,15-*syn* after absorption of a photon. Figure generated by Franziska Wolff.

Since there is no crystal structure available, Franziska Wolff built several homology models of HKR.[16] Based on sequence similarities, the structures of six rhodopsins (PDB IDs in brackets) served as templates: Bacteriorhodopsin mutant M56A (1PXS)[159], Halorhodopsin (3VVK)[160], Sensory Rhodopsin (1XIO)[161], Sensory Rhodopsin II (2KSY)[162], Rhodopsin I (5AWZ)[163] and Channelrhodopsin C1C2 (3UG9)[164]. Each model was considered as a dimer, embedded in a POPC lipid bilayer membrane and solvated with water. Subsequently, MM and QM/MM equilibrations and QM/MM production simulations were performed. The QM region comprised the retinal including the Schiff base, D239 and a nearby water molecule which were described with DFTB and 3OB parameters.

To asses the quality of the models, the excitation energies were calculated with the OM2/MRCI method and compared to the excitation energy of the bR dark state which was also calculated with OM2/MRCI. The experimental shift between the HKR Rh-Bl state (487 nm; 2.55 eV) and the bR dark state (570 nm; 2.18 eV) is 0.37 eV, which was best reproduced by the homology model based on Channelrhodopsin C1C2 with a shift of 0.28 eV. For this model, the isomerization pathway from 13-*cis*,15-*anti* to 13-*cis*,15-*syn* was calculated by the CASSCF method. The final structure served as starting structure for five QM/MM molecular dynamics replicas with different initial velocities of the atoms, which were randomly assigned from the Maxwell distribution at 300 K.

This work aims to validate the homology model further by identifying the intermediates P550 and P570 as well as the stable Rh-UV state. We show that P550 is formed after 2 ns QM/MM MD simulations by calculating the red shift in energy compared to the starting structure. The transition to P570 is reached by heating the protein up to 450 K for short time intervals. Due to the high temperature, the configurational sampling is accelerated and a water wire is formed between the Schiff base and D239. In order to reach the Rh-UV state, the PT from the Schiff base to D239 along the water wire is simulated with umbrella sampling.

8.2. Computational Details

System setup and QM/MM simulations.
In Ref. [16], the HKR dimer was embedded in a bilayer of 274 POPC lipids and solvated with 11552 TIP3P water molecules. Each monomer consists of 216 residues. Here, all QM/MM simulations of the system, including umbrella sampling, were performed with the GROMACS simulation package[122] interfaced with DFTB+ 19.1[132,133] and patched with Plumed 2.5.1.[123,134]. If not stated otherwise, the settings described in the following were used in all simulations.

The QM region comprised the retinal including the K243 side chain to which it is bound, the side chain of D239, and a water molecule in close proximity to the two molecules. Bonds between Cα and Cβ were treated with link atoms. The QM region was only set up for one monomer and consisted of 73 atoms which were described with semiempirical DFTB3 using the 3OB parameter set.[54,135] All other atoms in the system were described and treated with the CHARMM36 force field.[113] A temperature of 300 K and a pressure of 1 bar were set with the Nosé-Hoover thermostat and the Parrinello–Rahman barostat. Equations of motion were solved with the leap-frog integrator using a time step of 1 fs.

Excited state calculations.
Excitation energies were calculated with the OM2/MRCI method as described in Ref. [16], which follows the application for channelrhodopsin-2 described in Ref. [165], where 20 electrons and 20 orbitals were considered for the active space. The MNDO2005 program package was used.[166,167] Not that in Ref. [165] a shift of 0.3 ev was observed when compared to SORCI calculations, probably due to polarization effects.

In this study, the excitation energies were calculated for every 1 ps along the investigated trajectories. Only the retinal was considered quantum mechanically and the atoms of the protein and water environment were included as MM point charges. Since OM2/MRCI is limited to certain amount of MM point charges, the lipid bilayer was not considered in the calculations. In a similar study of bR, the neglect of the membrane led to a red shift of 0.1 eV. The absorption maxima were determined by fitting a Gaussian function to the obtained excitation energies after they were reweighed with the oscillator strength.

Heating of the protein.
To accelerate the configurational sampling, the protein including the QM region was heated from 300 K to 450 K over 50 ps. The temperature was kept at 450 K for additional 50 ps, then cooled down to 300 K over 50 ps, where it stayed 50 ps. This heating and cooling cycle was repeated five times, resulting in a total simulation time of 1 ns. The equilibrated structures of the five QM/MM MDs (replica 1 to 5, performed by Franziska Wolff in Ref. [16]) served as starting structures, i.e., five simulations were performed. The temperature of the solvent and the lipids remained at 300 K during the simulations, since higher temperatures led to water molecules entering the bilayer and to the formation of pores in the bilayer.

In two simulations, a water wire between the Schiff base and D239 was formed (replica 1 and 5). The additional water molecules forming the water wire or water molecules which were in close proximity to the Schiff base and D239 were put into the QM region and described with DFTB3 using the 3OB paramter set. Five additional QM water were considered in replica 1 and six additional QM water in replica 5, resulting in a total of six and seven QM water molecules in the QM region, respectively. To prevent an outflow of the QM water molecules from the QM region, spherical harmonic restraints were applied to the oxygen atoms. They were allowed to move freely within a sphere centered in their initial positions. Beyond a defined radius r, harmonic restraints with a force constant of 100000 kJ/(mol·nm^2) set in. For both replicas, a radius of r = 1 Å and 2 Å was considered, i.e., four independent 2 ns long equilibration simulations were performed. The excitation energies of the last nanosecond were calculated and the red shift evaluated. Only the equilibrated structures of replica 5 with r = 1 Å and 2 Å served as starting structures for subsequent umbrella sampling.

Umbrella Sampling.
For each starting structure, the respective radii of r = 1 Å and 2 Å for the spherical restraints of the QM water molecules were either used further or removed so that the molecules can move freely. Thus, four sets of simulations were performed.

To describe the PT from the Schiff base to D239 along the water wire, the modified center of excess charge (mCEC) coordinate ζ described in Sec. 7.2.2 was used. The nitrogen atom

of the Schiff base was considered as the initial proton donor, D239-Cγ as the final proton acceptor, and in addition five QM water which connect both groups were considered in the description.

For umbrella sampling, 38 windows were considered with ζ values ranging from -1.00 to 0.85 with an increment of 0.05. A harmonic restraint with a force constant of 2000 kJ/(mol·nm^2) was applied to keep ζ at the respective values. Each window was simulated for 1 ns, resulting in a total simulation time of 38 ns. The PMFs of the reactions were obtained by the weighted histogram analysis method using the implementation of Alan Grossfield.[168] For the analysis, only the last 500 ps of the trajectories were considered.

8.3. Results

P550 state.
The retinal of the HKR homology was isomerized from 13-*cis*,15-*anti* to 13-*cis*,15-*syn* with the CASSCF method in Ref. [16]. Afterwards, 2 ns long QM/MM simulations were performed for five replicas using the same starting structure but different initial velocities. In experiments the P550 state is formed after ~458 ps and the subsequent P570 state after additional 2.5 ms. Hence, the simulation times of 2 ns should be long enough to reach and equilibrate the P550 state.

In this work, the excitation energies of each replica were calculated using the OM2/MRCI method for snapshots of the last 1 ns. The obtained energies and the red shift compared to the excitation energy of the initial Rh-Bl state (2.89 eV with OM2/MRCI) are shown in Tab. 8.1. Moreover, the averaged N–Cγ distance between the Schiff base and D239 are

Table 8.1.: Excitation energies of the P550 state obtained from OM2/MRCI calculations and the average N–Cγ distances between the Schiff base and D239 based on 1 ns QM/MM simulations for each replica.

replica	excitation energy [eV]	red shift to Rh-Bl [eV]	N–Cγ dist. [Å]
1	2.73	0.16	4.73
2	2.66	0.23	4.84
3	2.62	0.27	4.59
4	2.58	0.31	4.91
5	2.57	0.32	4.93

shown. The experimental red shift of 0.29 eV is qualitatively reproduced by replica 1 and 2 with 0.16 and 0.23 eV. With red shifts of 0.27, 0.31 and 0.32 eV, replica 3, 4 and 5 show even better agreement with experiments. The initial N–Cγ distance of 3.28 Å in the Rh-Bl state increases in all replicas to values around 4.59 to 4.93 Å. The active site molecules are shown in Fig. 8.2 for the final structure of each replica, i.e., after 2 ns QM/MM molecular dynamics. In replica 1 and 4, two MM water molecules moved near the active site. Histograms of the excitation energies and N–Cγ distances are shown in Fig. B.1.

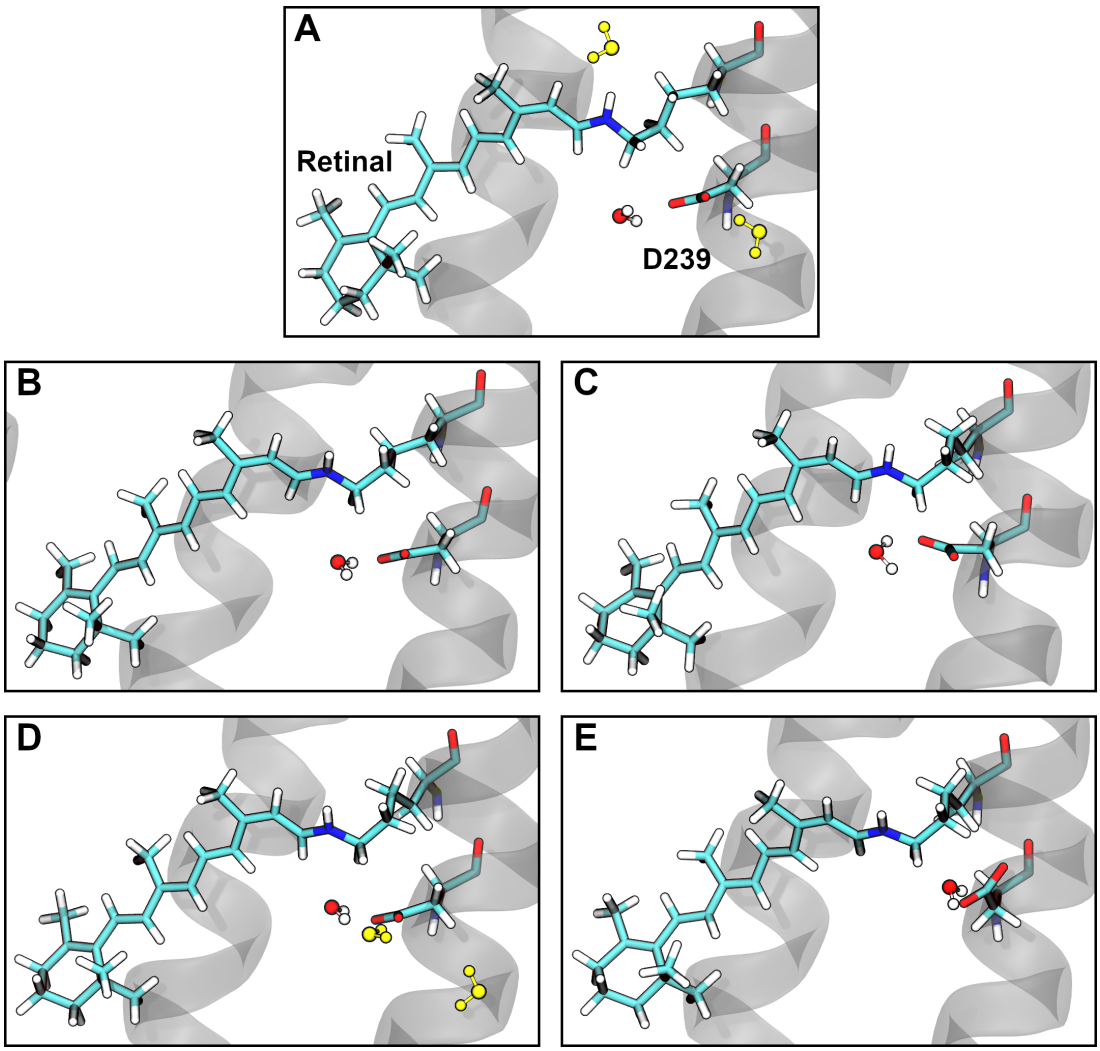

Figure 8.2.: (A-F) P550 structures of the active site molecules including one QM water for replica 1 to 5. MM water molecules near the active site are colored in yellow.

P570 state

The transition from P550 to P570 with a time constant of 2.5 ms exceeds the limit of MD simulation. Thus, in each replica the protein was repeatedly heated up to 450 K and cooled down to 300 K. Higher temperatures accelerate the configurational sampling and P570 should be formed faster. After the heating and cooling cycles, the active sites of replica 2, 3 and 4 look very similar to the P550 state active sites in Fig. 8.2. The Schiff base and D239 are spatially separated and the QM water always remains close to D239, therefore a subsequent PT seems unlikely. In replica 1 and 5, several water molecules entered the active site and formed a water wire between the Schiff base and D239. It seems possible that a PT may occur in these structures along the newly formed water wire. Thus, only replica 1 and 5 were considered for further investigation. The additional water molecules were treated as QM water molecules and an outflow of the molecules from the the active site was prevented by applying spherical restraints. Thus, the water molecules could move freely within a sphere of radius r and beyond the radius they were pushed back closer to

their initial positions. Both systems were duplicated and subsequently equilibrated for 2 ns, setting r to 1 Å in the first duplicate and to 2 Å in the second duplicate. For the last 1 ns, the excitation energies were calculated with OM2/MRCI. The obtained maxima are listed in Tab. 8.2.

The experimental red shift of 0.37 eV for the Rh-Bl to P570 transition is not reproduced. Replica 1 exhibits excitation energies of 2.88 and 2.83 eV, for r to 1 and 2 Å, which corresponds to a red shift of 0.01 and 0.06 eV. Thus, the excitation energy is nearly the same as for Rh-Bl. Replica 5 agrees qualitatively better with experiments. The obtained excitation energies of 2.67 and 2.70 eV correspond to red shifts of 0.22 and 0.19 eV. Still, the obtained energies are higher than those for the P550 state which could be attributed to insufficient sampling resulting in an inaccurate fit of the Gaussian. The obtained energies for replica 5 in the P570 state are more scattered (compare histograms in Fig. B.2C and D) compared to those in Fig. B.1 for replica 5 in the P550 state. A longer sampling may lead to a more accurate fit and therefore a red shift that is in better agreement with experiments. Nonetheless, the obtained structures look promising for a potential PT from the Schiff base to D239 along the water wire.

Table 8.2.: Excitation energies of the P570 state obtained from OM2/MRCI calculations.

replica	excitation energy [eV]	red shift to Rh-Bl [eV]
1 (r = 1 Å)	2.88	0.01
1 (r = 2 Å)	2.83	0.06
5 (r = 1 Å)	2.67	0.22
5 (r = 2 Å)	2.70	0.19

Proton transfer during P570→Rh-UV transition – Umbrella Sampling.
The final structures of replica 5 with r = 1 and 2 Å, referred to as $r_{1\,\text{Å}}$ and $r_{2\,\text{Å}}$ in the following, were used to simulate the PT from the Schiff base to D239 with Umbrella sampling (US). Four sets of umbrella sampling were performed in which spherical restraints were either applied (restr.) on the QM water oxygen or removed (free), so that the water molecules could move freely. The respective US sets are referred to as $r_{1\,\text{Å}}$ (restr.), $r_{1\,\text{Å}}$ (free), $r_{2\,\text{Å}}$ (restr.), and $r_{2\,\text{Å}}$ (free) in the following.

The obtained PMFs are shown in Fig. 8.3 together with exemplary structures of the reactant state (P570), transition state and product state (Rh-UV). The snapshots were taken from respective trajectories of $r_{2\,\text{Å}}$ (restr.).

In all four PMFs, the protonated Schiff base (P570) at $\zeta = -1$ corresponds to the global minimum and the deprotonated Schiff base (Rh-UV) around $\zeta = 0.7$ to a local minimum. The PT proceeds via a proton hole mechanism in all simulations. For the $r_{1\,\text{Å}}$ simulations, the reaction barriers are 72 and 82 kJ/mol, with and without restraints on the QM water, respectively. The local minima of the Rh-UV state are only slightly smaller in energy with 69 and 73 kJ/mol. Thus, the P570 is clearly more stable which contradicts the experimental findings where Rh-UV is stable for more than 24 h. In the $r_{2\,\text{Å}}$ simulations, the barrier heights are 64 and 58 kJ/mol which is in good agreement with the estimated barrier based

on the experimental time constant of 27 ms using transition state theory. The energies of the local minima are 42 and 38 kJ/mol, therefore the protonated Schiff base of P570 is still more stable than the deprotonated Schiff base of Rh-UV.

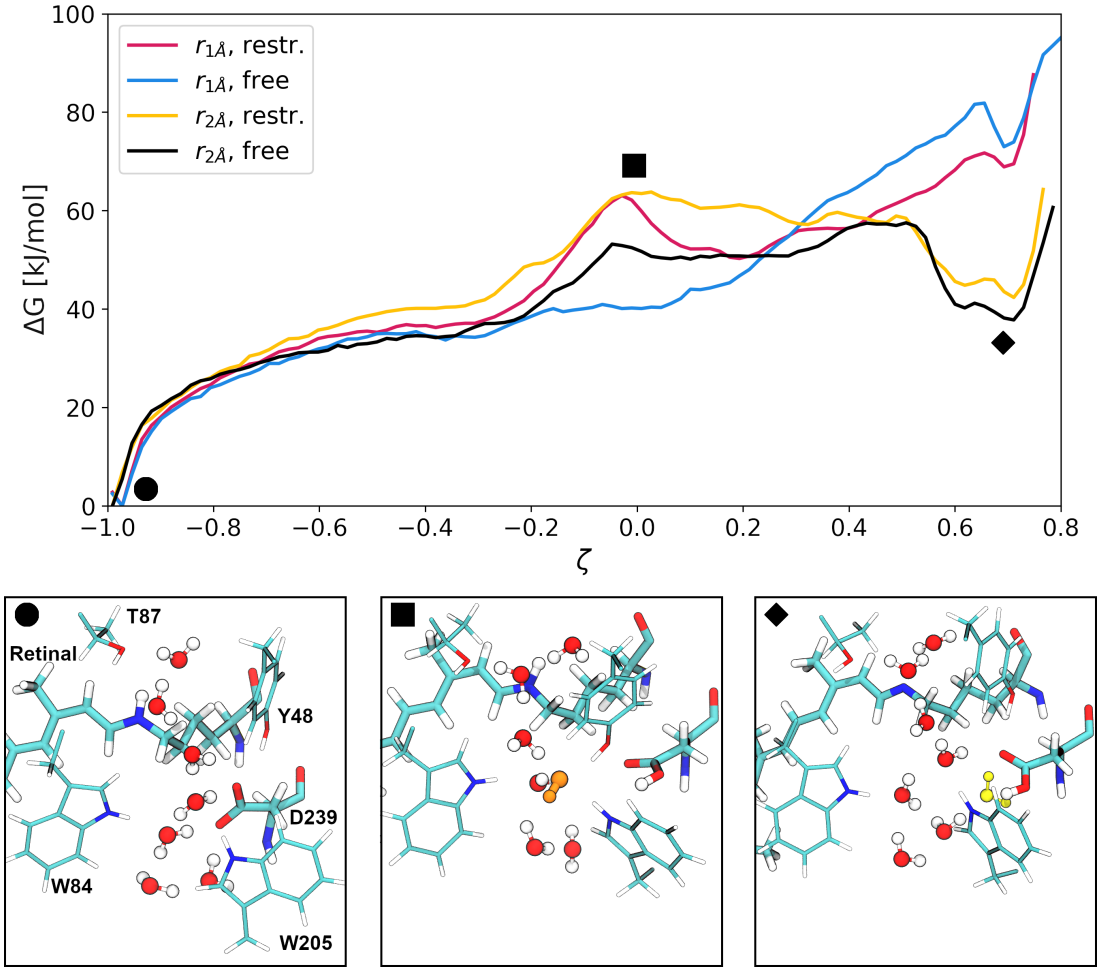

Figure 8.3.: PMFs of PT from the Schiff base to D239 obtained with Umbrella Sampling together with exemplary structures of the active site of the reactant state, transition state and product state. Hydroxide ion shown in orange and MM water in proximity to the active site are shown in yellow.

8.4. Conclusion and outlook

Franziska Wolff built a homology model of HKR since no crystal structure is available. This work aimed to verify the model by simulating the Rh-Bl to Rh-UV transition with QM/MM MD simulations.

After isomerization of the retinal with the CASSCF method, an ensemble of five QM/MM MD simulations was performed to reach the P550 state. The experimentally observed red shift is qualitatively reproduced in two simulations and even quantitatively in three simulations. The next intermediate is formed after 2.5 ms which exceeds the limit of computer simulations. Thus, the protein was repeatedly heated and cooled down to accelerate the configurational sampling. We observed the formation of a water wire between the Schiff base and D239 in two of the five simulations. In the other three simulations, the Schiff base and D239 were spatially separated with no additional water molecules in between. The final step of the Rh-Bl→Rh-UV involves a PT from the Schiff base to D239, which we hypothesize to occur via a water wire. Thus, only the simulations with a water wire were considered further as P570 state. The experimental red shift of the P570 was only qualitatively matched by one simulation which was then used as basis for umbrella sampling to simulate the proton transfer. The obtained free energy profiles show that the protonated Schiff base is the global minimum and the deprotonated Schiff base merely a local minimum. Since the RH-UV state is stable for more than 24 h, one would expect that deprotonation of the Schiff base would lead to a significantly deep global minimum.

Hence, the starting structures obtained from the heating/cooling cycles may not resemble the P570 state or were not equilibrated long enough. It could also be that the umbrella sampling simulations did not converge. In some of the windows MM water molecules disrupted the QM water wire which might have lead to inaccurate sampling of the PT pathway. Moreover, the overlap of some windows me not be sufficient, compare Fig. B.3 and B.4. Another possibility might be, that the water wire collapses once the Schiff base is deprotonated. Thus, the backtransfer cannot proceed since the Schiff base and D239 too far away. This hypothesis can be tested by using the previously described heating and cooling approach for structures in which the Schiff base is deprotonated.

9. Metadynamics simulations of proton-coupled electron transfers using coupled-perturbed equations in a DFTB3 and DFTB3/MM setup.

Author Contributions:
Tomáš Kubař implemented the coupled-perturbed equations for DFTB3 and the derivative of QM atomic charges with respect to MM atomic coordinates. Denis Maag created the test systems and different setups and supervised Josua Böser, who performed the metadynamics simulations for his Master's thesis (Ref. [169]).

9.1. Introduction

In a recent work, Gillet et al.[35] introduced a new concept of free energy calculations. They used Mulliken charges in the density-functional tight binding (DFTB) method as one or more collective variables S

$$S = S(Q) = S(Q(\vec{r})) \tag{9.1}$$

in biasing potential simulations, such as metadynamics or Umbrella Sampling. The atomic charges Q are readily available in DFTB and depend on the molecular geometry, i.e., the atomic positions $\vec{r}$. The forces on atoms due to an applied biasing potential $V(S)$ are obtained as the derivatives of the biasing potential with respect to atomic coordinates a. By chain rule, this leads to:

$$F_a = \frac{\partial V(S(Q(\vec{r})))}{\partial a} = \frac{\mathrm{d}V(S)}{\mathrm{d}S} \cdot \frac{\mathrm{d}S(Q)}{\mathrm{d}Q} \cdot \frac{\partial Q(\vec{r})}{\partial a}. \tag{9.2}$$

Since V is usually a quadratic function or a sum of Gaussian functions, the first derivative $dV(S)/dS$ can be easily determined, and the second derivative $dS(Q)/dQ$ is also calculated easily. The derivatives of atomic charges with respect to atomic coordinates $\partial Q(\vec{r})/\partial a$ have to be calculated by means of coupled-perturbed (CP) equations, originally derived, implemented and tested by Witek et al.[55]

In order to use the Mulliken charges as CVs, the solution of the CP-DFTB equations were adopted by Gillet et al. and implemented in the DFTB code within the QM/MM framework of Gromacs[130] and the plumed software[123]. The new framework was applied to a small model system that consists of the side chains of two tyrosines with one proton removed from an oxygen atom. First, they performed a series of unbiased QM simulations

and metadynamics QM simulations of the PT, ET and PCET between the molecules in the gas phase. The simulations were stable and identified the free energy minima correctly. Subsequently, the tyrosine side chains were solvated in a box of TIP3P water molecules and the transfer reactions simulated in a QM/MM setup. These simulations, however, were unstable and the free energy profiles did not converge. It turned out that the derivatives of the QM atomic charges with respect to the MM atomic coordinates were not considered in the calculation of the CP-DFTB equations.

In this work, the CP equations were extended to third order DFTB. In addition, the derivatives of QM charges with respect to MM coordinates were implemented which makes it possible to perform hybrid QM/MM simulations. In order to thoroughly test the implementation, we performed QM and QM/MM metadynamics of PCET reactions between two tyrosine sidechains. The potentials of the mean force were obtained for 32 different setups which differ in their charge state, conformations and environment, compare Fig. 9.1.

9.2. Computational Details

Density-functional tight binding and coupled-perturbed equations.
DFTB and the coupled-perturbed equations for DFTB are described in Sec. 3.4. The derivatives of QM charges with respect to coordinates of MM atoms were adopted from Benjamin Hourahine.[56]

System Setup.
The two tyrosine side chains in the test systems, hereafter referred to only as tyrosines, either carry a negative charge (Tyr_2^- anion) or have an unpaired electron ($Tyr_2^\bullet$ radical). In addition, they were considered in two different conformations, "flipped" or "stacked". The names refer to the configurations of αY731 with βY356 and αY731 with αY730 in RNR, respectively.[21] A PCET across the α/β interface in RNR probably only occurs when αY731 is flipped-out and facing βY356, whereas a PCET between αY731 and αY730 probably only occurs in a stacked configuration. Initial structures for the test systems were taken from from Ref. [170], where QM/MM simulations of RNR based on a docking model[171] were performed.

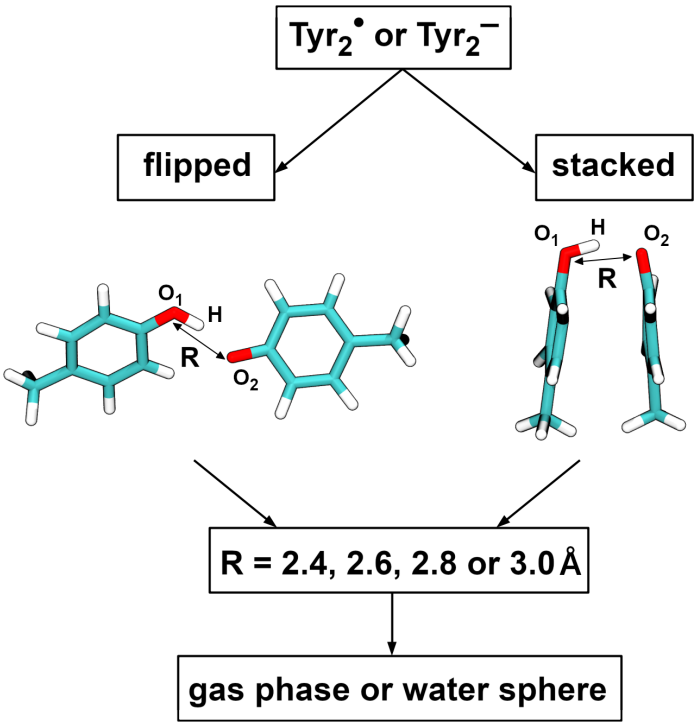

Figure 9.1.: Overview of the different systems used for testing the CP equations implementation in DFTB3. The systems consist of the side chains of two tyrosines starting from Cβ with one proton removed from oxygen atom O_2. Two charge states were considered, negatively charged (Tyr_2^-) or neutral with an unpaired electron ($Tyr_2^\bullet$), as well as two different conformations, called flipped and stacked. In addition, the O_1–O_2 distances were restrained to 2.4, 2.6, 2.8 or 3.0 Å. The systems were then simulated either in the gas phase or in a sphere of water molecules, resulting in a total of 32 different setups.

The two obtained structures of the flipped and stacked conformations were centered in a cubic box sized $100 \times 100 \times 100$ nm^3. Next, the positions of all atoms were manually sifted to adopt O_1–O_2 distances of R = 2.4, 2.6, 2.8 or 3.0 Å, on which harmonic restraints with a force constant of 100 000 kJ mol^{-1} nm^{-2} were applied. To maintain the overall conformations while leaving enough room for changing the atomic positions to modify the atomic charges during the simulations, spherical restraints were applied to the carbon atoms of the aromatic rings and the oxygen atoms. This allowed the atoms to move freely within a distance of 0.5 Å from their initial positions. Beyond the set radius, a harmonic restraint with a force constant of 100 000 kJ mol^{-1} nm^{-2} set in and pulled the atoms back into the sphere. Additional harmonic restraints with a force constant of 1000 kJ mol^{-1} nm^{-2} were applied to the sum of the O_1H and O_2H distances, when $|O_1H| + |O_2H| > (R + 0.2\,\text{Å})$, to favor the hydrogen bonded configurations.

Finally, the 16 generated setups were duplicated for two different sets of simulations. In the first set, the tyrosines were kept in the gas-phase. In the second set, a solvation shell of water molecules was formed around the molecules. Both sets of simulations were performed with a local Gromacs 2020 version[122] patched with Plumed 2.5.1[123,134] and interfaced with DFTB+ 19.1[132,133] with the implemented CP-DFTB equations[55,133] for DFTB3.

Simulations in the gas-phase.
The gas phase systems were described solely with DFTB3 using the 3OB parameter set. They were equilibrated for 1 ns using the leap-frog integrator with a time step of 0.5 fs. Electrostatics and Van der Waals interactions were cut-off after 2 nm. Subsequently, multiple walker metadynamics[59,60,131] of the proton-coupled electron transfers were performed using two collective variables.

The first CV is the difference of O–H distances,

$$\Delta d = |O_1H| - |O_2H| \tag{9.3}$$

and the second CV the difference of the total charge of each tyrosine excluding the transferable H atom,

$$\Delta Q = \underbrace{\sum_{i}^{\text{mol\#1}} \Delta q_i}_{Q_1} - \underbrace{\sum_{j}^{\text{mol\#2}} \Delta q_j}_{Q_2} \tag{9.4}$$

where Δq_i and Δq_j are the Mulliken charges of atoms belonging to the first tyrosine and the second tyrosine, respectively.

All gas phase simulations used 16 walkers in which Gaussian biasing potential were added every 500 steps with a height of 0.5 kJ mol^{-1} and a width of 0.05 Å for the first CV and 0.02 e for the second CV. The bias between walkers was exchanged every 500 steps. All $Tyr_2{}^-$ systems used a time step of 0.5 fs and each of the 8 setups yielded a total simulation time of 2 ns. For the $Tyr_2{}^\bullet$ systems, the time step was increased to 1 fs and each setup yielded a total simulation time of at least 11.2 ns and up to 16.5 ns.

Simulation in aqueous solution.
For the simulations in aqueous solution, 35 water molecules were placed around the

flipped conformations and 30 water molecules around the stacked conformations such that a solvation shell is formed. The simulations were performed in a QM/MM approach, i.e., the tyrosines were described with DFTB3 using the 3OB parameter set and the water molecules were modeled as TIP3P water described with MM. To keep the water molecules in place, spherical restraints were applied to the oxygen atoms, i.e., after moving further away than 1 Å from their initial positions, harmonic restraints with a force constant of 100 000 kJ mol^{-1} nm^{-2} set in pushing the atoms back.

Analogously to the simulations in the gas phase, the solvated tyrosine systems were equilibrated for 1 ns with a time step of 0.5 fs, electrostatics and Van der Waals interactions were cut off after 2 nm. Next, metadynamics simulations were performed with the same set of reaction coordinates, Δd and ΔQ, using the same gaussian heights, widths, deposition rates and bias exchange strides. However, the time step was set to 1 fs for all simulations and 24 walker were used in each setup. The flipped and stacked Tyr_2^- systems yielded a total simulation time of 2.2, 3.2, 4 and 7 ns for R = 2.4, 2.6, 2.8 and 3.0 Å. The flipped $Tyr_2^\bullet$ systems yielded a total simulation time of 3.9 ns for R = 2.4 Å, and 4.2 ns for R = 2.6, 2.8 and 3.0 Å. The stacked $Tyr_2^\bullet$ systems yielded a total simulation time of 9.2, 7.8, 9.4 and 6.8 ns for R = 2.4, 2.6, 2.8 and 3.0 Å.

9.3. Results

The two dimensional free energy landscapes of the proton-coupled electron transfers are shown in Fig. 9.2, 9.3, 9.4 and 9.5. In all PMFs, a negative Δd value corresponds to a O_1–H bond, and a positive Δd value to a O_2–H bond. The ΔQ value depends on the total charge of the systems.

In the anionic Tyr_2^- systems, the H atom carries a partial charge of about 0.4 e. When it is bonded to O_1, then $Q_1 \approx -0.4$ e. Tyrosine #2 carries the negative charge, therefore $Q_2 \approx -1.0$ e and consequently $\Delta Q = 0.6$ e. The same applies vice versa when the hydrogen is bonded to tyrosine #2.

In the radical $Tyr_2^\bullet$ systems the transferable hydrogen atom carries a partial charge of about 0.2 e and when it is bonded to tyrosine #1, the sum of the atomic partial charges of tyrosine #1 is about −0.2 e. Since the system is neutral, the total charge of tyrosine #2 is about 0.0 e. It follows that $\Delta Q = -0.2$ e and vice versa when H is bonded to O_2.

9.3.1. Anionic systems

Fig. 9.2 and 9.3 show the two dimensional free energy landscapes of the PCET for the Tyr_2^- systems in the flipped and stacked conformations. Qualitatively, they all look almost the same, except for the gas phase simulations with R = 2.4 Å. In these two simulations, there is one narrow minimum around $\Delta d = 0$ and $\Delta Q = 0$, which corresponds to a shared proton between O_1 and O_2. In all other simulations, there are two narrow minima corresponding to the O_1–H and O_2–H bonds. The minima should be of equal depth, which is the case for most simulations, but not all. This can be attributed due to a lack of convergence which can be improved by performing or extending the simulations with the well-tempered variant of metadynamics. Nonetheless, all simulations ran without any problems and show

the expected behaviour, such as an increase of the barrier height with increasing $|O_1 - O_2|$ distances. The obtained transition state energies are visualized in Fig. 9.6 and listed in Tab. 9.1.

As mentioned before, for R = 2.4 Å, the gas phase simulations in the flipped and stacked conformation share the hydrogen atom and consequently there is no barrier. At larger distances, the proton is no longer shared and the height of the barriers increase with the O_1–O_2 distances. In the gas phase, the transition state energy $\Delta E^{\ddagger}$ of the stacked conformation is 6 and 3 kJ/mol lower than for the flipped conformation at intermediate distances of R = 2.6 and 2.8 Å. At R = 3.0 Å, however, the barrier is 12 kJ/mol higher in the stacked conformation.

When the systems are solvated in a small sphere of MM water, the reaction barriers increase significantly. For R = 2.4 Å, the proton is no longer shared between O_1 and O_2, but bonded to either one of them. For a PCET, a barrier of 21 and 16 kJ/mol has to be overcome in the flipped and stacked conformation, respectively At $|O_1 - O_2| = 2.6$ Å, the barrier heights are 39 and 33 kJ/mol. Hence, a PCET occurs more likely at short O_1–O_2 distances when the tyrosines are stacked. In contrast, at larger O_1–O_2 distances of 2.8 Å and 3.0 Å the barriers heights in the flipped conformations are lower by 6 and 9 kJ/mol than in the stacked conformation, respectively.

Note that for comparable O_1–O_2 distances, the reaction barriers obtained in this work are higher than those obtained in the earlier work of Gillet et al.[35]. In Ref. [35] the tyrosines were less restrained than in this work, which led to larger fluctuations towards shorter O_1–O_2 distances and consequently smaller barriers.

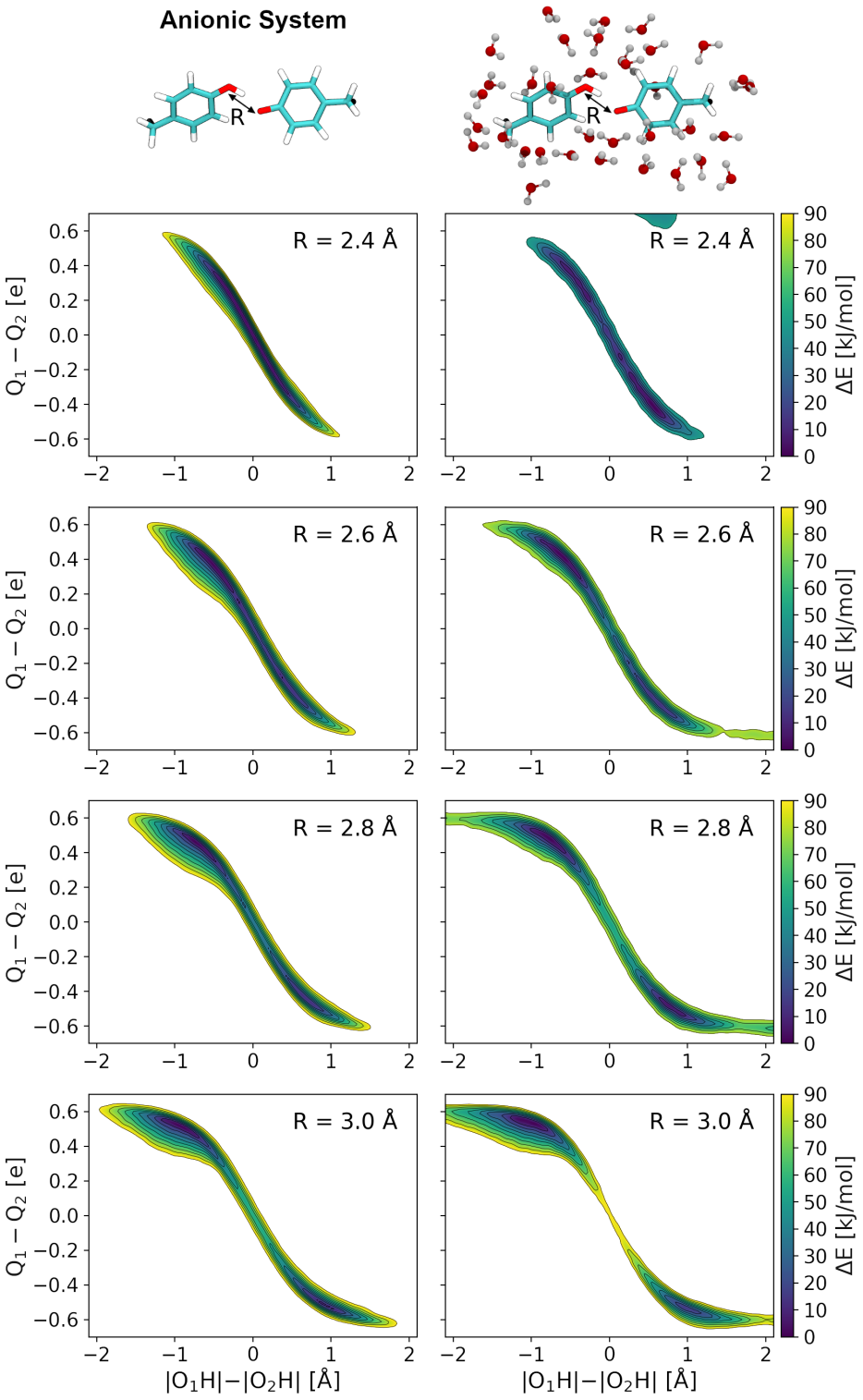

Figure 9.2.: Potentials of the mean force of PCET reactions in the flipped Tyr_2^- conformation in the gas phase (left) or in a sphere of water molecules (right) for different O_1–O_2 distances of R = 2.4, 2.6, 2.8 and 3.0 Å (top to bottom). The horizontal axis represents the PT coordinate and the vertical axis the ET coordinate. Contour lines are drawn every 10 kJ/mol.

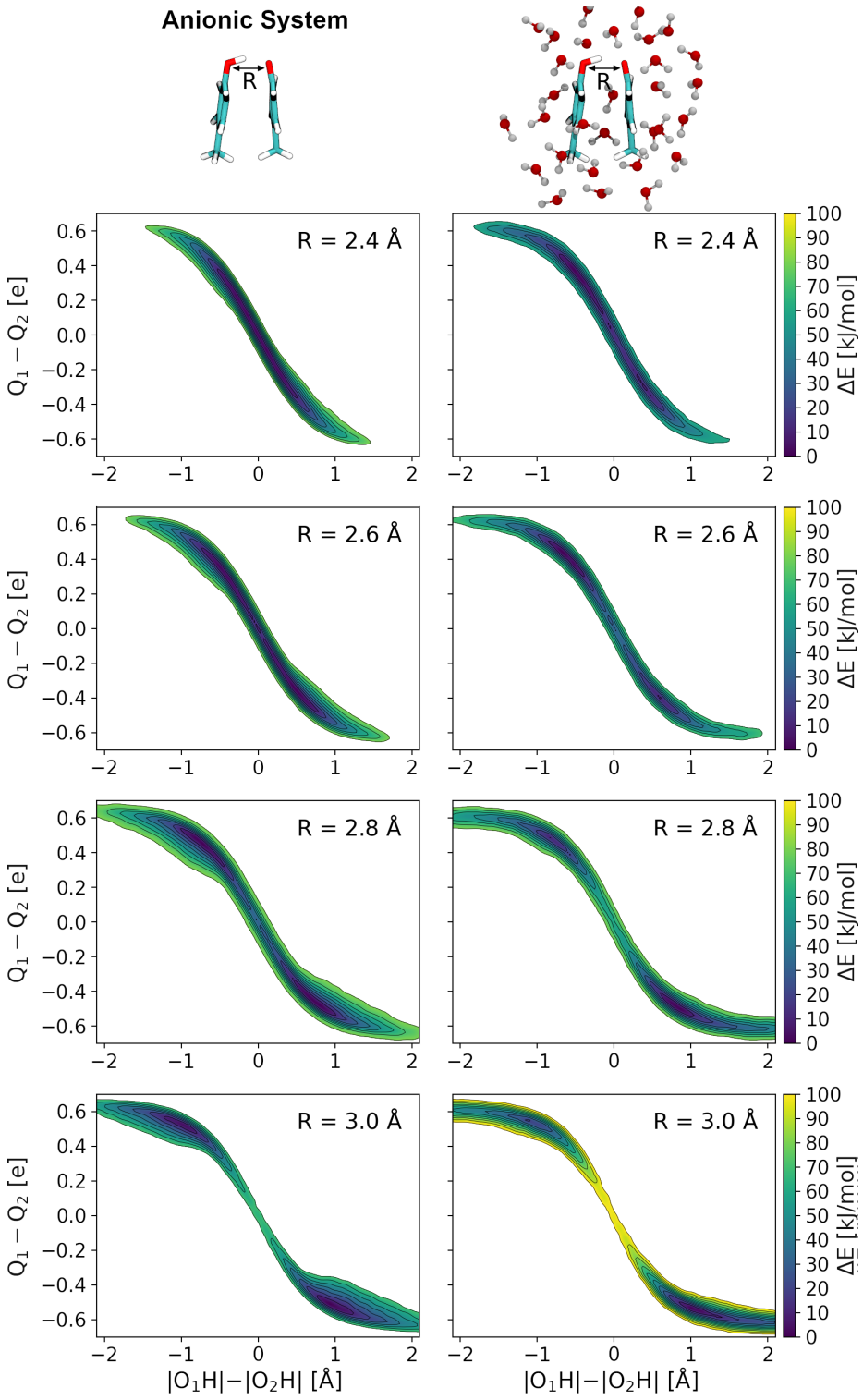

Figure 9.3.: Potentials of the mean force of PCET reactions in the stacked Tyr_2^- conformation in the gas phase (left) or in a sphere of water molecules (right) for different O_1–O_2 distances of R = 2.4, 2.6, 2.8 and 3.0 Å (top to bottom). The horizontal axis represents the PT coordinate and the vertical axis the ET coordinate. Contour lines are drawn every 10 kJ/mol.

9.3.2. Radical systems

The two dimensional PMFs of the PCET reactions for the $Tyr_2^{\bullet}$ systems are shown in Fig. 9.4 and 9.5. Not all simulations converged yet due to the broad free energy landscape. Compared to the narrow PMFs of the Tyr_2^{-} systems, more gaussian biasing potentials are needed to fill the free energy landscape. Hence, especially the simulations in aqueous solution at O_1–O_2 distances of 2.8 Å and 3.0 Å need need to be extended, as well as the gas phase simulation with R = 3.0 Å in the stacked conformation.

The converged simulations in the gas phase exhibit two broad minima of equal depth corresponding to the O_1–H and O_2–H bonds. At short O_1–O_2 distances of 2.4 Å and 2.6 Å, the barrier heights for a PCET in the flipped and stacked conformation are 11 and 37 kJ/mol (flipped) and 12 and 36 kJ/mol (stacked). Hence, they only differ by 1 kJ/mol. At R = 2.8 Å, the barrier of 57 kJ/mol in the flipped conformation is 6 kJ/mol smaller than in the stacked conformation. Since the metadynamics simulation of the stacked conformation at R = 3.0 Å did not converge, the barrier heights cannot be compared for this O_1–O_2 distance.

When the tyrosines are embedded in a sphere of water molecules several interesting things happen. The reaction barriers do not increase significantly for R = 2.4 Å, only by 3 and 2 kJ/mol in the flipped and stacked conformation, respectively. At R = 2.6 Å the heights of the barriers even decrease by 8 and 9 kJ/mol.

Moreover, the topography of the free energy landscapes change compared to the gas phase simulations. In the flipped conformation with R = 2.4 Å, the two minima are broader in direction of the ET coordinate. At longer O_1–O_2 distances of 2.6, 2.8 and 2.8 Å, the minima are very narrow and shift more negative/positive values of the ET coordinate, i.e., the charges are more localized.

In the stacked conformation, the minima are slightly broader compared to the gas phase simulations. Thy are shifted closer to ΔQ = 0 e, i.e., the sum of the partial charges of tyrosine #1 and #2 are nearly the same. However, as the O_1–O_2 distance increases, the charge differences appear to become larger. A definite statement can only be made when the simulations for R = 2.8 and 3.0 Å have converged. Note that the minima for R = 2.4 Å are not of equal depth, due to a nearby water molecule forming a hydrogen bond with O_2 and therefore stabilizing the deprotonated state of tyrosine #2.

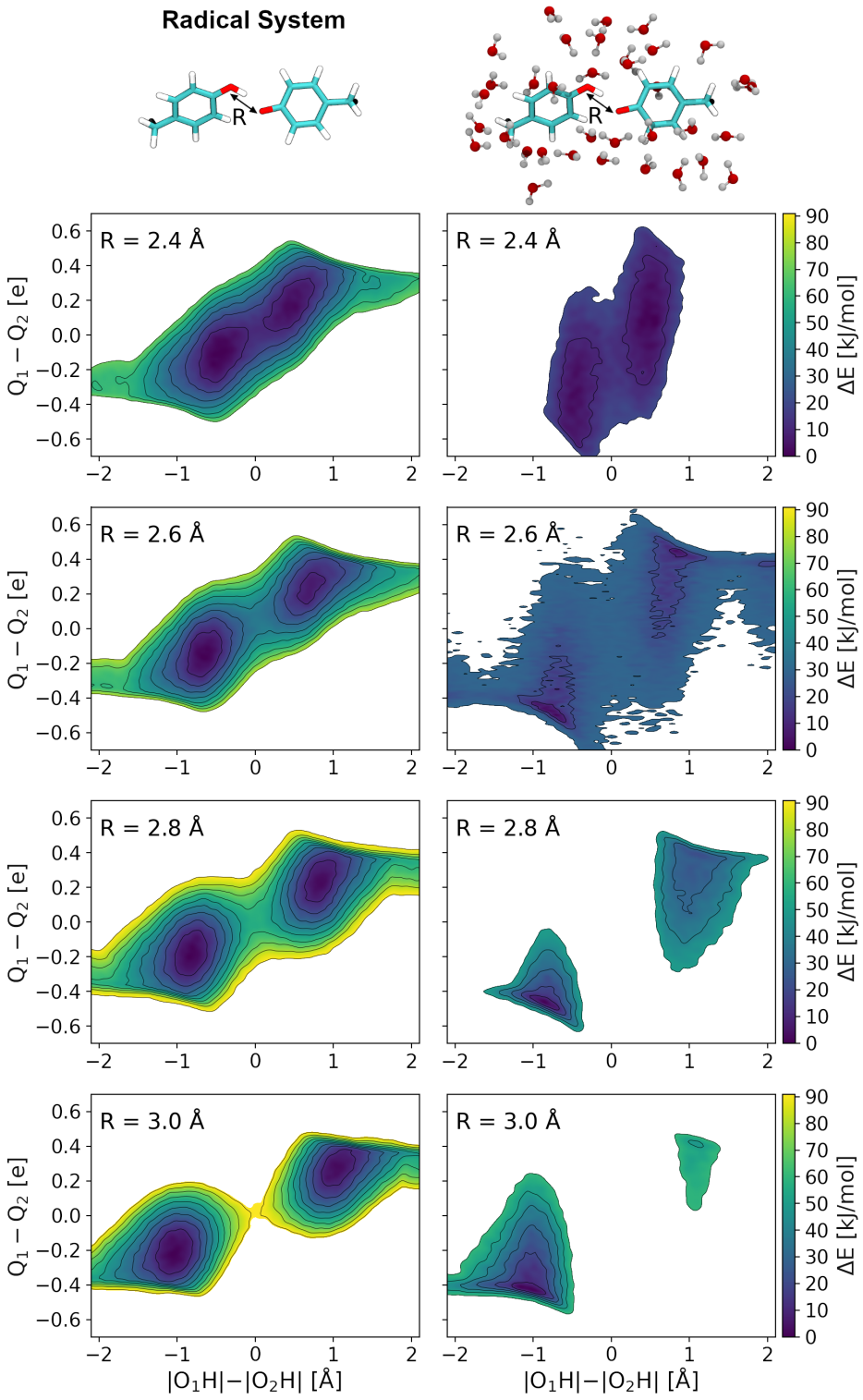

Figure 9.4.: Potentials of the mean force of PCET reactions in the flipped $Tyr_2^{\bullet}$ conformation in the gas phase (left) or in a sphere of water molecules (right) for different O_1–O_2 distances of R = 2.4, 2.6, 2.8 and 3.0 Å (top to bottom). The horizontal axis represents the PT coordinate and the vertical axis the ET coordinate. Contour lines are drawn every 10 kJ/mol.

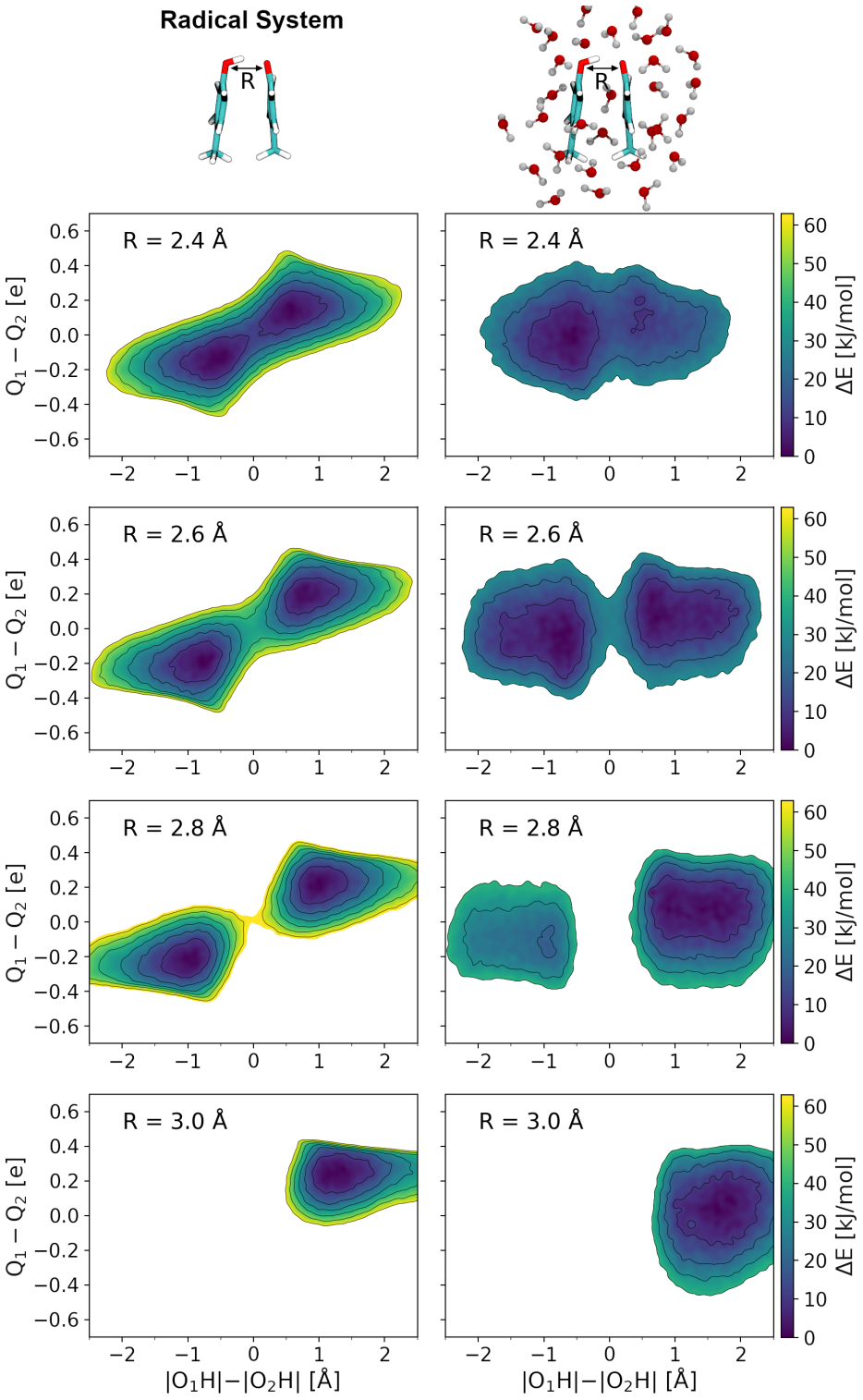

Figure 9.5.: Potentials of the mean force of PCET reactions in the stacked $Tyr_2^{\bullet}$ conformation in the gas phase (left) or in a sphere of water molecules (right) for different O_1–O_2 distances of R = 2.4, 2.6, 2.8 and 3.0 Å (top to bottom). The horizontal axis represents the PT coordinate and the vertical axis the ET coordinate. Contour lines are drawn every 10 kJ/mol.

Table 9.1.: Reaction barriers of proton-coupled electron transfers for the Tyr_2^- and $Tyr_2^\bullet$ systems performed in gas phase (subscript g) and a sphere of water (subscript w) in the flipped (f) and stacked (s) conformations. Energies are given in kJ/mol. Missing values correspond to unconverged simulations where no energy barrier could be determined.

setup	2.4 Å	2.6 Å	2.8 Å	3.0 Å
$(Tyr_2^-)_g$, f	0	19	36	56
$(Tyr_2^-)_g$, s	0	12	33	68
$(Tyr_2^-)_w$, f	21	39	54	89
$(Tyr_2^-)_w$, s	16	33	60	98
$(Tyr_2^\bullet)_g$, f	11	37	57	91
$(Tyr_2^\bullet)_g$, s	12	36	63	
$(Tyr_2^\bullet)_w$, f	14	29		
$(Tyr_2^\bullet)_w$, s	14	27		

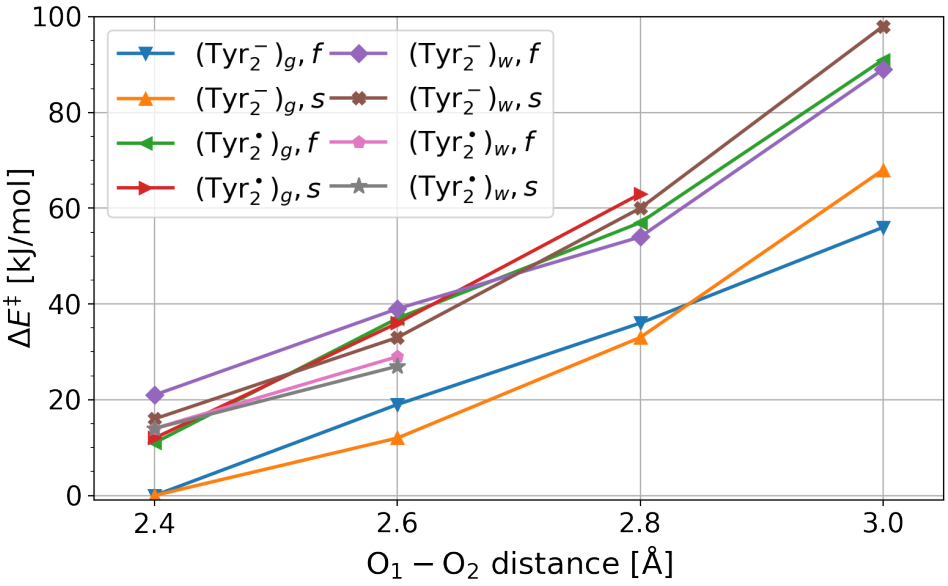

Figure 9.6.: Reaction barriers of proton-coupled electron transfers performed in gas phase (subscript g) and a sphere of water (subscript w) in the flipped (f) and stacked (s) conformations. Missing values correspond to unconverged simulations where no energy barrier could be determined.

9.3.3. Computational efficiency

The presented timings in the following are taken from short unbiased QM and QM/MM simulations. We find that the computational cost increases significantly by solving the CP-DFTB equations. For the anionic Tyr_2^- systems, one MD step in gas phase takes 2.18 s on average on a single Intel Xeon Silver 4214 CPU. Only 1.6 % account for the self-consistent-charges DFTB3 calculation and 98.3 % for the solution of the CP-DFTB equations, which is a 59-fold increase, A MD step of the $Tyr_2^\bullet$ system in gas phase takes 5.18 s, where the DFTB3 calculation takes 1.11 % of the time and the CP-DFTB calculation takes the remaining 98.89 %. Consequently, the cost is increased by a factor of 90.

When the systems are solvated in a sphere of water, the computational cost increases even more because of the additional derivatives due to the MM atoms. One MD step of the Tyr_2^- systems takes 6.71 s, 0.55 % for DFTB3 and 99.45 % for the CP-DFTB equations. That is a 181-fold increase. For the $Tyr_2^\bullet$ systems, one step takes 15.08 s where 0.3 % accounts for DFTB3 and the remaining 99.7 % for the solution of the CP-DFTB equations, i.e., a 333-fold increase.

As mentioned in Ref [35], the calculations of the CP-DFTB equations leave room for optimization by means of parallelization, for example with OpenMP.

9.4. Conclusion and outlook

We implemented the CP equations into DFTB3 as an extension to the previous work by Gillet et al.[35] where biasing potentials were applied on partial atomic charges in extended sampling MD simulations. The gradients of the potentials are calculated by solving the CP-DFTB equations which were originally developed by Witek et al.[55] The previous scheme worked well for pure QM systems, however, hybrid QM/MM simulations were unstable and failed to converge because the derivatives of QM atomic charges with respect to MM atomic coordinates were missing. Based on an earlier development by Benjamin Hourahine,[56] we implemented the missing gradients.

In order to test the new framework, we performed 32 metadynamics simulations of PCETs in a small test system. The systems differed in their charge states and conformations. In addition, the systems were either simulated in gas phase in a QM setup or in a sphere of water molecules in a QM/MM setup. We considered two reaction coordinates, one for the ET process composed of Mulliken atomic charges as introduced in Ref. [35], and one for the PT process.

All QM and QM/MM simulations were stable and converged nicely when simulated long enough. The minima of the systems were correctly identified and meaningful transition state energies were obtained. The only drawback was the computational cost which increased substantially due to the CP equations. Nonetheless, the implementation of the CP-DFTB equations and the additional gradients were successful. The scheme may be used in more complex molecular systems, for example, PCET reactions involving more than two molecules or a PCET reaction in a protein environment.

10. Electrostatic interactions contribute to the control of intramolecular thiol–disulfide isomerization in a protein

Chapter 10 is reproduced from Ref. [172] with permission from the PCCP Owner Societies:

- Maag, D.; Putzu, M.; Gómez-Flores, C. L.; Gräter, F.; Elstner, M.; Kubař, T., Electrostatic interactions contribute to the control of intramolecular thiol–disulfide isomerization in a protein., Phys. Chem. Chem. Phys. 2021.

Author Contributions:
This work was done in cooperation with Marina Putzu, who prepared and performed the 334 QM/MM MD simulation of I27*. Denis Maag analyzed the trajectories and performed QM/MM metadynamics simulations of I27* and the model system with an applied external electrostatic potential.

10.1. Introduction

In a novel approach, Alegre-Cebollada et al. investigated protein unfolding and disulfide isomerization of a mutated I27 immunoglobulin domain (I27*) in real time with force-clamp atomic-force microscopy (AFM).[173] I27* was engineered to have a disulfide bond between Cys24 and Cys55 and a free reactive cysteine Cys32, see Fig. 10.1A. Due to a constant pulling force of 250 pN, far below the force necessary to break covalent bonds (above 1 nN),[174,175] the protein unfolded up to the disulfide bond. Residues 25 to 54, including Cys32, were located on a flexible loop which did not stretch because of the disulfide bond, see Fig. 10.1B. Thus, Cys32 remained in the vicinity of the disulfide bond after the first unfolding step and was able to engage in a nucleophilic attack on Cys24 or Cys55, see Fig. 10.1C and 10.1D. A reaction with Cys55 occurred 3.8 times more frequently than with Cys24. No conclusive explanation of this regioselectivity was found, and it was called for systematic studies on how the reactivity of disulfide bond is affected by their environments.

To this end, Kolšek et al. carried out force-clamp MD simulations on I27* using a molecular mechanics-based framework that allowed for disulfide bond rearrangements through Monte Carlo-controlled topology exchanges.[177] This approach reproduced the regioselectivity observed by the experimental AFM setup with the advantage of an atomistic

description of the process. In the simulations, Cys55 was more readily spatially accessible for a nucleophilic attack than Cys24, and consequently, it was approached by Cys32 more often, leading to the reaction Cys32→Cys55 occurring more frequently. Thus, steric factors play an important role in disulfide shuffling.

So do electrostatic interactions. The rate of thiol-disulfide exchange, an S_N2 reaction, depends on the nucleophilicity of the attacking thiolate S_{nuc}, the electrophilicity of the attacked sulfur S_{ctr} as well as the stability of the leaving group S_{lg}.[40,178] These factors are not solely determined by the reactive species themselves; rather, they are affected by the steric and electrostatic interactions with the environment. Notably, the lowest-energy state of a symmetric molecule (R_1=R_2=R_3) in the gas phase is a linear trisulfide arrangement with the negative charge delocalized over all three sulfurs.[179] This is reflected by the general observation that thiol-disulfide exchange is best catalyzed by hydrophobic, aprotic environments – conditions in which the charge is quite delocalized.[43] On the other hand, polar solvents induce a localization of the charges, favoring arrangements with separated molecules, a thiolate and a disulfide.

The charge distribution on a thiol-disulfide center – and consequently, the nucleophilicity and the electrophilicity of the sulfur atoms – are modulated not only by the solvent but also by the microenvironment, e.g. the neighboring functional groups or amino acids. Wu et al. demonstrated that ionic residues in close proximity to the reactants have a major impact on disulfide exchange reaction rates.[180] They investigated the reaction between a cysteine as a nucleophile and several small disulfide-bonded peptide homodimers. Net charges ranging from −2 to +2 were introduced in each peptide by putting glutamate or arginine residues in positions adjacent to the disulfide-bonded cysteines. The reactivity showed a linear dependence on the introduced net charges, −2 showing the least and +2 the highest reactivity. Similar effects of the environment on the reactivity were observed in other studies, also.[181–184]

This work aims to explain the regioselectivity of the disulfide shuffling in proteins, considering the mutated immunoglobulin domain I27* as an example, providing more detail than the previous work in Ref. [177]. To this end, we perform 334 QM/MM force-clamp simulations of I27*, with an accumulated sampling of ∼5.7 µs. The QM/MM simulations are set up in order to cover a possible nucleophilic attack of the deprotonated reduced Cys32, located on a flexible loop, on both Cys24 and Cys55. To elucidate the prerequisites for a successful disulfide exchange, we analyze 10 ps prior to the formation of transition state in trajectories where a reaction does take place, and analyze potentials of mean force as function of the sulfur–sulfur distances based on all of the trajectories. Also, we perform QM/MM metadynamics simulation of the two different disulfide exchange reactions in I27* but observe difficult convergence. Finally, we demonstrate the impact of electrostatics on disulfide exchange on the basis of metadynamics simulation of a small model system.

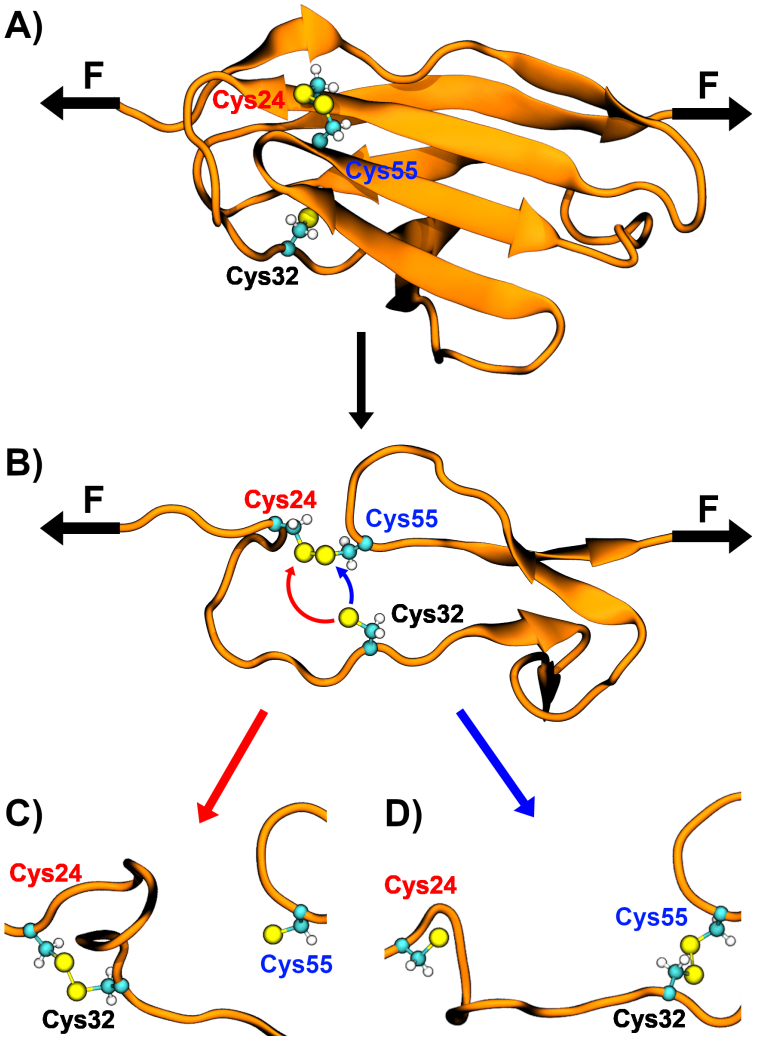

Figure 10.1.: Constant pulling force on the termini of the I27* domain (A) leads to unfolding up to the disulfide bond Cys24–Cys55 (B). This uncages Cys32 so that it can perform a nucleophilic attack on Cys24 (C) or Cys55 (D). Sulfur atoms are depicted by yellow balls. Image created by the authors, based on PDB ID 1WAA.[176]

10.2. Computational Details

10.2.1. QM/MM force-clamp simulations of I27*

System Setup.
We performed 334 QM/MM simulations of an immunoglobulin I27 domain (PDB ID 1WAA),[176] which was engineered to have two oxidized cysteines at positions 24 and 55 forming a disulfide bond and a free reactive cysteine at position 32 by Kolšek et al.[177] Snapshots from their force-clamp swapping simulations were selected as starting structures. In 160 of the selected structures S32 was closer to S24, and in 174 structures closer to S55.

Due to an applied external pulling force on the termini, the protein was already unfolded up to the disulfide bond between S24 and S55. The complete I27* domain consists of 93 amino acids, 21 of which are charged, and carries a net charge of −7 e, see Tab. 10.1. Under mechanical stretching stress, the protein strand unfolds up to the disulfide bond between S24 and S55, with the end-to-end distance increasing to ca. 230 Å, see Fig. 10.2. Hence, an extremely large (long) simulation box would be required to enclose the entire protein strand.

To reduce the box size and thus the computational cost, the molecular system was reduced by removing most of the (completely extended) termini – N-terminus up to Ala19 and C-terminus starting at Gly66. This is justified because the contribution of these termini to the electrostatic potential on the disulfide-bonded sulfur atoms (as discussed in the following) is negligible, and so it their influence on the outcome of the intramolecular thiol–disulfide exchange. After the removal of the first 22 N-terminal and the last 24 C-terminal residues, the truncated protein strand exhibits an end-to-end distance of ca. 60 Å, i.e. one third of the original length.

All of the charged residues that are close to the three reactive cysteines Cys24, Cys55 and Cys32 are located on the loop formed by residues 25 to 54 and are therefore retained in the truncated I27* domain. The nearest charged residue in the termini (that are to be truncated) is Glu17, and its distance from the closest reactive Cys24 is at least $6 \times 4 = 24$ Å, and the other charged residues are further than that. The electrostatic interaction over such a long distance in a system immersed in strongly polarizable aqueous environment vanishes nearly completely.

The net charge of the truncated I27* domain is −5 e. Five positively charged residues were introduced by means of mutations, in order (i) to reproduce the electrostatic relationships in the complete protein as closely as possible, and additionally (ii) to achieve charge neutrality of the system. (The more common way of electroneutralization by means of adding counterions was tested in pilot simulations, which oftentimes showed instability whenever the counterions approached the reaction center.) None of the mutations are located (i) within 30 Å of Cys24 and Cys55, nor (ii) on the flexible loop between Cys24 and Cys55. In the original I27* domain, the N-terminal segment (residues −3 to 23) carries a net charge of $Q_{\mathrm{N-term}} = -3$ e, and the C-terminal segment (residues 56 to 89) is elcetroneutral, $Q_{\mathrm{C-term}} = 0$ e. The charge difference between the two segments is $\Delta Q = Q_{\mathrm{N-term}} - Q_{\mathrm{C-term}} = -3$ e.

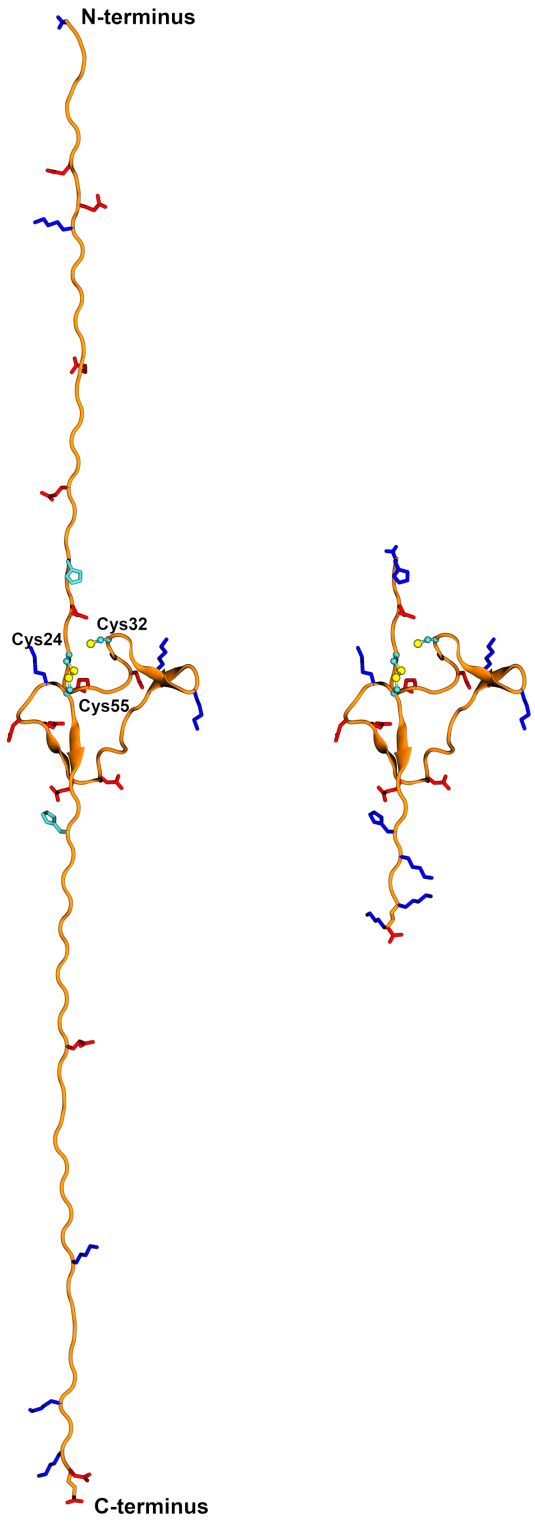

Figure 10.2.: Left: The complete I27* domain after pulling the termini in opposite directions. Right: The truncated I27* domain with mutations introduced at positions 62, 64 and 65. Negatively charged residues – red, positively charged residues – blue, neutral histidines in the full I27* – cyan (these are protonated thus positively charged in the truncated I27*).

Without any treatment of charges, the truncated I27*, which consists of the residues 20–65 and has both termini ionized, both the N-terminal and C-terminal chains carry one negative charge each, $Q_{\mathrm{N-term}} = Q_{\mathrm{C-term}} = -1$ e, so that $\Delta Q = 0$ e. The original charge difference and charge neutrality are achieved by (i) mutating the remote residues Asn62, Gln64, and Leu65 (located at least $6 \times 4 = 24$ Å away from the nearest reactive Cys55) to lysines, and (ii) protonation of the two histidines His20 and His61, see Fig. 10.2 and Tab. 10.1. After these modifications, the truncated I27* domain exhibits $Q_{\mathrm{N-term}} = 0$ e and $Q_{\mathrm{C-term}} = +3$ e, so that the charge difference of $\Delta Q = -3$ e is restored.

The protein was centered in a rectangular box sized $15.0 \times 4.8 \times 4.8$ nm^3 and solvated with 11 125 water molecules.

MM equilibration.
Prior to QM/MM simulations, the structures were equilibrated with classical force field molecular dynamics using Gromacs 5.0.1 patched with Plumed 2.1.1.[122,123]. The AMBER99SB-ILDN forcefield[185] and TIP3P water were used. Periodic boundary conditions were employed; electrostatics treated with the particle–mesh Ewald method. Lennard-Jones interactions were cut-off at 1 nm and the neighbour list updated every 10 MD steps. The leap-frog integrator was used with a time step of 2 fs. Initial velocities of the atoms were assigned from the Maxwell-Boltzmann distribution at 10 K and the system was heated up to 300 K linearly over an interval of 10 ps. Subsequently, an NVT equilibration with the Bussi thermostat[126] at 300 K was performed over 100 ps, followed by an NPT equilibration with the Parrinello–Rahman barostat[186] at 1 bar over 1 ns. During both steps, harmonic position restraints were applied to the heavy atoms of the peptide with a force constant of 1000 kJ mol^{-1} nm^{-2}.

QM/MM equilibration.
Next, QM/MM equilibrations over 100 ps were performed with Gromacs 5.0.1 including a local DFTB3 implementation,[130] additionally patched with Plumed 2.1.1. The QM region comprised the side chains of Cys24, Cys55 and Cys32 up to Cβ. This choice was motivated by the lack of any electronic effects (like, e.g., coordination) to other amino-acid side chains, charge transfer from/to other side chains or other phenomena calling for additional side chains or waters being included in the QM region, and also by the need for efficient computation required to achieve microsecond sampling. Bonds between Cα and Cβ were treated with the link atom approach, i.e. the QM region is capped with a hydrogen atom placed at a fixed position along the bond. In total, the QM region consisted of 15 atoms described with the semi-empirical density functional theory method DFTB3 and 3OB parameters.[54,135] The rest of the system was described with the AMBER99SB-ILDN forcefield[185] and TIP3P water. The previously applied positions restraints were lifted and the time steps reduced to 0.5 fs. Temperature and pressure were kept at 300 K and 1 bar with the Nosé-Hoover thermostat and the Parrinello–Rahman barostat, respectively. Additionally, the two centers of mass of the terminal amino acids were pulled away from each other along the x-axis with a constant force of 500 kJ mol^{-1} nm^{-1} (830 pN). Electrostatic interactions between the rather localized negative charge of the QM region and the MM system were scaled down by the factor of 0.75 corresponding to the inverse square root of the optical dielectric constant. This is an effective approach to compensate for the missing

Table 10.1.: Amino acid sequence of the full I27* domain and the truncated I27* domain considered in this work. Negatively charged residues – red, positively charged residues – blue, disulfide-bonded cysteines – brown.

#	full	trunc.	#	full	trunc.	#	full	trunc.
			20	HIS	HIS	66	GLY	
			21	PHE	PHE	67	MET	
			22	GLU	GLU	68	THR	
			23	ILE	ILE	69	GLY	
			24	CYS	CYS	70	GLU	
			25	LEU	LEU	71	VAL	
			26	SER	SER	72	SER	
			27	GLU	GLU	73	PHE	
			28	PRO	PRO	74	GLN	
			29	ASP	ASP	75	ALA	
			30	VAL	VAL	76	ALA	
			31	HIS	HIS	77	GLN	
			32	CYS	CYS	78	THR	
			33	GLN	GLN	79	LYS	
			34	TRP	TRP	80	SER	
			35	LYS	LYS	81	ALA	
			36	LEU	LEU	82	ALA	
			37	LYS	LYS	83	ASN	
			38	GLY	GLY	84	LEU	
			39	GLN	GLN	85	LYS	
			40	PRO	PRO	86	VAL	
			41	LEU	LEU	87	LYS	
			42	ALA	ALA	88	GLU	
-3	GLY		43	ALA	ALA	89	LEU	
-2	ALA		44	SER	SER			
-1	MET		45	PRO	PRO			
0	ALA		46	ASP	ASP			
1	LEU		47	CYS	CYS			
2	ILE		48	GLU	GLU			
3	GLU		49	ILE	ILE			
4	VAL		50	ILE	ILE			
5	GLU		51	GLU	GLU			
6	LYS		52	ASP	ASP			
7	PRO		53	GLY	GLY			
8	LEU		54	LYS	LYS			
9	TYR		55	CYS	CYS			
10	GLY		56	HIS	HIS			
11	VAL		57	ILE	ILE			
12	GLU		58	LEU	LEU			
13	VAL		59	ILE	ILE			
14	PHE		60	LEU	LEU			
15	VAL		61	HIS	HIS			
16	GLY		62	ASN	LYS			
17	GLU		63	CYS	CYS			
18	THR		64	GLN	LYS			
19	ALA		65	LEU	LYS			

electronic polarization of the MM environment as recommended by Stuchebrukhov.[187,188]

QM/MM production simulation.
Finally, the 334 force-clamp simulations were performed over 20 ns each with the same setup. When a disulfide reaction occurred, the simulation stopped due to the protein termini leaving the simulation box at both sides. Thus, instead of a theoretical maximum of ca. 6.68 µs, the total simulation time was ca. 5.7 µs. Snapshots of the trajectories were saved every 0.5 ps.

Analysis.
Distances and angles between the sulfurs were measured with Plumed[123] in all trajectories. Charges of the QM atoms were calculated with DFTB+.[132] Additionally, the electrostatic potential arising from the MM environment and by the QM sulfur atoms on each QM sulfur atom was calculated.

The electrostatic potential on the sulfur atoms arising from all of the MM atoms is calculated as

$$V_{\mathrm{E}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \sum_i \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|}$$

(q_i – charge of MM atom i, $\mathbf{r}_i$ – coordinates of MM atom i, $\mathbf{r}$ – coordinates of the atom on which the ESP is calculated).

The ESP caused on a QM sulfur atom by another QM atom corresponds to the DFTB Hamilton shift contribution stemming from the interaction between the atoms,

$$\Phi_a = \Delta q_b \gamma_{ab} + \frac{2}{3}\Delta q_a \Delta q_b \Gamma_{ab} + \frac{1}{3}\Delta q_b^2 \Gamma_{ba}$$

(a – atom on which the ESP is calculated, b – atom which causes the ESP, Δq_a and Δq_b – charges of atoms a and b), where γ_{ab} and Γ_{ab} are analytical functions of interatomic distance; see Ref. [135] for details.

10.2.2. QM/MM metadynamics of disulfide shuffling in I27*.

Well-tempered multiple-walker metadynamics[59,60,131] QM/MM simulations of the nucleophilic attack of S32 on S24 and S55 were performed to obtain the potentials of mean force of the reactions. The simulations were performed with a local version of Gromacs 2020 patched with Plumed 2.5.1 and interfaced with DFTB+ 19.1. Starting structures for the metadynamics simulations were selected from the 334 QM/MM equilibrated structures matching the following criteria: (a) a reaction occurred in the free QM/MM molecular dynamics simulation with this starting structure; (b) the S_{nuc}–S_{ctr}–S_{lg} angle is greater than 130°. We chose 12 different structures for the S32–S24 reaction setup (I) and 12 different structures for the S32–S55 reaction setup (II). Both metadynamics simulations used 24 walkers, thus the starting structures were duplicated. Each walker was simulated over 2 ns, thus 48 ns in total. A time step of 1 fs (leap-frog integrator) was employed and a temperature of 300 K (Bussi thermostat) and a pressure of 1 bar (Parrinello–Rahman barostat) were maintained. Electrostatic interactions between the QM and MM regions were

scaled down by a factor of 0.75. The S_{ctr}–S_{lg} and S_{nuc}–S_{ctr} distances were used as collective variables (CV) to drive the reactions, i.e. (I) S24–S55 and S32–S24, and (II) S24–S55 and S32–S55 Gaussian biasing potentials with an initial height of 1.2 kJ/mol and a width of 0.2 Å were deposited every 1 ps in each walker with a bias factor of 20. The bias exchange period of the walkers was set to 1 ps.

Additional restraints.

The configurational space of each reaction was reduced by applying harmonic restraints to S_{ctr}–S_{lg} distances > 3.5 Å and S_{nuc}–S_{ctr} distances > 10 Å with a force constant of 100 000 kJ mol^{-1} nm^{-2}. Additionally, the S_{nuc}–S_{ctr}–S_{lg} angle was restrained to values > 130° with a force constant of 100 000 kJ mol^{-1} rad^{-2}. Since metadynamics puts biases on both distances, the disulfide bond will extend over time and eventually break even without the sulfur anion S_{nuc} being close enough for a reaction. Hence, additional restraints were applied to the sum of switching functions applied to the three sulfur–sulfur distances, in order to avoid bond breaking while the sulfur anion is too far away. The switching functions were defined as

$$s(r_{ij}) = \frac{1 - \left(\frac{r_{ij}}{r_0}\right)^n}{1 - \left(\frac{r_{ij}}{r_0}\right)^m} \tag{10.1}$$

with the parameters taking values of $r_0 = 2.9$ Å, $n = 10$ and $m = 20$ for all considered combinations; $s_S(r_{S_{32}S_{24}})$, $s_S(r_{S_{32}S_{55}})$ and $s_S(r_{S_{24}S_{55}})$. The parameters where chosen in a way, so that the restraints do not interfere with the transition state. When the sum of all three coordination number was > 1.82, i.e. a disulfide bond length of ca. 2.3 Å without the sulfur anion in proximity, harmonic restraints with a force constant of 20 000 kJ mol^{-1} nm^{-2} were applied to avoid further elongation of the bond. In spite of the accumulated sampling of 48 ns, this simulation failed to converge.

10.2.3. Metadynamics simulation of disulfide shuffling in a symmetric aqueous model system

We performed QM/MM metadynamics simulations of a system composed of a dimethyl disulfide molecule and a methylthiolate anion using DFTB3 with the 3OB parameter set. An additional, artificial ESP of either −0.5 V, −0.25 V, 0 V, +0.25 V, or +0.5 V was imposed on one of the sulfur atoms. The simulations were performed with a local version of Gromacs 2020 patched with Plumed 2.5.1 and interfaced with DFTB+ 19.1.[122,123,132]

Setup.

First, the system was put in a rectangular periodic box of 3.0 × 3.0 × 3.0 nm^3, solvated with 877 TIP3P waters, and neutralized with one sodium ion. Subsequently, an energy minimization with the steepest descent methods was conducted with GROMACS/DFTB+, followed by an NVT equilibration with the Bussi thermostat at 300 K over 100 ps. For the NVT equilibration the leap frog integrator was used with a time step of 1 fs. Periodic boundary conditions were set and electrostatics were treated with particle–mesh Ewald. Lennard-Jones interactions were cut-off at 1 nm and the neighbour list was updated every 10 MD steps. The 15 atoms of the dimethyl disulfide molecule and methylthiolate were

treated with QM and the rest of the system with MM. Electrostatic interactions between the charged QM region and the MM system were scaled down by the factor of 0.75. The sulfur–sulfur distances were restrained to values smaller than 6 Å with a force constant of 100 000 kJ mol^{-1} nm^{-2} to keep the molecules together and to reduce the configurational space for the reaction. Additionally, the distances between the sulfurs and the sodium ion were restrained to values greater than 12 Å with a force constant of 100 000 kJ mol^{-1} nm^{-2}.

Metadynamics.
Subsequently, the potential of the mean force of the disulfide shuffling was obtained with well-tempered multiple walker metadynamics.[59,60,131] We used 24 walkers, each simulated for 10 ns at 300 K with the Bussi thermostat and at 1 bar with the Parrinello–Rahman barostat, yielding a total simulation time of 240 ns. All other settings were the same as for the NVT equilibration. The three distances between the sulfurs were used as collective variables (CV) to drive the reactions. A gaussian potential with an initial height of 0.5 kJ mol^{-1} and a width of 0.2 Å was deposited every 250 fs along the trajectory. Deposited biases from all other walker were read every 500 fs.

Additional restraints.
The configurational space of each reaction was reduced by means of harmonic restraints to S_{ctr}–S_{lg} and S_{nuc}–S_{ctr} distances above 6 Å with a force constant of 100 000 kJ mol^{-1} nm^{-2}. Analogously to the I27* metadynamics simulation, the sum of the three sulfur-sulfur coordination numbers was retrained to > 1.82 using the same parameters and force constant. Also, in pilot metadynamics simulations, the molecules irreversibly reacted to one of several chemically non-sensical species whenever all three sulfur–sulfur distances were very short (below 3 Å), i.e. in a triangular configuration. These structures lie very high in energy and thus are irrelevant for the investigated disulfide shuffling. To prevent such erroneous reactions, restraints were placed on the coordination numbers that were introduced for every sulfur atom as the sum of the switching functions applied to the distances from each of the two other sulfurs, e.g., $c(S_1) = s(S_1{-}S_2) + s(S_1{-}S_3)$. Each of the three coordination numbers was restrained to values below 1.8 with a force constant of 50 000 kJ mol^{-1} nm^{-2}, which penalizes triangular structures with short S–S distances. The sum of all three sulfur–sulfur distances was restrained to values above 9 Å with a force constant of 100 000 kJ mol^{-1} nm^{-2}, which also prevents the three sulfurs from approaching too closely.

In other high-energy conformations, the sulfur anion came very close to the carbons of the dimethyl disulfide and deprotonated them. Thus, additional restraints were employed to avoid such occurrences. The number of bonded hydrogen atoms was defined as the sum of the switching functions $s(C{-}H)$ with $r_0 = 1.3$ Å, $n = 45$ and $m = 90$, for each carbon atom separately. Each of these quantities was restrained to values above 2.5 with a force constant of 50 000 kJ mol^{-1} nm^{-2}. All non-covalent sulfur–carbon distances were restrained to values above 3 Å with a force constant of 50 000 kJ mol^{-1} nm^{-2}. Sulfur–hydrogen distances further than two covalent bonds away, were also restrained to values above 3 Å with a force constant of 50 000 kJ mol^{-1} nm^{-2}.

10.3. Results

10.3.1. Detailed view of the approach of the free thiolate

It was previously pointed out that spatial accessibility controls the reactivity on the disulfide bond in I27*. Therefore, it appears necessary to analyze how often and how closely S32 approaches S24 or S55 during the simulations, in detail. To this end, we obtained a 2D histogram of the distances S32–S24 and S32–S55 from all of the 334 QM/MM molecular dynamics simulations. The interval of distances from 2.4 to 30 Å was divided into bins 0.1 Å wide, and the 2D histogram was then converted to a potential of the mean force (PMF). The resulting PMF for S–S distances up to 10 Å is shown in Fig. 10.3 together with exemplary S–S–S configurations and exemplary pathways. The histogram and PMF over the full range of distances are shown in Fig. C.1.

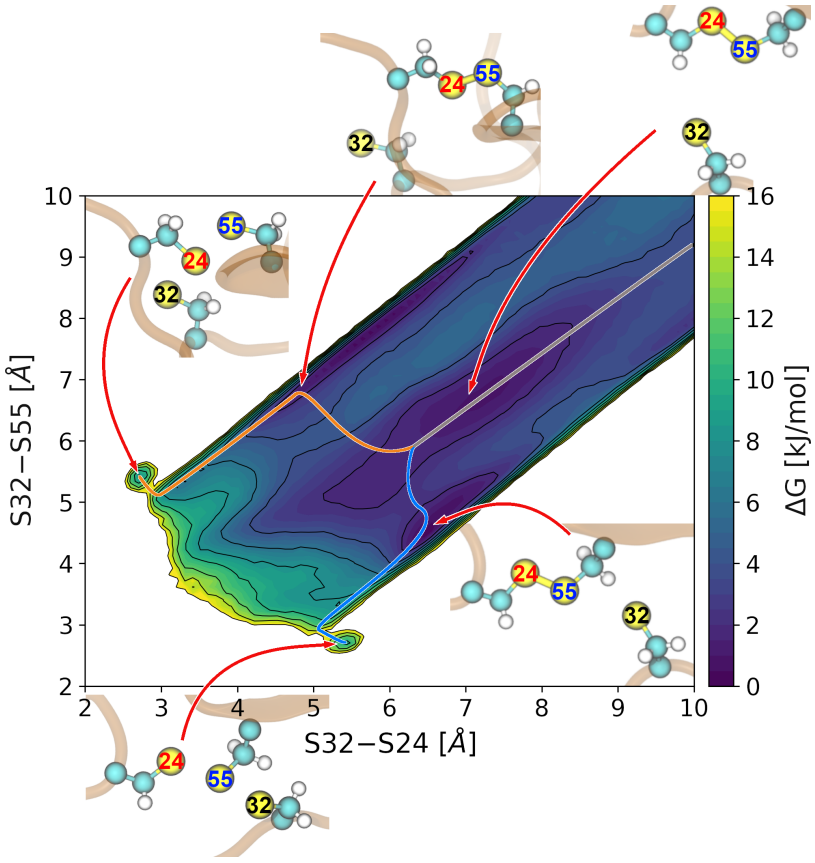

Figure 10.3.: Potential of the mean force as function of the S32–S24 and S32–S55 distances, with the S24-S55 distances integrated out. Exemplary pathways for a reaction with S24 (orange) and with S55 (light blue) are drawn on the surface coming from large distances (grey). Contour lines are drawn every 2 kJ/mol.

There are three minima of an equal depth that was set to 0 kJ/mol. The long, narrow minimum at the upper edge of the energy profile corresponds to nearly-linear S32–S24–S55 configurations, while the similar minimum at the lower edge corresponds to nearly-linear S32–S55–S24 configurations. The third deep minimum is found around the distances

S32–S24 and S32–S55 of 7 Å, and corresponds to a triangular configuration with S32 located in similar distances from S55 and S24.

Two shallower minima with free energies of ca. 10 kJ/mol are located at the distances S32–S24/S32–S55 of 2.7/5.4 Å and 5.4/2.7 Å. These correspond to the transition state structures of the two disulfide exchange reactions. As such, these should be saddle points on the free energy surface, and the observed shallow minima are an artifact of DFTB3/3OB, which underestimates the energy and overestimates the bond lengths of trisulfide species, as discussed previously.[189] Nevertheless, this systematic error affects both reactions, S32→S24 and S32→S55, to exactly the same extent, therefore any qualitative conclusions from this study will be unaffected. The height of energy barriers to both reactions is similar, ca. 15 kJ/mol. A tiny difference in barrier height is expected considering the rather small regioselectivity that was observed in experiments as well as in simulations, however it appears impossible to resolve using free MD simulations like here.

To learn how often S32 approaches S24 or S55, the histogram in Fig. 10.3 was split into 2 regions, the "upper" region in which S32 is closer to S24, and the "lower" region where S32 is closer to S55. All probabilities in the "upper" region were summed up and converted to a free energy value, and the same was done for the "lower" region, see Fig. 10.4. In the "upper" region with 3 Å < |S32–S24| < 10 Å and |S32–S24| < |S32–S55|, S32 was considered to be nearer to S24. In the "lower" region with 3 Å < |S32–S55| < 10 Å and |S32–S24| > |S32–S55|, S32 was considered to be nearer to S55. We found $\mathcal{P}_{\text{lower}}/\mathcal{P}_{\text{upper}} = 1.4$, or in

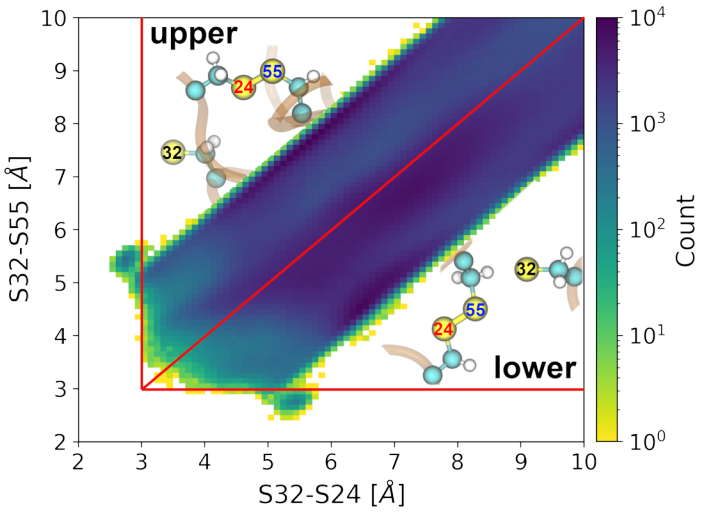

Figure 10.4.: In the upper region, S32 is nearer to S24 and in the lower region nearer to S55.

terms of free energy, $G_{\text{lower}} - G_{\text{upper}} = -0.8$ kJ/mol. This means that S32 is closer to S55 on average, therefore a reaction may occur more frequently with S55 than with S24, as stated previously in Ref. [177] on basis of simpler simulations.

To see how the distances between S32 and the disulfide bond correlate with the length of that bond, histograms of the distances S32–S24 and S32–S55 were generated for different S24–S55 bond lengths observed, see Fig. 10.5. All 2D histograms obtained are shown in Figs. C.2–C.4. It appears that S32 is increasingly more likely to be closer to S55 with

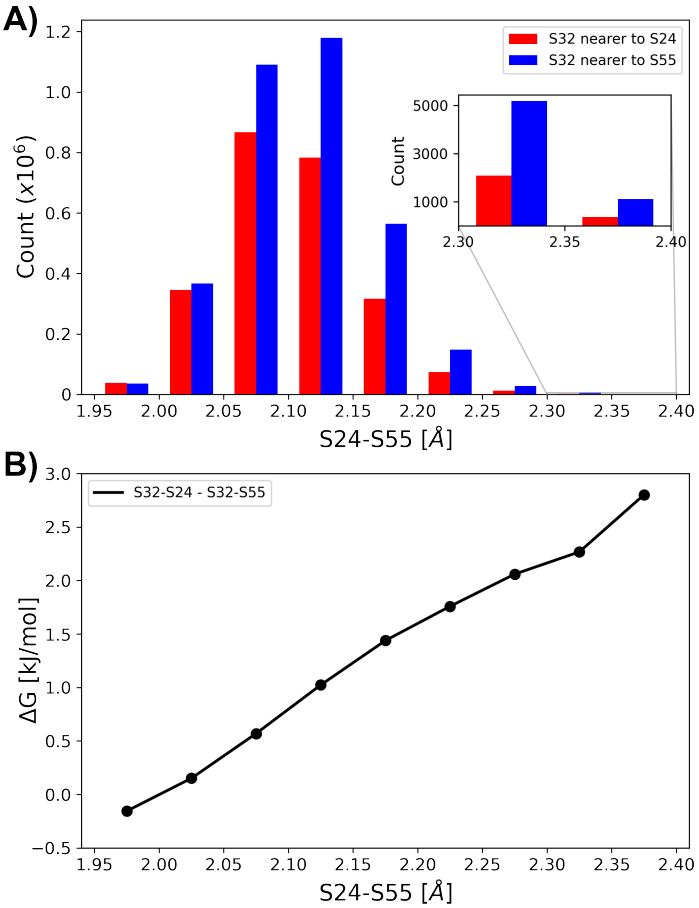

Figure 10.5.: A: Probabilities of finding S32 closer to S24 or S55, depending on the bond length S24–S55. B: The ratio of probabilities in each bin, converted to free energy difference.

increasing S24–S55 distance. Viewed from the other side: whenever S32 is closer to S55, a longer S24–S55 bond is favored. Consequently, it may be easier for the system to stretch the bond S24–S55 further to pass to a transition state. By contrast, whenever S32 is closer to S24, a shorter bond is favored, thus a transition state is less likely to form.

10.3.2. Metadynamics simulation of disulfide shuffling in I27*

Multiple-walker well-tempered QM/MM metadynamics simulations were performed with the intention to resolve the free energy profile of both thiol–disulfide shuffling reactions and to estimate the genuine energy barriers. However, the corresponding free energies do not converge even after a combined simulation time of 48 ns for each reaction, although the biases being added are very small. The energy profiles after a simulation time of 24 ns, 30 ns, 36 ns, 42 ns, and 48 ns are shown in Fig. 10.6 and the height of the barriers are in Fig. C.5. The barrier heights for S32→S24 fluctuate between 19–22 kJ/mol, and for S32→S55 between 22–27 kJ/mol. Nevertheless, they are in the same order of magnitude with those obtained from the 334 free QM/MM molecular dynamics, cf. Fig. 10.3. Such a barrier height is also compatible with the many occurrences of the reactions observed on a nanosecond

timescale. The location of free energy minima for S32 approaching S24 or S55 also agree with those obtained from the unbiased MD simulations well.

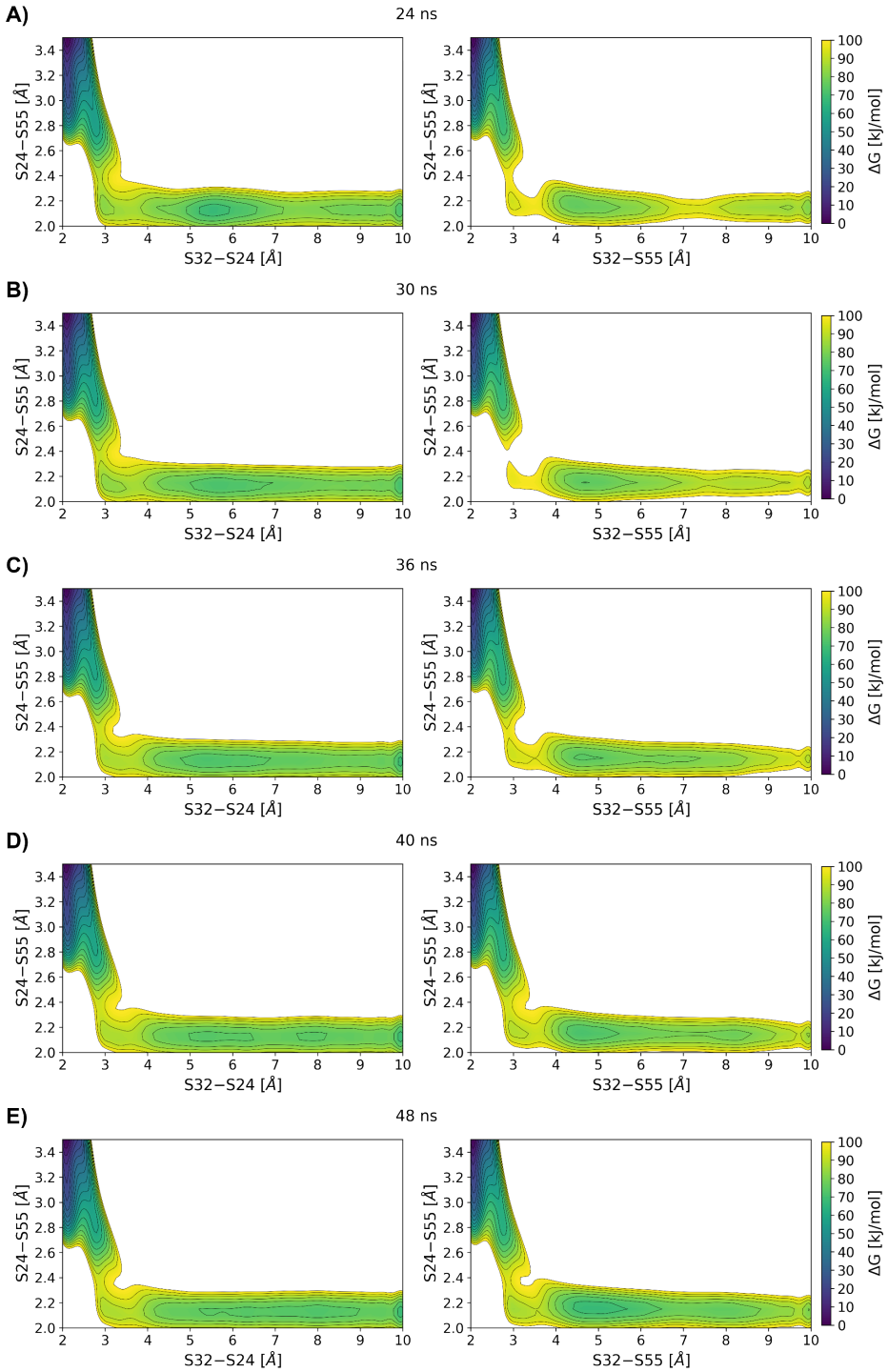

Figure 10.6.: Free energy profiles of the 2D metadynamics simulations of S32 reacting with S24 (left) and S55 (right) after a total simulation time of 24 ns (A), 30 ns (B), 36 ns (C), 40 ns (D), and 48 ns (E).

10.3.3. Analysis of observed reactions

QM/MM MD captures the experimentally observed regioselectivity.
We performed an ensemble of QM/MM force-clamp simulations of I27*, starting from 334 structures generated by Kolšek et al.[177] The termini of protein chain were pulled in opposite directions with a constant force of 500 kJ mol^{-1} nm^{-1} = 830 pN. A disulfide exchange reaction was possible by means of an attack of Cys32, present in the deprotonated thiolate form, on either Cys24 or Cys55. Each simulation was stopped after a disulfide exchange has taken place or after 20 ns, whichever occurred first, and the total simulation time was ∼5.7 µs.

In spite of the restricted time scale of the QM/MM simulations, a reaction occurred 66 times, with a preference for Cys32 attacking Cys55 (48 reactions) over Cys24 (18 reactions). The preference for Cys32 agrees with the experimental observations, and the Cys55/Cys24 ratio of 2.7 agrees is remarkably similar to the experimental ratio of 3.8. Now, the question arises why one of the reactions is favored over the other. In an attempt to answer this question, we analyzed selected structural and electrostatic parameters in the interval of 10 ps prior to the formation of the transition state in the trajectories where a reaction occurred.

Disulfide shuffling correlates with distances, angles, charges and ESP.
In every trajectory in which a disulfide shuffling occurred, the last 10 ps (20 snapshots) before the formation of a transition state were analyzed. The three distances between the sulfurs were measured, as well as the angle between the nucleophilic sulfur anion S32 (S_{nuc}), the central sulfur under attack (S_{ctr}), and the respective sulfur of the leaving group (S_{lg}). In addition, the Mulliken atomic charges of the sulfurs were recorded. To assess the influence of the molecular environment on the outcome of the reaction, the electrostatic potential (ESP) on each of the three QM sulfurs caused by all of the QM and MM atoms was monitored. The temporal course of the described quantities for all of the observed reaction is shown in Fig. 10.7. The mean values and standard deviations of these quantities are listed in Tab. 10.2.

Table 10.2.: Distances and angles between the three sulfur atoms, charges of sulfur atoms, and electrostatic potentials (ESP) on the sulfur atoms. S24 is S_{ctr} and S55 is S_{lg} in the reaction S32→S24; S24 is S_{lg} and S55 is S_{ctr} in the reaction S32→S55. All data given as mean value and standard deviation.

Reaction	S32→S24	S32→S55
|S32–S24| [Å]	3.78 (0.54)	5.48 (0.53)
|S32–S55| [Å]	5.65 (0.51)	3.68 (0.52)
|S24–S55| [Å]	2.14 (0.08)	2.14 (0.08)
Angle [°]	149 (19)	143 (18)
Q(S24) [e]	−0.04 (0.03)	−0.14 (0.04)
Q(S55) [e]	−0.12 (0.04)	−0.02 (0.04)
Q(S32) [e]	−1.09 (0.06)	−1.10 (0.05)
ESP(S24) [V]	−2.22 (0.35)	−1.91 (0.33)
ESP(S55) [V]	−1.94 (0.34)	−2.25 (0.36)
ESP(S32) [V]	−3.26 (0.35)	−3.07 (0.36)

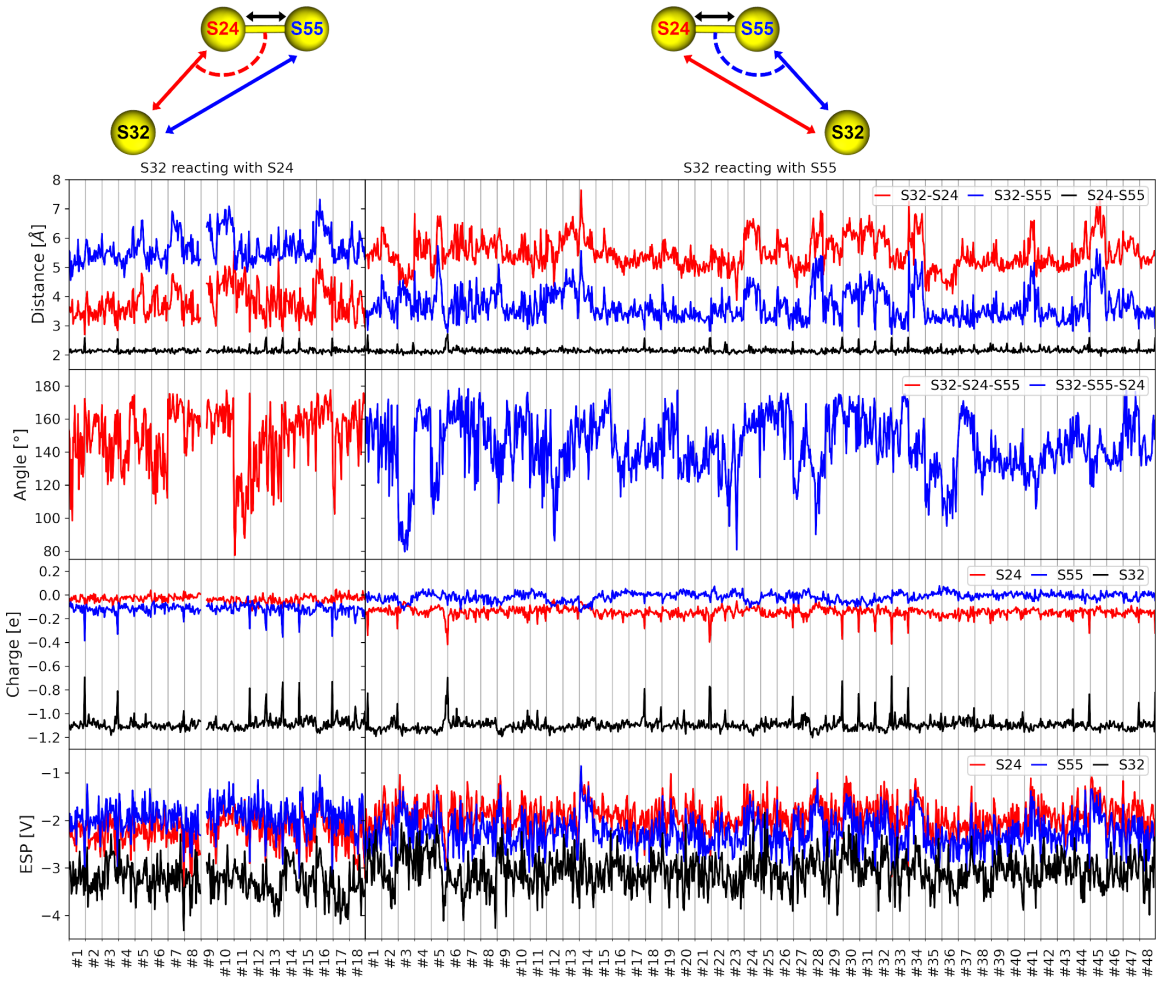

Figure 10.7.: From top to bottom: Distances and angles between the three sulfur atoms, charges of sulfur atoms, and the electrostatic potential (ESP) on each sulfur atom caused by all of the MM and QM atoms. Each column represents one occurrence of a disulfide shuffling reaction with either S24 (18 times, left) or S55 (48 times, right), showing data from the interval of 10 ps preceding the formation of the transition state. The simulation time of #9 on the left side was shorter than 10 ps. Peaks in the S_{nuc}–S_{ctr} distances, sulfur charges, and ESPs resemble structures where the transition state is approached but not fully formed yet.

Distances and angles prior to the reaction.
The distance between S_{nuc} (S32) and the respective attacked sulfur S_{ctr} (S24 or S55) fluctuates between 3–5 Å whereas S_{nuc} is further away from the leaving sulfur S_{lg}, at 4.5–7 Å. The TS is formed as soon as $|S_{nuc}–S_{ctr}|$ has decreased to ~2.75 Å and $|S_{ctr}–S_{lg}|$ has increased to ~2.75 Å, while $|S_{nuc}–S_{lg}|$~5.4 Å indicating a linear arrangement.[189] The temporal course of the distances for two example reactions, one with S24 and the other with S32, is shown in Fig. C.6 for the section of the trajectory immediately preceding and following the reaction.

The angle $S_{nuc}–S_{ctr}–S_{lg}$ oscillates between 80–180° but this range narrows down to 120–170° right before the formation of TS. At small S32–S24 distances below 3.4 Å, the preferred angle lies between 150–170°, whereas the preferred angle at small S32–S55 distances shows

more variance of 120–170°, see also histograms in Fig. 10.8. Thus, the S–S–S arrangement deviates further from linearity in the 10 ps prior to the disulfide exchange with S55, as compared to the same time frame preceding an exchange with S24. This observed larger flexibility supports the hypothesis that S55 is better accessible for a nucleophilic attack by S32 than S24.[177]

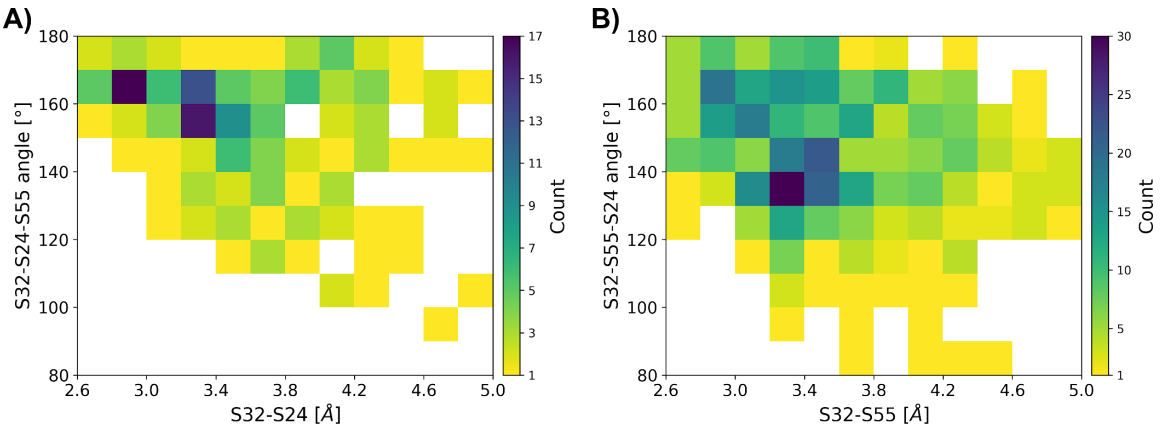

Figure 10.8.: Histograms of the distance and angle distribution based on snapshots 10 ps before a transition state is formed for S32⟶S24 (A), and S32⟶S55 (B).

Charges.
The target of the S_N2 attack, S_{ctr} is slightly more positively charged than S_{lg} in nearly all observed reactions, see Fig. 10.7. The S32 anion initially has a charge of ca. −1.1 e, and neutralizes as S32 approaches S_{ctr} gradually. In the transition state, the negative charge is equally distributed between S_{nuc} and S_{lg}. By contrast, the charge of S_{ctr} remains around zero the entire time. All in all, the charges of S32 and the respective S_{lg} correlate with the distance of the approaching nucleophile, S32, from the target, S_{ctr}. Also, the negative charge is transferred from S_{nuc} to S_{lg} during the reaction directly, without any transient accumulation on S_{ctr}, which was already observed in previous studies.[45,190,191]

Electrostatic Potentials.
The electron density on the individual atoms, expressed in terms of atomic charges in DFTB3, is determined to a large extent by the electric field experienced by the atoms. Therefore, in search for the mechanism that controls the disulfide shuffling, it is necessary to analyze the ESP on the sulfur atoms arising from their molecular environment. The ESP on S_{ctr} and S_{lg} is substantially negative due to the close proximity of the S32 anion. Since the distance $|S_{nuc}–S_{ctr}| < |S_{nuc}–S_{lg}|$, the ESP on S_{ctr} is generally more negative. than on S_{lg}, with few exceptions. In the transition state, the ESP on the S32 anion decreases as its charge is being transferred to S_{lg}, and the ESP on S_{lg} increases. Individual contributions arising from all MM atoms and the QM atoms are shown in Fig. C.7.

Influence of electrostatics on regioselectivity.
As mentioned above, $Q(S_{ctr}) > Q(S_{lg})$, and the negative charges of S_{nuc} and S_{lg} are interchanged without accumulating at S_{ctr} during the reaction. Thus, two assumptions can be made: (i) a more positive $Q(S_{ctr})$ favors the nucleophilic attack on S_{ctr} more; (ii) a more negative $Q(S_{lg})$ makes S_{lg} a better leaving group. These statements may be expressed in terms of ESP, with which the charges correlate. To investigate this, we calculated the differences $\Delta Q = Q(S_{ctr}) - Q(S_{lg})$ and $\Delta\text{ESP} = \text{ESP}(S_{lg}) - \text{ESP}(S_{ctr})$ for both reactions, and took averages over the intervals of 10 ps prior to the formation of the transition state. These results are visualized in Fig. 10.9.

It turns out that both ΔQ and ΔESP are larger for the reaction S32→S55. This means that S_{ctr} is, on average, a somewhat better target of an S_N2 attack, and S_{lg} is a better leaving group, in that reaction compared to S32→S24. Thus, electrostatic interactions contribute to the observed regioselectivity of the disulfide exchange reaction.

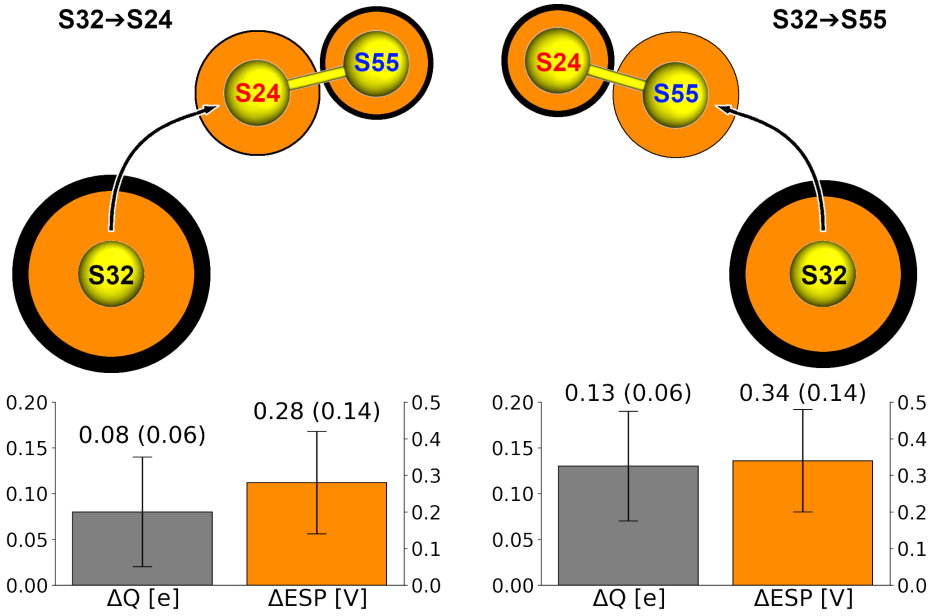

Figure 10.9.: Top: Atomic charges of the sulfur atoms and ESP arising from all of the QM and MM atoms, averaged over the respective ensembles of disulfide exchange reactions as observed in the QM/MM simulations. ESP coded by the radius of balls; charge coded by the outline thickness, scaled down by a factor of 3 for S32 for clarity. Bottom: Averages of charges and ESP represented by differences $\Delta Q = Q(S_{ctr}) - Q(S_{lg})$ and $\Delta\text{ESP} = \text{ESP}(S_{lg}) - \text{ESP}(S_{ctr})$.

10.3.4. Ensemble of starting structures – is there any bias?

To determine whether the choice of initial configurations leads to a bias on the reaction outcome, we examined the distribution of the S–S distances in the ensemble of starting structures as well as in the ensemble of structures obtained from the QM/MM equilibration simulations are shown in Tab. 10.3 and Fig. 10.10. In both ensembles, S32 is not necessarily significantly closer to either S24 or S55; rather, there are structures in which the distances

|S32–S24| and |S32–S55| are similar. Furthermore, in most of the structures, S32 is either too far from or not linearly aligned with S24 and S55, so a reasonable S–S–S transition state cannot form quickly. The distribution of S–S distances is even more scattered and shifted to larger distances after the initial QM/MM equilibration phase.

In the "start S24" batch (S32 closer to S24 in the starting structure), the reaction S32→S24 occurred 17× while S32→S55 occurred 22× for a total of 39 reactions. In the "start S55" batch (S32 closer to S55 in the starting structure), the reaction S32→S24 occurred only once while S32→S55 occurred 26× for a total of 27 reactions. Hence, there are two important observations: (i) even though "start S24" contains 14 fewer structures than "start S55", more reactions occurred; (ii) preference for S32→S55 is found even in "start S24" where S32 is closer to S55 initially. We can conclude that ratio of the reaction outcome is not influenced (biased) by the unequal numbers of starting structures in the ensembles.

Table 10.3.

	taken from Ref. [177]		after QM/MM equil.	
	start S24	start S55	start S24	start S55
|S32–S24| < |S32–S55|	126	0	109	29
|S32–S24| > |S32–S55|	0	152	40	129
|S32–S24| ≈ |S32–S55|	34	22	11	16

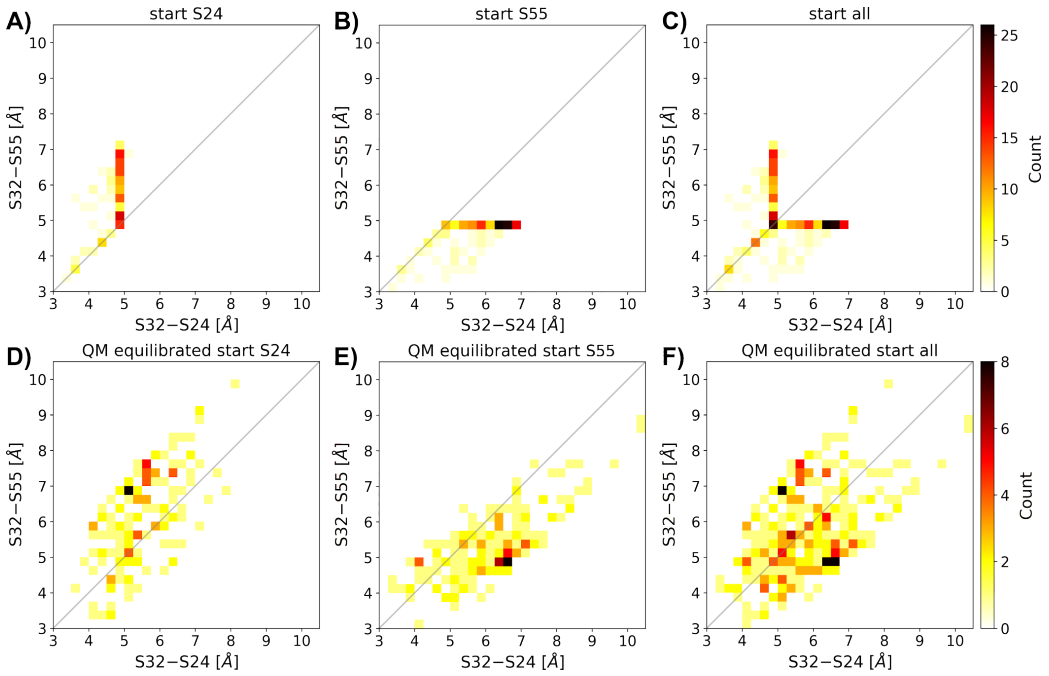

Figure 10.10.: Histograms of the S32–S24 and S32–S55 distances in the ensemble of structures taken from Ref. [177]. (A–C) and among the structures yielded by QM/MM equilibrations (D–F). Left: 160 structures where S32 is closer to S24, center: 174 structures where S32 is closer to S55, right: histograms over the whole ensemble of structures. The bin width is 0.25 Å.

10.3.5. Effect of external electric potential on the reaction

According to our above analysis of charges and ESP, the polarization of the nucleophile, of the target as well as that of the leaving group dictates whether and how an S_N2 reaction proceeds. The question arises if this polarization of the disulfide is a consequence of or a reason for the preferential attack. The polarization itself is driven by the electrostatic interactions with the surrounding atoms, which may be quantified by the ESP. To investigate how the electrostatics influence disulfide exchange reactions in general, additional QM/MM metadynamics simulations of a model system with an external electrostatic field of varying strength were performed. The system comprised a dimethyl disulfide molecule and a methylthiolate anion in aqueous solution.

An advantage of this small, simple model is that the PMF is completely symmetric as long as no external potential is applied, and this knowledge may be used for a convenient convergence check. The minimum energies for bonds between S_1–S_2 and S_1–S_3 are 0 kJ/mol, and for a bond between S_2–S_3 the energy is 2 kJ/mol. The energy barriers to the three disulfide exchange reactions lie in the range of 49–52 kJ/mol. All this illustrates the good convergence of the simulation, with a statistical error of at most 2 kJ/mol.

Simulations were performed with an additional, external ESP of −0.50 V, −0.25 V, 0 V, +0.25 V, and +0.50 V, respectively, imposed on the atom S_1; this additional potential will be denoted ESP_{ext} in the following. The simulation setup was designed to sample all three disulfide bonding patterns: S_1–S_2, S_1–S_3, and S_2–S_3, with the respective remaining sulfur atom in the deprotonated reduced (anionic) state. The reaction energies and height of energy barriers to disulfide shuffling are plotted in Fig. 10.11. The 2D representations of the PMF expressed as function of the S_1–S_2 and S_1–S_3 distances (with the S_2–S_3 distance integrated out) are shown in Fig. C.8 together with exemplary molecular structures, and the numerical values of energy barriers are shown in Tab. C.1.

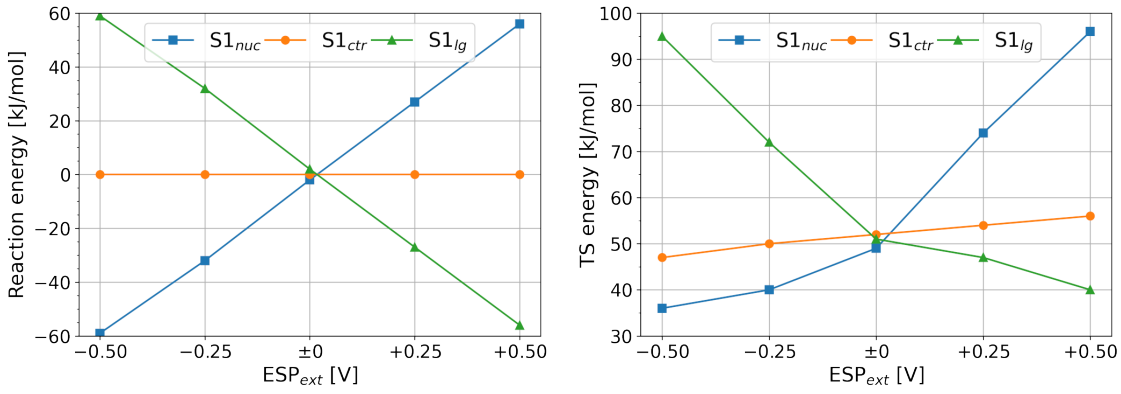

Figure 10.11.: Thermodynamics and kinetics of the disulfide exchange reaction in the model system with additional external potential, ESP_{ext} imposed on atom S_1. Left – reaction energy; right – height of energy barrier. The data points are labeled by the role of S_1 in the reaction. Note: all three respective data points for ESP_{ext} = 0 V correspond to the same situation.

The reaction energies in Fig. 10.11 (left) exhibit clear trends. First, ESP_{ext} applied to the sulfur being attacked, S_{ctr}, has no influence on the reaction energy because the reactant and the product are identical – one of the disulfide-bonded atoms carries ESP_{ext}. The other two cases are in fact one: a reaction with ESP_{ext} on S_{nuc} is the reverse process to a reaction with ESP_{ext} on S_{lg}. Thus, the reaction energy of one equals the negative of the reaction energy of the other. The stabilization or destabilization of the negative charge of the thiolate (which is S_{nuc} prior to the reaction, and S_{lg} thereafter) by ESP_{ext} may be considered to rationalize the numerical value of reaction energies: The contribution to total energy from the interaction of that charge $Q(S^-)$ with the electrostatic environment (represented here by ESP_{ext}) is well approximated as $E' = ESP_{ext} \cdot Q(S^-)$. Since $Q(S^-) \approx -1.1$ e always, E' is a linear function of ESP_{ext}. As an example, for S_1 being S_{nuc} and $ESP_{ext}= -0.5$ V, $E' = -0.5 \cdot (-1.1)$ eV $= +0.55$ eV $= +53$ kJ mol^{-1}, in a good agreement with the actual observation.

The reaction energies are not quite important in the context of the current work, however. Due to the stretching force applied on I27*, the new free thiolate is immediately pulled away from the newly formed disulfide bond as soon as the first disulfide exchange has taken place. Therefore, the energy of the product and the thermodynamics of the reaction do not play any role. The crucial phenomenon will rather be the effect of ESP, or ESP_{ext} in the study of the model system, on the reaction rates.

Let us turn our attention to the heights of energy barrier in Fig. 10.11 (right). The barrier heights with ESP_{ext} on S_{nuc} or on S_{lg} change in a way that is very similar to the reaction energies: the barrier is elevated whenever the reaction energy is positive, while lower barriers are seen in cases that have negative reaction energies. This is a simple consequence of the shape of the corresponding energy landscapes as depicted in Fig. C.8. Most interesting in the current context will be the case where S_1 is S_{ctr} because this is the kind of data that we have measured in our simulations of I27*. There is a roughly linear dependence of energy barrier on ESP_{ext}, with a slope of ca. 10 kJ mol^{-1} V^{-1} and positive ESP_{ext} giving higher barriers. This is explained easily as positive ESP_{ext} leads to a decrease of charge of S_{ctr}, which thus becomes a worse nucleophilic target, and the other way around for negative ESP_{ext}.

It has to be pointed out that a certain ESP_{ext} applied on S_1 in the model system shall have the same effect on the electronic structure of the disulfide bond as the same value of ΔESP in the simulations of I27*. Recall that the ΔESP values found for the reactions were 0.28 and 0.34 V, respectively. The effect of this difference may be compared to the difference of $ESP_{ext} = 0$ and 0.06 V: Multiplication by the slope of the dependence of energy barrier on ESP_{ext} leads to the difference of energy barriers of ca. 0.6 kJ mol^{-1}, which would cause a ratio of reaction rates of ca. 1.3 from Arrhenius' equation. This factor contributes to the above observed ratio of reaction rates of 2.7, while the remainder of this ratio (of ca. 2) is probably due to other effects like spatial accessibility as discussed previously.[177]

10.4. Concluding Discussion

The disulfide shuffling in the I27* domain was investigated by generating an extensive ensemble of trajectories using unbiased semiempirical QM/MM MD simulation. Of two possible disulfide shuffling reactions, S32→S55 was preferred over S32→S24, in agreement with experimental observations as well as previous computational results.

Next, we asked what structural factors contribute to the preferential attack. The distances and angles between the cysteine sulfur atoms in the trajectories were measured. It was found that S32 can approach S55 over a wider range of angles than S24, therefore S55 is the more easily accessible target of a nucleophilic attack. Further, S32 is located more often closer to S55 than to S24, making a nucleophilic attack on S55 more likely. All that agrees with the previous observations by Kolšek et al.[177]

Clearly, steric factors play a very important role in disulfide shuffling, but this may not be the complete explanation. Rather, electrostatic interactions may contribute to the reaction control. Thus, we decided to analyze the electron density of the trisulfide system as well as electrostatic interactions in the protein, to see if we find any significant effects. Note that electron density is represented by Mulliken atomic charges within DFTB.

S55 in the role of nucleophilic target carried a more positive charge than S24, and S24 carried a more negative charge as a leaving group than S55 did. This means that S55 is the better nucleophilic target, and S24 the better leaving group of the two. The charges were averaged over two separate ensembles of simulations, and thus there is no bias towards the ensemble with a larger number of simulations.

The observed difference of atomic charges may be accounted to the electrostatic potentials on the sulfur atoms caused by the molecular environment (amino acid side chains, peptide backbones and solvent), which are slightly different for each sulfur atom. Consequently, it is the electrostatic effects of the molecular environment that support the reaction S32→S55 more than S32→S24. This is an additional explanation of the outcome of the force-clamp experiments on I27*, in addition to the previous concept of regioselectivity via accessibility.

In terms of the transition state theory, the steric factors make the approach frequency and thus the pre-exponential factor higher for the reaction S32→S55. Also, the different polarization of electron density results in a lower energy barrier for the same reaction. These two effects act in the same direction, favoring the reaction S32→S55.

A possible electrostatic control of regioselectivity was demonstrated on a model system featuring a symmetric free energy landscape of the disulfide exchange reactions. As soon as an external electric potential is imposed on one of the sulfur atoms, the charges of the sulfurs change, and consequently, so do the free energies: A negative applied ESP results in a more positive charge, which makes the touched atom a better nucleophilic target but a worse leaving group. On the other hand, a positive applied ESP results in a more negative charge, making the atom a better leaving group but a worse nucleophilic target.

We provided a quantitative measure of this effect on the reaction energies and barriers. Electrostatic potential arising from the protein and water environment may polarize the disulfide bond slightly, such that the nucleophile attacks one of the sulfur atoms preferentially. Thus, electrostatics may break the symmetry of the disulfide system. This either induces regioselectivity, or contributes to the regioselectivity due to steric factors.

This model study shows how an external electric field affects the kinetics of disulfide shuffling. The magnitude of ESP applied here, in the order of tenths of volt, corresponds to the differences of potentials observed in protein systems like the I27* domain. In a protein, the "external field" arises from the protein and solvent environment – the surrounding amino acid side chains, peptide backbone as well as any water present. Such an electric field brings on a variation of energy barriers of few kJ mol^{-1}. This modulates the reaction rates by a small factor, and it turns out that the kinetics of disulfide shuffling in proteins is affected by electrostatic effects of the close environment of the disulfide moiety.

Active sites of enzymes and other proteins feature perfectly positioned and patterns of specific interactions. The case of disulfide exchange reaction investigated here is different but still remarkably similar in the working principle: Even though there is no real active site, the selectivity of the reaction is still achieved through the interactions with the environment. All of this likely matters for the disulfide exchange reactions as known in proteins like VWF, integrins and others.

11. Accurate free energies for complex condensed phase reactions using an artificial neural network corrected DFTB/MM methodology

Chapter 11 is reprinted with permission from Ref. [192]:

- Gómez-Flores, C. L.; Maag, D.; Kansari, M.; Vuong, V. Q.; Irle, S.; Gräter, F.; Kubař, T.; Elstner, M.; Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural Network Corrected DFTB/MM Methodology., J. Chem. Theory Comput. 2022.

Author Contributions: This work was done in cooperation with Claudia Leticia Goméz-Flores and Mayukh Kansari. Claudia Leticia Goméz-Flores optimized and trained the neural network. Denis Maag generated the I27* test set and performed QM/MM simulations. Mayukh Kansari reparameterized the sulfur-sulfur parameters. Tomáš Kubař implemented the neural network in DFTB+.

11.1. Introduction

As for other S_N2 reactions, a hydrophobic environment is catalytic for thiol–disulfide exchange reactions because the charge of the sulfurs is more delocalized.[43] In the gas phase, the charge is completely delocalized along the three sulfurs when the molecules are symmetric and form a nearly linear trisulfide complex.[44,45] In a polar environment, e.g. in water and/or a protein, the charge is more localized. Consequently, the thiolate and the disulfide states are stabilized whereas the trisulfide state is the transition state, as sketched in the energy scheme in Fig. 11.1.

Thus, the mechanism of thiol–disulfide exchange is more complicated than it seems at the first glance, and it turns out to be a major challenge for quantum chemical methods. A detailed study, benchmarking 92 density functionals for the thiolate-disulfide exchange between a methylthiolate and a dimethyldisulfide, was presented by Neves et al.[193] A proper inclusion of electron correlation is crucial for this reaction, as illustrated by the entirely different Hartree–Fock and CCSD(T) energetics of the reaction. While the reference method MP2/aug-cc-pVTZ yielded a linear structure with S–S bond lengths of 2.40 and 2.42 Å (see

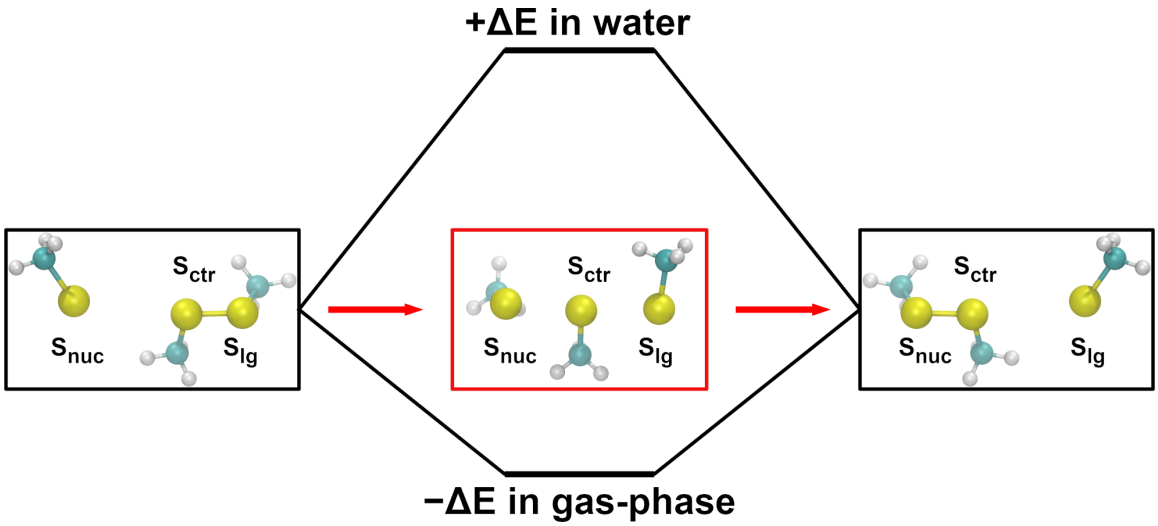

Figure 11.1.: Thiol-disulfide exchange between a methylthiolate and a dimethyldisulfide (black boxes). The trisulfide state (red box) is the global minimum in gas-phase and the transition state in aqueous solution.

the trisulfide structure in Fig. 11.1), both LDA and GGA functionals showed deviations from linearity and significantly longer bonds. The potential energy was compared in detail for the gas-phase situation: Potential energy differences depend sensitively on the level of theory and the basis set size. This is reflected by the energy differences (depths of the minimum) deviating among the different methods: For instance, CCSD(T)/aug-cc-pVTZ yields an energy difference of $\Delta E = -2.13$ kcal/mol in the gas phase, while $\Delta E = -5.05$ kcal/mol with MP2/aug-cc-pVDZ, and $\Delta E = -5.65$ kcal/mol with B3LYP/TZVP. An error of up to 10 kcal/mol was reported for the LDA functionals, while the GGAs still show errors of more than 5 kcal/mol (e.g. 6.7 kcal/mol with PBE). Including the exact exchange improves the situation, however the error of the widely used B3LYP functional is still ca. 3 kcal/mol. In fact, the performance of hybrid functionals depends on the amount of exact exchange sensitively, and the optimal amount was found to be 40–50%.

In solution, the symmetric conformation becomes a transition state with a significant barrier: CCSD(T)/aug-cc-pVTZ shows a barrier of 9.28 kcal/mol, which decreases with decreasing basis set size and level of theory to 6.24 kcal/mol with MP2/aug-cc-pVDZ. These energy estimates are based on continuum solvent models, which yield substantial contributions to the thermal free energy and zero-point energy of ca. 8 kcal/mol. Such free energy contributions in solution can be taken into account in a straightforward manner by sampling the configurational space using molecular dynamics (MD) simulations.

To study the reaction, hybrid quantum chemistry/molecular mechanics (QM/MM) approaches can be applied, describing the reactive part with quantum chemical methods (QM) methods and the remainder of the system with classical force fields, as introduced by Warshel & Levitt in 1976.[194] However, using ab initio or DFT approaches for the QM region, sampling is usually restricted and mostly done by searching reaction pathways, as done by Neves et al. where two consecutive thiol-disulfide exchanges in the enzymatic reduction of gluthathione disulfide by a disulfide isomerase protein were investigated.[195]

The reaction profile was calculated using one of the best performing DFT functionals from the previous study.[193] Barrier heights of 18.7 kcal/mol and 7.2 kcal/mol were estimated for the first and second disulfide exchange reaction, respectively.

As an alternative, fast semi-empirical (SE) methods, such as the Density Functional Tight-Binding (DFTB) approach, are capable of extensive sampling when applied in suitable QM/MM frameworks. However, the lower computational cost of DFTB comes at the expense of lower accuracy. DFTB approximates and neglects certain energy contributions,[196] and the remaining ones are obtained using the DFT-PBE functional. While geometries are reproduced very well in most cases, the accuracy of reaction energies, barriers and vibrational frequencies is generally lower than for DFT methods. Hence, DFTB is far from chemical accuracy due to quantitative errors.[54] Moreover, there are qualitative errors, such as the inaccurate transition states in thiol–disulfide exchange reactions.[189] In order to deal with qualitative errors, specific reaction parametrizations (SRP)[196] can be developed which usually leads to good description of the problem at hand.

Another approach for the description of potential energy surfaces (PESs) are machine learning (ML) potentials. The first ML potential was introduced by Doren et al. in 1995, who fitted a DFT PES with the help of an Artificial Neural Network (ANN).[197] However, the model was limited to a few atoms, and it would take 12 more years before larger atomic systems could be described. In 2007, Behler and Parrinello presented an ML potential in which the total energy is constructed as the sum of the atomic energies and the environment of each atom is considered within a cutoff radius.[67] Since then, many new developments have followed, such as the ANN TensorMol[198] model which also considers long-range electrostatics and van der Waals energies, or the SchNet[199] architecture which can be used as an energy-conserving force field in molecular dynamics simulations.

SE methods can also be combined with ML algorithms, which are designed correct the energy difference between the SE method and a high level QM method. This so-called Δ-ML approach was first incorporated in a QM/MM framework by Shen and Yang[200] in 2018 with an ANN to correct the PES of peptide building blocks. Since then, similar approaches have been further developed, for example by Böselt et al.[201], Gastegger et al.[202], Zeng et al.[203], or Pan et al.[204,205]

In this work, we aim at developing a Δ-ML based approach for the description of thiol–disulfide exchange within a QM/MM framework. This approach is designed to correct both quantitative and qualitative errors in the description of the thiol–disulfide exchange reaction. As a reference, we use DFT and CCSD(T) data, aiming at chemical accuracy of a computational scheme, which at the same time allows for a sufficient sampling of the conformational space in order to compute free energy surfaces. The approach is then compared to a traditional SRP methodology, where specific repulsive potentials are adapted to the reaction at hand. With both methods we were able to improve the accuracy of thiol–disulfide exchange reactions, with the Δ-ML approach even up to CCSD(T) accuracy, with a computation time that is 3–4 orders of magnitude faster than DFT-GGA.

11.2. Computational Details

11.2.1. Artificial Neural Network for Δ-Learning

In the Δ-ML approach,[200,206,207] the energy differences between two QM levels of theory are learned with a Behler–Parrinello ANN.[67] The general structure of ANNs and how molecules are encoded in order to use them as input for ANNs are described in chapter 6. For more details about the implementation, compare Ref. [192].

11.2.1.1. Training Set of Molecular Structures

The ANN was learned on a small system comprising a dimethyl disulfide and a methylthiolate anion, i.e. 15 atoms in total, as shown in Fig. 11.2A. The complete data set was generated from two different subsets which were obtained by different approaches. For the first subset, the two molecules were considered in the gas phase, and were geometry optimized on the B3LYP/TZVPP level of theory. Subsequently, an unrelaxed potential energy scan was performed as described in Ref. [189], which yielded 5112 different structures. All obtained structures are in a nearly linear S–S–S configuration and therefore this subset only covers a small part of the configurational space of the two molecules.

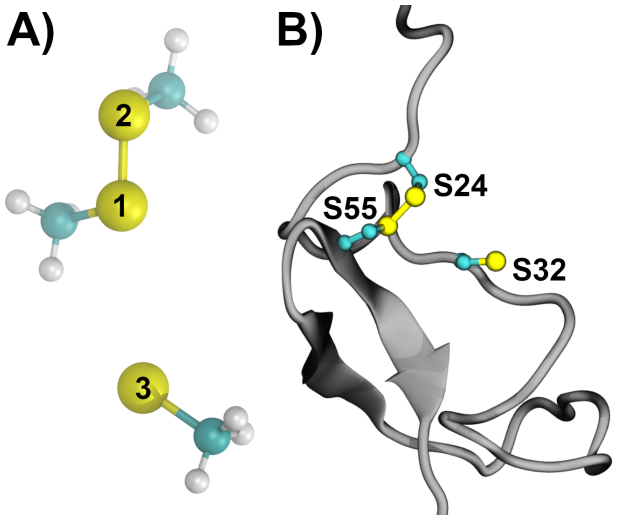

Figure 11.2.: (A) Exemplary structure of the methylthiolate and dimethyldisulfide (B) Exemplary structure of the I27* protein with the three cysteines at position 24, 55 and 32. Sulfur atoms are depicted by yellow balls.

In order to obtain more, especially non-linear configurations we performed QM/MM simulations of a small protein consisting of the residues 20 to 65 of a mutated immunoglobulin domain (I27*),[172] compare Fig. 11.2B and chapter 10. The protein contains two disulfide-bonded cysteines and a deprotonated cysteine anion, which were described quantum chemically with DFTB3/3OB. Covalent bonds between Cα and Cβ were treated with the link atom approach, i.e. capped with hydrogen atoms. Hence, the QM region comprises a methylthiolate and a dimethyldisulfide, the same two molecules that were used for the unrelaxed potential energy scan. To obtain a large number of different structures of the

system in a primarily non-linear configuration, thiol–disulfide exchange reactions were enforced by employing a metadynamics protocol. The distances between the three sulfurs were used as reaction coordinates, analogously to the setup for the thiol–disulfide exchange between the methylthiolate and the dimethyldisulfide in aqueous solution. Snapshots of the QM region were extracted from the trajectories and binned regarding the three sulfur–sulfur distance combinations. Only distances with a minimum length of 1.85 Å, a maximum bond length of 6.95 Å and a bin width of 0.15 Å were considered. This yielded 8436 bins, fewer than the theoretically possible number because not all distance combinations are chemically feasible or present in the QM/MM metadynamics. One structure from each bin was added to the second subset, and the complete training set comprises 13,548 unique structures. Histograms of the distribution of the structures as a function two S–S distances are shown in Fig. 11.3.

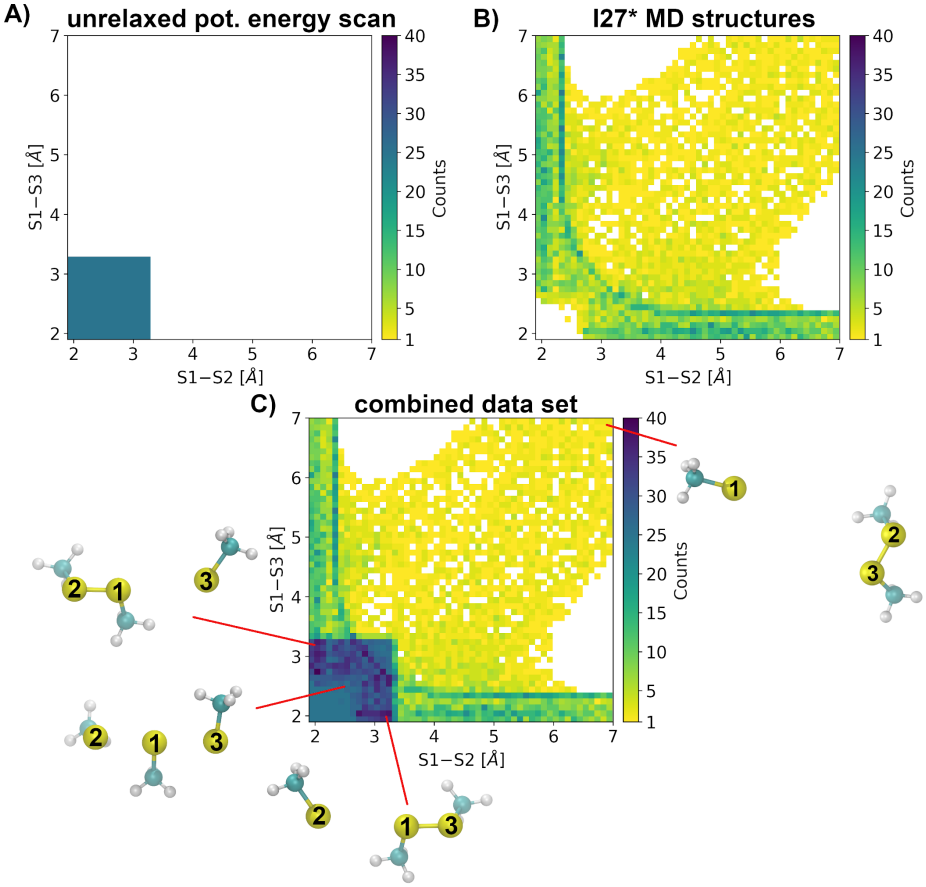

Figure 11.3.: Distribution of the training structures as a function of the S1–S2 and S1–S3 of the (A) unrelaxed potential energy scan (5112 structures), (B) QM/MM simulations of I27* (8436 structures), (C) combined data set (13,548 structures), A minimum bond length of 1.9 Å, a maximum bond length of 7.0 Å and a bin width of 0.1 Å was considered for the histograms.

11.2.1.2. Quantum Chemical Calculations

The DFTB energies for the training structures were calculated with DFTB3 employing the 3OB parametrization, using the DFTB+ software package.[133] The energy of the methylthiolate–dimethyldisulfide system at infinite separation was calculated as the sum of energies of the two isolated molecules, yielding a value of −9931.149 kcal/mol. This value was subtracted to obtain the relative energies throughout the entire work.

Reference ab initio calculations of energies for the same set of training structures in gas phase were performed at two different levels of theory, DFT-B3LYP and coupled clusters (CC). B3LYP energies were calculated by TurboMole 6.5 using the aug-cc-pVTZ basis set.[208] The sum of energies of reactans at infinite separation was −824854.937 kcal/mol. CC energies were calculated by ORCA 4.2.1[209,210] with CCSD(T)[211] at normal accuracy with the DLPNO approximation and the aug-cc-pVTZ basis set.[208] The sum of energies of reactants at infinite separation was −823715.672 kcal/mol.

11.2.1.3. Training of the Artificial Neural Network

The training of the ANN was performed by C. L. Gómez-Flores and is detailed in Ref. [192].

11.2.2. Test calculations

11.2.2.1. Dimethyl Disulfide and Methylthiolate in Water.

We performed QM/MM multiple walker metadynamics[59,60,131] of a system composed of a dimethyl disulfide molecule and a methylthiolate anion with different QM methods/parameters to benchmark the reaction energies of thiol-disulfide exchange. The system setup is the same as for the metadynamics simulation of disulfide shuffling in a symmetric aqueous model system described in Ref. [172] (cf. subsec. 10.2.3), however without an imposed additional artificial ESP. The methods are

(I) DFTB3 with the parameter set 3OB,[135,196]

(II) DFTB3/3OB complemented with the machine learned ΔE corrections to B3LYP,

(III) DFTB3/3OB including the SRP for the S–S interaction parametrized with a B3LYP/aug-cc-pVTZ unrelaxed potential energy scan,

(IV) DFTB3/3OB complemented with the machine learned ΔE corrections to CCSD(T),

(V) DFTB3/3OB including the SRP based on G3B3 data.

For better readability, method (I) will be denoted as DFTB/3OB, method (II) as ML/B3LYP, method (III) as DFTB/B3LYP, method (IV) as ML/CC and method (V) as DFTB/G3B3 in the following. The SRP of the S–S repulsive potentials (III) and (V) are detailed in Ref. [192]. The three distances between the sulfurs were used as collective variables (CV) to drive the reactions. All simulations were performed with a local Gromacs 2020 version[122] patched with Plumed 2.5.1[123,134] and interfaced with DFTB+ 19.1,[132,133] which itself was additionally modified for simulations using the ML corrections.

11.2.2.2. C4 Domain of von Willebrand Factor.

Setup.
The Von Willebrand Factor (VWF) is glycoprotein present in blood plasma comprising several domains. The C4 domain contains five disulfide bonds which might undergo disulfide shuffling with each other resulting in many different exchange combinations.[212,213] Based on MD structures taken from Ref.[213], a disulfide bond between cysteine 57 and cysteine 79 was cut, cysteine 57 protonated and cysteine 79 designed as an anion. The protein, described with the AMBER99SB-ILDN forcefield[185], was placed in a cubic box of $9 \times 9 \times 9$ nm^3. The box was then filled with 23477 TIP3P water molecules and neutralised with 2 sodium atoms. Periodic boundary conditions were set and electrostatics were treated with particle-mesh Ewald. Lennard-Jones interactions were cut-off at 1 nm. All following steps were performed with Gromacs 2020-dev patched with Plumed 2.5.1 and interfaced with DFTB+ 19.1. First, the system was energy minimised with the steepest descent method. Subsequently, harmonic position restraints were applied to the heavy atoms of the protein with a force constant of 1000 kJ mol^{-1} nm^{-2}. A NVT equilibration with the Bussi thermostat at 300 K over 100 ps was performed followed by a NPT equilibration at 300 K (Bussi thermostat[126]) and 1 bar (Parrinello-Rahman barostat[186]) over 10 ns. The leap-frog integrator was used with a time step of 1 fs and the neighbour list was updated every 10 MD steps. The harmonic position restraints were lifted and the system additional 100 ps NPT equilibrated.

QM Equilibration and Metadynamics.
We investigated the disulfide exchange reactions between the cysteines at position 36, 78 and 79. The three cysteines were described quantum mechanically and bonds between Cα and Cβ were treated with the link atom approach, i.e. the QM region is capped with a hydrogen atom placed along the bond. The QM region consists of 15 atoms and electrostatic interactions between the charged QM region and the MM system were also scaled down by the factor of 0.75.[187,188] A hydrogen-bond correction was employed with the *damping* method[54] and the BJ implementation of the D3 correction[129] was used with parameters for the 3OB parameter set. The system was NPT equilibrated over a short period of 40 ps using 3OB parameters.

Next, pilot metadynamics simulations were performed with the three sulfur–sulfur distances as reaction coordinates and the modified S–S parameters based on G3B3 data. However, the 3D setup is difficult to converge and therefore we considered the following three setups:

$$\mathrm{S79^-} + \mathrm{S36{-}S78} \rightleftharpoons \mathrm{S79{-}S36} + \mathrm{S78^-}$$
$$\mathrm{S78^-} + \mathrm{S79{-}S36} \rightleftharpoons \mathrm{S78{-}S79} + \mathrm{S36^-}$$
$$\mathrm{S36^-} + \mathrm{S78{-}S79} \rightleftharpoons \mathrm{S36{-}S78} + \mathrm{S79^-}$$

where the sulfurs are in an nearly linear configuration. The two respective distances between the central and outer sulfurs were used as reaction coordinates, i.e. S79–S36/S36–S78, S36–S78/S78–S79, and S78–S79/S79–S36. Starting structures for each of the three reactions were selected from snapshots of the pilot 3D metadynamics. Prior to the 2D metadynamics, each starting structure was duplicated and equilibrated over 40 ps. During

the equilibration one duplicate in each setup was forced into the reactant state and the other into the product state. The S–S–S angles of the respective linear configurations were restrained to values > 160° with a force constant of 100 000 kJ mol^{-1} rad^{-1}.

Finally, well-tempered multiple walker metadynamics[59,60,131] of each reaction was performed with the previously described DFTB methods and parameters. In each simulation 24 walkers were used and simulated over 500 ps, 12 walkers starting in the reactant state and 12 walkers in the product state. S–S–S Angles were kept > 160°, gaussian hills with a height of 1 kJ/mol and a width of 0.2 Å were deposited every 500 fs. A bias factor of 20 was set and deposited biases from all other walker were read every 1000 fs. The configurational space of the reaction was reduced by restraining the respective CV distances < 6.0 Å with a force constant of 50 000 kJ mol^{-1} nm^{-1}.

11.3. Results and Discussion

The hyperparameter optimization and training results are detailed in Ref. [192]. The focus of this chapter lies on the benchmark simulations.

11.3.1. Benchmark: Free Energies of Aqueous Molecular Systems

As discussed above, the PES of thiol–disulfide exchange for a solvated system differs significantly from a gas-phase system, and a transition state appears where there is a minimum in the gas phase (compare Fig. 11.1). To investigate the performance of the DFTB+/ML approach for solvated systems, we performed QM/MM metadynamics simulations of a dimethyl disulfide–methylthiolate system immersed in water that was described by an MM force field. The metadynamics setup was designed to sample all three disulfide bond patterns, i.e. S1–S2, S1–S3, and S2–S3 with the respective third sulfur in a deprotonated anionic state. The free energy profile of the exchange reactions is completely symmetric and therefore ideally suited for comparing the different levels of theory.

Free Energy Profiles – Convergence and Transition State Geometry.
The 2D representations of the three-dimensional (3D) free energy landscape, expressed as a function of the S1–S2 and S1–S3 distances with the S2–S3 distance integrated out, are shown in Fig. 11.4 together with exemplary molecular structures and pathways. All PMFs are symmetrical and show the three expected minima of equal depth. Moreover, the transition states within the respective PMFs have the same energy, which illustrates the good convergence of the simulations.

The PMF obtained with uncorrected DFTB/3OB (Fig. 11.4A) shows two significant problems: (i) the bonds S1–S2 and S1–S3 in the transition state geometries are too long with ca. 2.8 Å, and (ii) the transition state geometries exhibit shallow minima on the free energy landscape, rather than saddle points[189]. As discussed above, the longer bonds in the TS geometry may be related to the deficiencies of the underlying PBE functional.[193] Since PBE shows a relative error of ca. 7 kcal/mol for the gas-phase structure, it can be assumed that the same error also occurs in solution, and thus PBE would fail to describe the local bonding structure of the S–S–S moiety. In fact, the barrier between the minima is

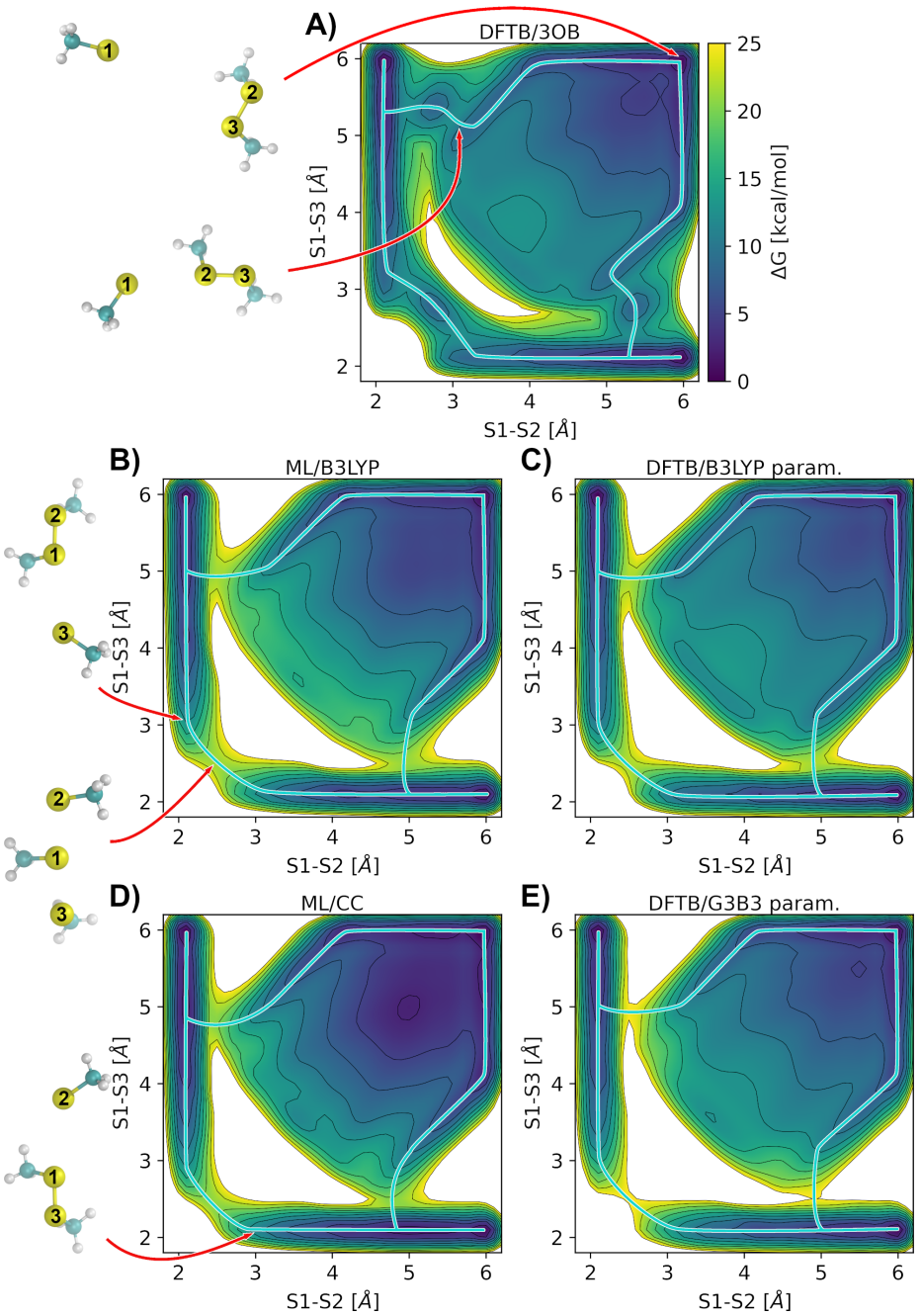

Figure 11.4.: Potential of the mean force (PMF) of disulfide shuffling between a dimethyl disulfide and a methylthiolate in aqueous solution obtained with different QM methods – (A) DFTB3 with 3OB parameters, (B) DFTB3 with 3OB parameters and a machine learned energy correction ΔE based on B3LYP, (C) DFTB3 with a reparameterized S–S repulsive potential fitted to B3LYP data, (D) DFTB3 with 3OB parameters and a machine learned energy correction ΔE based on CC, (E) DFTB3 with a reparameterized S–S repulsive potential fitted to G3B3 data. Contour lines are drawn every 2.5 kcal/mol. Exemplary pathways are drawn as light blue lines.

about 7 kcal/mol lower than in the calculations using the Δ-NN corrections and the SRPs, as discussed in the following.

Both problems are resolved by applying a machine learned energy contribution ΔE (Fig. 11.4B and D) or by reparametrizing the S–S repulsive potentials (Fig. 11.4C and E). The transition states now appear as saddle points at shorter bond lengths S1–S2 and S1–S3 of ca. 2.5 Å with DFTB/B3LYP, ML/B3LYP and DFTB/G3B3. This correlates with the potential energy scan in vacuum (Fig. 5B.2 in Ref. [192]). With ML/CC, the transition state appears at even shorter bonds of ca. 2.41 Å, similarly to the minimum energy structure of the potential energy scan in vacuum (Fig. 5C.3 in Ref. [192]) and with the MP2/aug-cc-pVTZ calculation by Neves et al.[193]

Remarkably, Δ-NN is able to correct the geometries and energies in aqueous solution even though the energy differences were learned for the gas-phase situation, and the interactions with the environment as calculated by DFTB were not corrected explicitly. The slight differences between the two methods, B3LYP and CCSD(T), regarding structures and energetics seem to be reproduced, i.e., these are not overwritten by any larger errors in the QM/MM interactions and sampling convergence. Note that we are comparing here total free energies from five independent free energy simulations.

Reaction Barriers.
Additionally, we obtained the heights of energy barriers to the thiol–disulfide shuffling. Two different configurations are considered as minima of free energy: The first structure is the global minimum on the 2D-PES, which is a triangular structure with S1–S2 and S1–S3 distances of ca. 6 Å; the second structure is a linear structure with a S1–S3 distance of 3 Å, similar to the exemplary structures shown in Fig. 11.4. When the sulfurs are linearly aligned, 3 Å is approximately the minimum distance that S3 and S1 can maintain before the disulfide bond between S1 and S2 elongates and eventually breaks.The latter choice of the structure is comparable to the benchmark model in Ref.[193], thus allowing for a comparison with their estimates of barrier heights from ab initio calculations employing implicit solvent calculations. The barrier heights obtained with both approaches are listed in Tab. 11.1 and plotted in Fig. 11.5.

Table 11.1.: Simulation time per nanosecond and reaction barriers w.r.t the global minima and the minima at S1–S3 = 3 Å. Energies are given in kcal/mol.

Method	h/ns	$\Delta G^{\ddagger}_{\text{global}}$	$\Delta G^{\ddagger}_{3\text{Å}}$
DFTB/3OB	6.8	12.4	3.3
ML/B3LYP	10.4	21.5	8.8
DFTB/B3LYP	6.0	22.0	11.8
ML/CC	11.1	20.7	9.8
DFTB/G3B3	6.2	24.9	10.8

Barrier Heights w.r.t. global minimum.
We choose the PMF obtained with DFTB-ML/CC as reference which yields a barrier

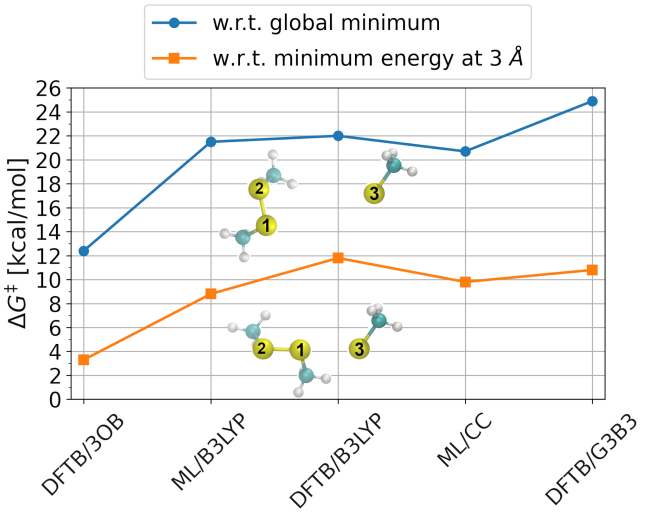

Figure 11.5.: Reaction barriers of disulfide shuffling between a dimethyl disulfide and a methylthiolate in aqueous solution obtained with different QM methods. Blue circles: Barriers calculated w.r.t the global minimum, i.e., triangular configuration; Orange squares: Barriers calculated w.r.t the energy of the linear structure with S1–S3 = 3Å.

$\Delta G^{\ddagger}_{\text{global}}$ = 20.7 kcal/mol. As mentioned before, the standard 3OB parametrization (DFTB/3OB) underestimated the barrier with 12.4 kcal/mol clearly. With the ML correction to B3LYP data (ML/B3LYP), the barrier lies 0.9 kcal/higher, at $\Delta G^{\ddagger}_{\text{global}}$ = 21.5 kcal/mol. The S–S repulsive potential (SRP) fitted to the same B3LYP data (DFTB/B3LYP) gives $\Delta G^{\ddagger}_{\text{global}}$ = 22.0 kcal/mol, which is 1.4 kcal/mol higher than the reference. The highest barrier is obtained with the S–S repulsive potential fitted to G3B3 data (DFTB/G3B3) with $\Delta G^{\ddagger}_{\text{global}}$ = 24.9 kcal/mol.

Overall, the SRP seems to work quite well. The correction of E_{rep} adjusts the S–S pair potentials, so only two-body contributions are affected. In cases where angular degrees of freedom, like bending angles or dihedrals need an additional correction, this approach may not work, as is the case of the peptide addressed in Ref. [200], in which the torsional barriers are underestimated with DFTB slightly.

Barrier Heights w.r.t. S1–S3 = 3 Å.
With this structure, we find barrier heights of around 10 kcal/mol, which agrees well with the CCSD(T) values of 9.28 kcal/mol.[193] Notably, that work applied an implicit solvation model for a restricted area of conformational space. The missing contribution to solvation energy and the entropic contributions were estimated at ca. 8 kcal/mol, which indeed brings the final result close to 20 kcal/mol as discussed above for the extended PES.

Computational Cost.
On average, calculations with DFTB required 6–7 h/ns on a single Xeon 2.20 GHz CPU core (Silver 4214), which increased to 10–11 h/ns when using the ML–ΔE correction. As stated in the previous section, most of the computation time of the ML correction accounts to the calculations of the forces (ca. 85%) whereas the calculation of ΔE is quite fast (ca. 15%). Note

that the calculation times include the metadynamics calculation and also the calculation of several restraints during the simulations. With DFTB/3OB the thiol-disulfide exchange barrier is underestimated by ∼10 kcal/mol, therefore the potential energy surface is "filled" faster and high energy regions are explored earlier than with the SRPs, DFTB/B3LYP and DFTB/G3B3. In these regions, DFTB+ calculations are more demanding and furthermore restraints which prevent the reaction to non-sensical chemical species set in and are calculated. Hence, the average calculation time with DFTB/3OB takes longer than with DFTB/B3LYP and DFTB/G3B3. In unbiased, free QM/MM MD simulations the simulation times should be the same.

11.3.2. Application: 2D metadynamics of thiol–disulfide exchange in the C4 domain of von Willebrand factor

In order to see how the different methods perform in more complex environments, we simulated thiol–disulfide exchanges between three cysteine residues located on two different flexible loops of the C4 domain of the von Willebrand factor. Each C domain (C1 to C6) contains at least 4 intramolecular disulfide bonds, some of which are partially reduced, facilitating intramolecular exchange of disulfide bonds. Recent findings suggest that that the reduction of the wt S36–S78 bond compromises the biological function of the C4 domain, that is, reinforcement of platelet binding.[213] Thus, a disulfide bond connecting the two subdomains in C4 is crucial for the functionality of the protein, which in our case corresponds to the wt bond S36–S78 and a disulfide bond between S79–S36. In this work, the disulfide bond between S79–S57 was reduced and subsequently the thiol-disulfide exchange between S79 and the wt disulfide bond S36–S78 was investigated. It turned out that a 3D setup with all three distances considered as reaction coordinates (like for the previously described small benchmark system) did not converge. Thus, we investigated each of the three exchange reactions individually in a 2D setup where the respective S1–S3 and S1–S2 distances were used as reaction coordinates, and the sulfurs were aligned linearly ($\geq 160°$) to reduce the configurational space of sampling. The obtained reaction barriers are shown in Tab. 11.2 and Fig. 11.6 together with exemplary structures of the protein and the radial distribution functions (RDF) of water. The RDFs are calculated w.r.t. the central sulfur S_{ctr} of the respective exchange reaction and averaged over all methods used to simulate the same setup.

The barrier heights obtained with the different methods show a similar trend as those in the benchmark system: DFTB/3OB yields the lowest barrier, DFTB/G3B3 the highest, and ML/CC, ML/B3LYP and DFTB/B3LYP agree well with each other. Thus, we only discuss the barrier heights and reaction energies obtained with ML/CC in the following.

We find two clusters of barrier heights, one with high values of 13.6 and 17.8 kcal/mol, and another with smaller values around 9 kcal/mol. The higher values correspond to the two reactions that lead to a vicinal disulfide bond between the adjacent cysteines S78–S79. Vicinal disulfides constitute the smallest possible intramolecular disulfide loops, which experience a high steric strain due to their small ring size.[214,215] Thus, a higher barrier has to be overcome to form the bond S78–S79, which also lies higher in energy than the other two disulfide bonds, S36–S78 and S79–S36. All other investigated thiol–disulfide

Table 11.2.: Energy barriers to thiol–disulfide shuffling between S36, S78 and S79 in the C4 domain in aqueous solution. Barriers are considered w.r.t. the global minimum of the respective bond. All values in kcal/mol.

	b1 ⇌ b2	b2 ⇌ b3	b3 ⇌ b1
method	$S79^-$ + S36–S78 ⇌ S79–S36 + $S78^-$	$S78^-$ + S79–S36 ⇌ S78–S79 + $S36^-$	$S36^-$ + S78–S79 ⇌ S36–S78 + $S79^-$
DFTB/3OB	4.5 / 3.5	13.3 / 3.1	2.2 / 9.4
ML/B3LYP	9.1 / 8.6	16.0 / 6.8	9.0 / 13.9
DFTB/B3LYP	8.7 / 8.4	14.8 / 7.8	8.0 / 14.3
ML/CC	9.9 / 9.3	17.8 / 7.4	10.0 / 13.6
DFTB/G3B3	11.9 / 10.3	19.8 / 9.5	10.1 / 17.5

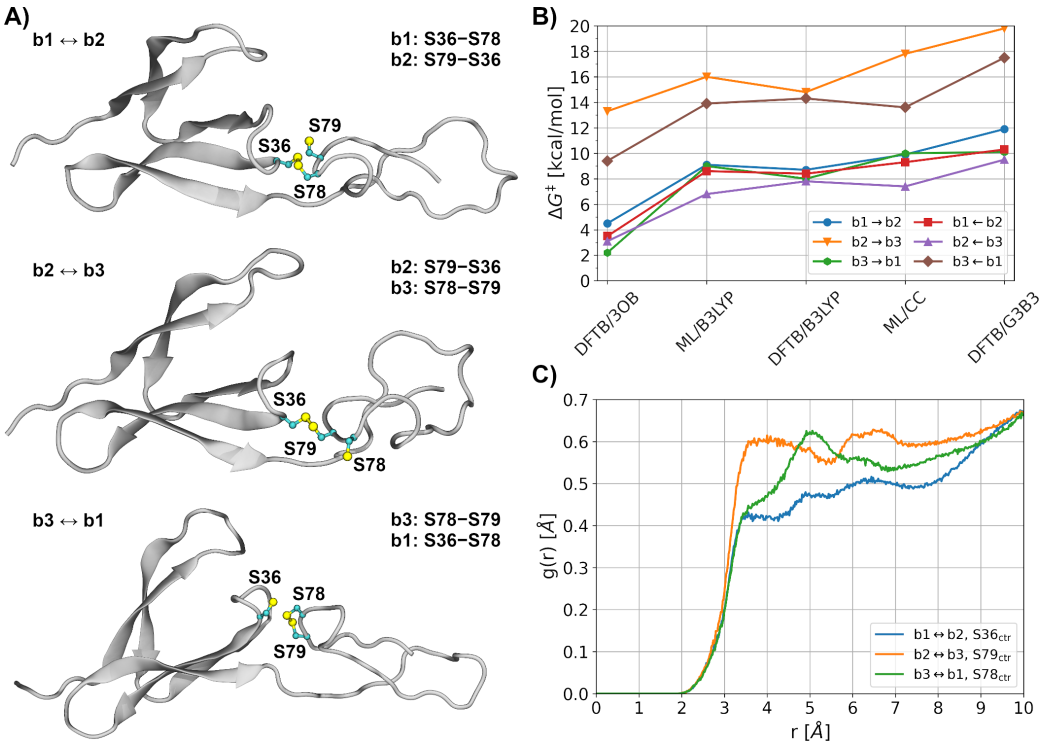

Figure 11.6.: Energy barriers to thiol–disulfide shuffling between Cys36, Cys78 and Cys79 in the C4 domain in aqueous solution obtained with the different QM methods. (A) Exemplary structures of the protein for three investigated cases. (B) Barriers w.r.t. the global minimum of each respective bond. (C) Radial distribution function of water around the respective central sulfur S_{ctr} of each setup, averaged over the simulations performed with the five different methods. The individual RDFs are shown in the *Appendix*.

exchange reactions have barriers around 9 kcal/mol. Note that the reaction energies are

obtained as the difference between the respective barrier heights. The difference of barrier heights in these two classes of reactions can be explained in terms of solvent accessibility, as illustrated by the RDF in Fig. 11.6C. In the b3⇌b1 reaction, the sulfurs are less exposed to water than in the b2⇌b3 reaction, which leads to a more delocalized charge and thus smaller barriers. As discussed above, the aqueous solvent introduces a significant barrier of about 20 kcal/mol, which is absent in the gas-phase energetics. The protein in the b3⇌b1 reaction shields the reaction from solvent access partially, which leads to a lowering of the barrier.

Compared to the methylthiolate–dimethyldisulfide system in aqueous solution, the symmetry of the reaction is broken due to the protein environment, resulting in different barrier heights and reaction energies. Furthermore, the protein environment is catalytic for the thiol–disulfide exchanges; the barrier heights of the two exchange reactions leading to the vicinal disulfide S78–S79 are reduced by 13% and 34%, and those of the other four exchanges are reduced by as much as 52% to 64%. This effect can be attributed to steric and electrostatic interactions.[172] These are addressed qualitatively in the *Appendix*.

Note, however, that the reactions in the C4 domain were kept linear whereas the trisulfide reactions in solution were not. Hence, the barriers might not be directly comparable. In C4, there might be a non-linear configuration which lies lower in energy than the linear ones, i.e. the reaction barrier could be higher than computed.

Interestingly, the barrier heights of the two clusters match those in Ref. [195], where Neves et al. investigated two consecutive thiol–disulfide exchange reactions between a protein disulfide isomerase and a glutathione disulfide as substrate with QM/MM, which yielded barrier heights of 18.7 and 7.2 kcal/mol.

11.4. Conclusion

Disulfide bonds have an important role for the function of many proteins, therefore, being able to address these reactions using accurate computational approaches is of great importance. These reactions were shown to be quite challenging for DFT methods[193], requiring costly computational approaches to be applied. Sampling, however, is then out of reach, which poses a further restriction on the accuracy of the results.

SE methods are 3–4 orders of magnitude faster than DFT-GGA using moderately sized basis sets, however, they may run into even greater difficulties for challenging reactions than DFT-GGA do. An appealing work-around is therefore the combination of SE with ML approaches, as shown in this work. The combination is fruitful in two respects: (i) First, the computational effort is comparable, therefore, this combination may be the fastest way to reach ab initio quality of PES, still allowing for extensive sampling. (ii) Second, the quality of the SE is high enough, so that both qualitative and quantitative errors seem to be small and consistent, so that the ML approach may be effective.

This has been demonstrated by training to two different ab initio levels, and the differences of the methods have been reproduced in all the applications considered. This indicates that the ab initio accuracy, trained for small systems and reference conformations, can be transferred into complex environments without loss of accuracy. Then, it is possible to obtain phase space sampling sufficient to generate the free energy profile of the

reaction. Obviously, this is only possible as long as the poor performance of the lower-level method (which is to be corrected) is rooted in local bonding effects. As an example, the thiol–disulfide exchange is sensitive to correlation effects,[193] which are difficult to capture with simple electronic structure methods. Importantly, these effects are local – localized to the bonds formed by the sulfur atoms, which can be learned in simplified gas-phase systems and then transferred to more complex environments.

The benchmarked SE methods were used to investigate the thiol–disulfide shuffling in the C4 domain of von Willebrand factor, which is a component of a large mega-dalton blood protein. The two subdomains of C4 are no longer connected when the vicinal disulfide bond S78–S79 between the two sequence adjacent cysteines is formed, for which we find high reaction barriers of 14–18 kcal/mol. All other exchange reactions show moderate barriers of ca. 9 kcal/mol. Moreover, the vicinal disulfide bond S78–S79 lies significantly higher in energy than the wt disulfide bond S36–S78 and the disulfide bond S79–S36. Thus, our results show that the biological function is controlled by steric contributions and electrostatic interactions. The protein effectively shields the solvent from the reaction center, thereby creating a less polar environment. As shown for the gas phase and solvent reaction profiles, the solvent exposure introduces a substantial barrier of ca. 20 kcal/mol, while the reaction is barrierless in gas phase. The protein environment therefore effectively modulates the energy barrier by desolvating the active site.

12. Force-clamp simulations of von Willebrand factor's C4 domain

The work in chapter 12 was done in cooperation with the Gräter and Hogg group:

- Kutzki, F.; Butera, D.; Lay, A. J.; Maag, D.; Chiu, J.; Wook, H.-G.; Kubař, T.; Elstner, M.; Aponte-Santamaría, C.; Hogg, P. J.; Gräter, F.; Force-Propelled Reduction and Exchange of Disulfide Bonds in Von Willebrand Factor's C4 Domain., *manuscript*.

12.1. Introduction

The von Willebrand Factor (vWF) is a large mutltimeric blood protein which is responsible for platelet adhesion to collagen as response to vascular injuries.[216–218] It consists of 2050 amino acids, divided into 12 domains, of which are many disulfide bonded cysteines. After and injury, vWF elongates under the shear stress of blood and the vWF-mediated platelet adhesion is initiated. This process is well understood, however, the mechanistic details of how the disulfide bonds can change under shear stress are still largely unknown. They are obviously essential for the structural integrity and function of the protein because many diseases that are associated with vWF show mutations of cysteines. The disulfide bonds are known to drive the mulitmerization of vWF and protect vWF from unfolding due to the tensile forces of flowing blood.[219] Moreover, the rearrangement of disulfide bonds might mediate different processes such as the multimer size of vWF[220] and platelet binding.[221,222]

In this study, we focus on the C4 domain of vWF which acts as an anchoring point for platelets.[217] C4 consists of 85 amino acids. The platelet binding site with and arginine-glycine-aspartate (RGD) motif lies at the tip of the first beta-hairpin, compare Fig. 12.1A.[212] The structure is stabilized by 5 intramolecular disulfide bridges which protect the protein from unfolding under the influence of shear stress. Since it has been shown that mechanical forces can catalyze thiol-disulfide exchanges, for example in the engineered Immunoglobulin domain I27* (compare Ref. [174] and chapter 10), the possibility of intramolecular disulfide bond rearrangements in C4 is investigated.

Force clamp simulations (performed by the Gräter group) find, that disulfide bonds S7–S41 and S36–S78 carry a higher mechanical load than the other three disulfide bonds when the termini are pulled into opposite directions with a constant force of ~500 pN. This indicates that S7–S41 and S36–S78 are more prone to rupture, which is confirmed by experiments (conducted by the Hogg group). Blood samples from 10 healthy donors showed that the disulfide bonds S7–S41 and S36–S78 are reduced by an average of 3.4% and 2.7%, respectively. Thus, a small fraction of the sulfurs are free thiols, which theoretically can perform a thiol-disulfide exchange. Indeed, a new disulfide bond S36–S41 was found *ex vivo* (resolved by HPLC and analysed by mass spectroscopy) as a result of a thiol-disulfide

exchange between S36 and S7–S41. Additional force-clamp simulations of the Gräter group, with a reduced S36–79 bond (Fig. 12.1B), even show that S36 can approach either S7 or S41 with similar probability. To test whether both attacks are chemically feasible, QM/MM metadynamics of the disulfide exchange reactions were performed and the PMFs and reaction barriers obtained.

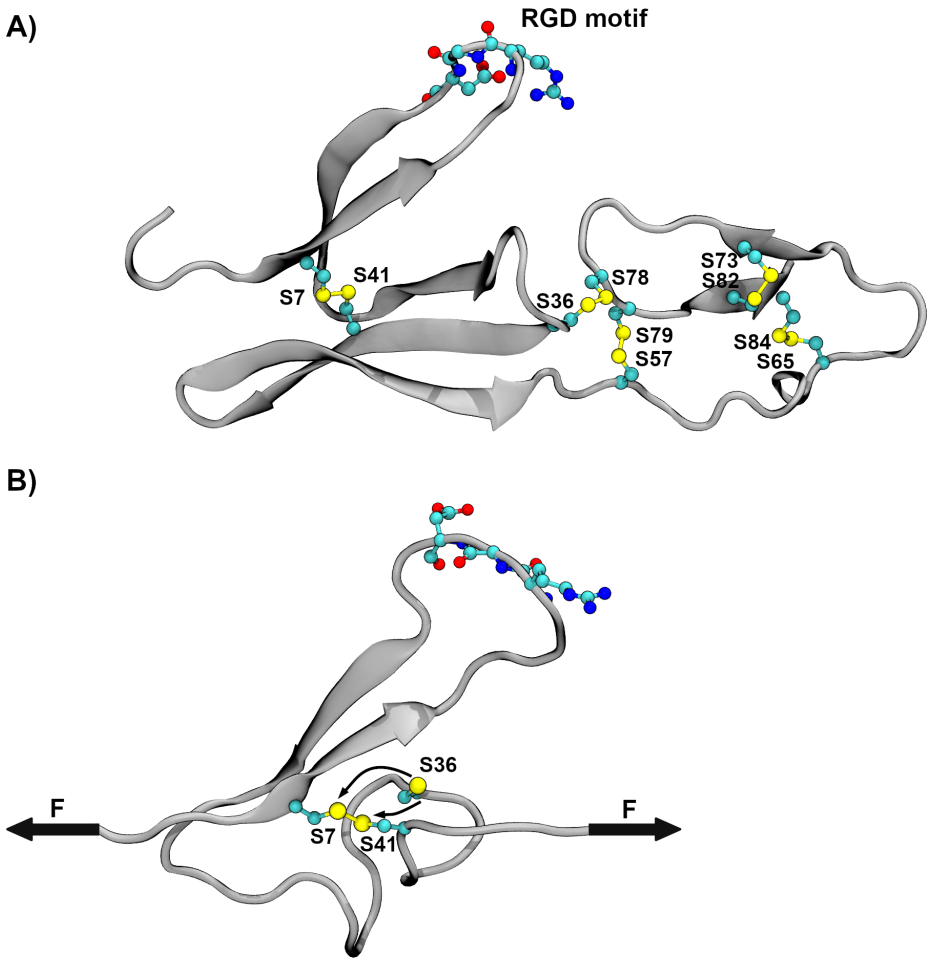

Figure 12.1.: (A) C4 domain of the von Willebrand Factor. The RGD motif located on the flexible loop serves as platelet binding site. (B) C4 domain with reduced S36–S78 bond after pulling the termini into opposite directions with a constant force. This leaves a free thiol S36 which can perform a thiol-disulfide exchange reaction with the disulfide bonded S7 or S41.

12.2. Computational Details

System Setup.
From the classical force-clamp MD simulations (performed by Fabian Kutzki), seven snapshots that showed high mutual RMSDs were selected as initial structures QM/MM simulations. The C-terminus, which is stretched by the applied pulling force, was removed

and only residues 1 through 46 were retained to reduce the computational cost. The truncated protein was placed in a box of $7.7 \times 5.9 \times 5.0$ nm^3 and solvated with 7222 TIP3P waters. The system was neutralized with 2 sodium ions and in addition a concentration of 150 mmol NaCl was added (20 sodium ions and 20 chloride ions). The structures were equilibrated over 500 ps in an NVT ensemble at 300 K with the leap-frog integrator using a time step of 1 fs. The simulations were performed with Gromacs 2020 patched with Plumed 2.5.1 and interfaced with DFTB+ 19.1.[122,123,132] The setup described in the following is analogous to the setups described in chapter 10 and 11. The side chains of the investigated cysteines (here: Cys7, Cys36, Cys41) were included in the QM region and described with DFTB3 using the 3OB parameter set[54,135] with the reparameterized sulfur-sulfur parameters based on G3B3 data.[192]. The rest of the system was described with the AMBER99SB-ILDN forcefield[185] and the TIP3P water model. The termini were pulled into opposite direction during the simulations with constant pulling force of 500 kJ/mol/nm (830 pN) along the x-axis. In order to compensate for missing electronic polarization of the MM region, the electrostatic interactions between the QM and MM region were scaled down by 0.75.[187,188]

Metadynamics:
The pulling force was reduced to 100 kJ/mol/nm (166 pN) and multiple walker well-tempered metadynamics[59,60,131] of the thiol-disulfide exchange reactions were performed. The final structures of the equilibrations served as initial starting structures for seven separate metadynamics simulations. We used 16 walker for each starting structure and every walker was simulated over 18.75 ns, yielding a simulation time of 300 ns per setup and over 2 µs in total. Following the setup described in Sec. 10.2.3, the sulfur-sulfur distances were used as reaction coordinates. Along the trajectories, gaussian biasing potentials with an initial height of 1.0 kJ/mol and a width of 0.2 Å were applied every 500 fs and exchanged every 1000 fs with all other walkers. A bias factor of 50 was set.

Additional restraints.
To reduce the configurational space of sampling, the S36–S7 and S36–S41 distances were kept below 1.1 Å and S7–S41 below 6 Å by means of harmonic restraints with a force constant of 100 000 kJ mol^{-1} nm^{-2}. To prevent S–S bond breaking without a third sulfur nearby or reactions to chemically non-sensical species, the same additional restraints described in section 10.2.3 were applied.

12.3. Results and Discussion

The 2D representation of the three-dimensional (3D) free energy landscape obtained for starting structure 1 is shown in Fig. 12.2 together with exemplary molecular structures, all seven free energy landscapes are shown in Fig. E.1. The barrier heights of both reactions, S36 →S7–S41 and S36 →S41–S7, are shown in Tab. 12.1.

For an attack of S36 on S7, a barrier of $\Delta F^{\ddagger}$ = 24.4 (1.0) kcal/mol has to be overcome on average. The attack of S36 on S41 exhibits a similar barrier height with $\Delta F^{\ddagger}$ = 24.3 (1.8) kcal/mol. Hence, both reactions are chemically feasible and can occur with similar probability.

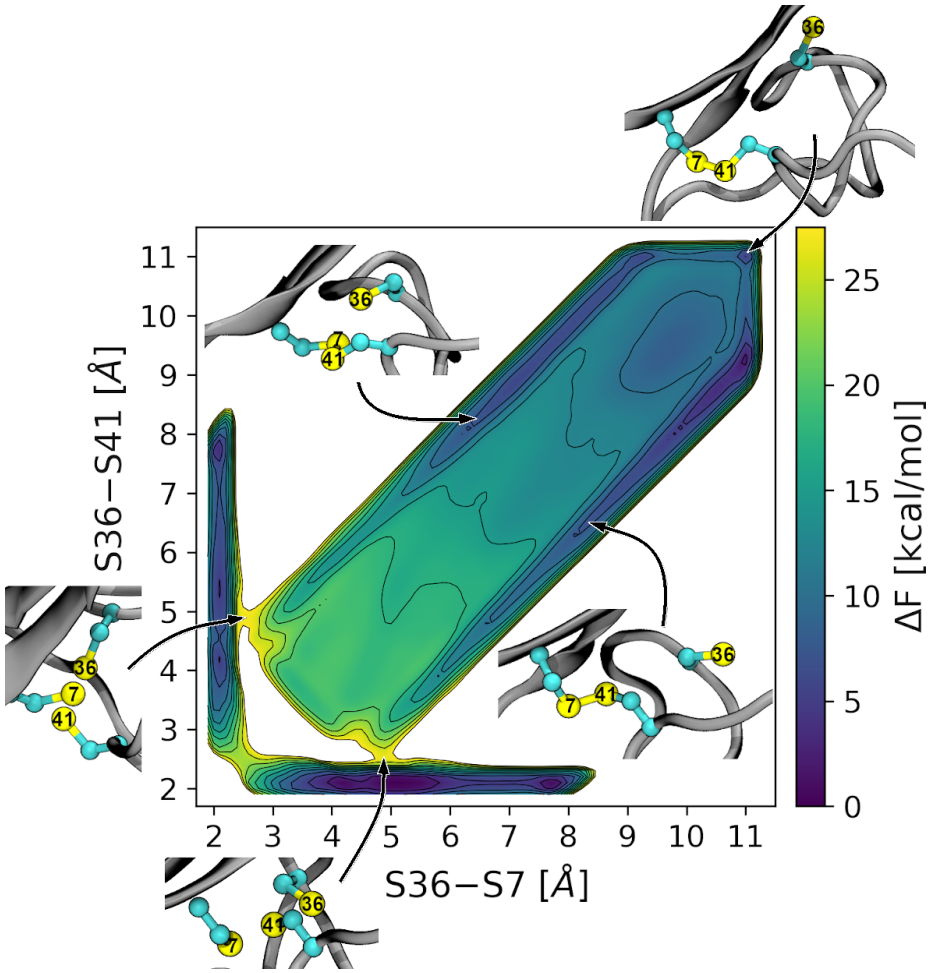

Figure 12.2.: Potential of the mean force as function of the S36–S7 and S36–S41 distances, with the S7-S41 distances integrated out, obtained with metadynamics for starting structure 1. Contour lines are drawn every 2.5 kcal/mol.

Table 12.1.: Reaction barriers in kcal/mol.

starting structure	S36 →S7–S41	S36 →S41–S7
1	24.9	24.4
2	25.6	25.8
3	22.7	26.8
4	23.4	25.8
5	25.6	23.7
6	24.1	21.3
7	24.9	22.2

13. Summary

In **chapter 7**, the final step of bacteriorhodopsin's photocycle was investigated. Despite more than 45 years of research, the underlying mechanism of the conversion from the **O** state back to the **bR** ground state remained largely unclear.

With multi-dimensional QM/MM metadynamics simulations we showed that the long-range PT during the **O→bR** transition proceeds via a proton hole mechanism. The protonation states of the involved key residues (D85, D212, E204, E194) are coupled to structural rearrangements of water molecules and an arginine sidechain (R82). In the **O** state, and a water wire connects the key residues and R82 is oriented downward to stabilize the negative charges of E204/E194. The water wire is disrupted when R82 swings upwards after the PT to stabilize the negative charges of D85/D212 in the **bR** state. Thus, the structural rearrangements of R82 modulate the hydration level of the active site. The positive charge of R82 also seems to facilitate the formation of the proton hole.

The obtained free energy profile in this work provides, for the first time, a mechanism that is consistent with with experimental data. Using transition state theory, we found that the **O** to **bR** transition occurs on a time scale of ~0.6 ms. Experiments suggest an estimated time scale of 0.5–5 ms. The calculated exergonicity of −3.6 kcal/mol also agrees with experimental estimates.

In **chapter 8**, the conversion of the Rh-Bl to the Rh-UV state was investigated in a homology model of histidine kinase rhodopsin. The model was validated by calculating the excitation energies of the intermediate states P550 and P570 which were obtained from QM/MM molecular dynamics simulations. The experimental red shift was quantitatively reproduced for P550 but only qualitatively for P570, where the calculated excitation energies of snapshots were unevenly distributed. This may have lead to an inaccurate absorption maximum, which is determined by fitting a Gaussian curve to the energy distribution.

In the P570 state, a water wire is formed along which the PT from the Schiff base to D239 was simulated with umbrella sampling. The computed free energy profile exhibits a global minimum for the protonated Schiff base and only a local minimum for the deprotonated Schiff base (RH-UV state). This finding does not agree with experiments since the RH-UV state is supposed to be stable for more than 24 h. However, using transition state theory, the obtained reaction barrier agrees well with the estimated time scale of the transition. Thus, it could be that the RH-UV state is not reproduced well in the simulation and therefore should be simulated longer, for example, to see if the water wire breaks down which would make a reprotonation of the Schiff base unlikely.

In **chapter 9**, the implementation of coupled-perturbed equations into third order DFTB was tested. The framework was applied to proton-coupled electron transfers in a model

system, using Mulliken charges as reaction coordinates to drive the ET. Different charge state and configurations of the model system in gas phase (QM only) and in aqueous solution (hybrid QM/MM) were considered. All simulations were stable and the obtained free energy profiles showed the expected topography and energy barriers. Hence, the method could next be applied to PCET reactions in more complex systems or in proteins.

In **chapter 10**, the intramolecular thiol-disulfide exchange between S32 and S24–S55 in an engineered I27 domain under mechanical stress was investigated. Experiments showed a preferential attack for S32→S55 over S32→S24, which we also observed in an ensemble of 334 unbiased QM/MM MD simulations. We analyzed the factors that contribute to the preferential attack and found that S32 can approach S55 over a wider range of angles than S24. Furthermore, S32 is more often found near S55 than S24, which makes a possible attack more likely.

In addition to the structural factors, we also analyzed how electrostatic interactions contribute to the regioselectivity. For this, the Mulliken charges of the sulfur atoms were examined within DFTB. On average, S55 carrierd a more positive charge than S24 as nucleophilic target, and S24 a more negative charge than S55 as leaving group. Hence, S55 is the better nucleophilic target and S24 the better leaving group. These differences may be accounted to the molecular environment which modulates the charges by means of electrostatic potentials imposed on the sulfur atoms.

The electrostatic effect on the regioselectivity was further demonstrated for thiol-disulfide exchanges in a model system featuring a symmetric free energy landscape. By imposing an external potential on one sulfur atom, the charge of the touched atom changes and the symmetry of the free energy profile broken. A negative applied ESP lead to a more positive charge and hence a better nucleophilic target, but a worse leaving group. In contrast, a positive ESP lead to a more negative charge and consequently to a better leaving group, but to a worse nucleophilic target.

After studying the electrostatic effects on thiol-disulfide exchange, we focused on the energetics in **chapter 11**. DFTB is well suited for fast and efficient sampling in a QM/MM framework, but exhibits quantitative and qualitative errors when describing thiol-disulfide exchange. With DFTB, the transition state is a local minimum instead of a saddle point, and in addition, the geometry is not well reproduced. These errors were corrected with a special reaction parameterization and an artificial neural network correction which learned the energy difference between DFTB and CCSD(T). By computing the multi-dimensional free energy profile of thiol-disulfide exchange in a model system, we show that both approaches improved the accuracy while retaining low computational cost. Subsequently, both methods were applied to intramolecular thiol-disulfide exchange between S36, S78 and S79 in the von Willebrand factor's C4 domain. We find different heights of barriers for the respective reactions due to structural and electrostatic factors.

In **chapter 12**, the reparamterized S–S potential was used to perform QM/MM metadynamics of intramolecular thiol-disulfide exchange between S7, S41 and S36 in the C4 domain under mechanical stress to validate experimental data. We show that both attacks exhibit comparable barrier heights and therefore may occur with similar probability.

Bibliography

(1) Oesterhelt, D.; Stoeckenius, W. *Proc. Natl. Acad. Sci. USA* **1973**, *70*, 2853–2857.

(2) Stoeckenius, W.; Rowen, R. *J. Cell Biol.* **1967**, *34*, 365–393.

(3) Stoeckenius, W.; Kunau, W. *J. Cell Biol.* **1968**, *38*, 337–357.

(4) Danon, A.; Stoeckenius, W. *Proc. Natl. Acad. Sci. USA* **1974**, *71*, 1234–8.

(5) Stoeckenius, W.; Lozier, R. H. *J. Supramol. Struct.* **1974**, *2*, 769–774.

(6) Henderson, R.; Unwin, P. N. *Nature* **1975**, *257*, 28–32.

(7) Schobert, B.; Lanyi, J. K. *J. Biol. Chem.* **1982**, *257*, 10306–10313.

(8) Neutze, R.; Pebay-Peyroula, E.; Edman, K.; Royant, A.; Navarro, J.; Landau, E. M. *Biochim. Biophys. Acta, Biomembranes* **2002**, *1565*, 144–167.

(9) Hasegawa, N.; Jonotsuka, H.; Miki, K.; Takeda, K. *Sci. Rep.* **2018**, *8*, 13123.

(10) Mathies, R. A. *Proc. Indian Acad. Sci. – Chem. Sci.* **1991**, *103*, 283–293.

(11) Lanyi, J. K. *Annu. Rev. Physiol.* **2004**, *66*, 665–688.

(12) Kateriya, S. *News Physiol. Sci.* **2004**, *19*, 133–137.

(13) Luck, M.; Mathes, T.; Bruun, S.; Fudim, R.; Hagedorn, R.; Tran Nguyen, T. M.; Kateriya, S.; Kennis, J. T. M.; Hildebrandt, P.; Hegemann, P. *J. Biol. Chem.* **2012**, *287*, 40083–40090.

(14) Luck, M.; Bruun, S.; Keidel, A.; Hegemann, P.; Hildebrandt, P. *FEBS letters* **2015**, *589*, 1067–71.

(15) Luck, M.; Hegemann, P. *J. Plant Physiol.* **2017**, *217*, 77–84.

(16) Wolff, F.-E. C. Computer simulations of light-triggered processes of chromophores in complex environments, Karlsruhe, 2018.

(17) Mayer, J. M. *Annu. Rev. Phys. Chem.* **2004**, *55*, 363–390.

(18) Hammes-Schiffer, S. *Energy Environ. Sci.* **2012**, *5*, 7696–7703.

(19) Stubbe, J.; Nocera, D. G.; Yee, C. S.; Chang, M. C. Y. *Chem. Rev.* **2003**, *103*, 2167–2202.

(20) Minnihan, E. C.; Nocera, D. G.; Stubbe, J. *Acc. Chem. Res* **2013**, *46*, 2524–2535.

(21) Reinhardt, C. R.; Li, P.; Kang, G.; Stubbe, J.; Drennan, C. L.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2020**, *142*, 13768–13778.

(22) Cukier, R. I. *J. Phys. Chem.* **1994**, *98*, 2377–2381.

(23) Cukier, R. I. *J. Phys. Chem.* **1995**, *99*, 16101–16115.

(24) Cukier, R. I. *J. Phys. Chem.* **1996**, *100*, 15428–15443.

(25) Soudackov, A.; Hammes-Schiffer, S. *J. Chem. Phys* **1999**, *111*, 4672–4687.

(26) Soudackov, A.; Hammes-Schiffer, S. *J. Chem. Phys* **2000**, *113*, 2385–2396.

(27) Decornez, H.; Hammes-Schiffer, S. *J. Phys. Chem. A* **2000**, *104*, 9370–9384.

(28) Huynh, M. H. V.; Meyer, T. J. *Chem. Rev.* **2007**, *107*, 5004–5064.

(29) Migliore, A.; Polizzi, N. F.; Therien, M. J.; Beratan, D. N. *Chem. Rev.* **2014**, *114*, 3381–3465.

(30) Kaila, V. R. I.; Verkhovsky, M. I.; Wikström, M. *Chem. Rev.* **2010**, *110*, 7062–7081.

(31) Weinberg, D. R.; Gagliardi, C. J.; Hull, J. F.; Murphy, C. F.; Kent, C. A.; Westlake, B. C.; Paul, A.; Ess, D. H.; McCafferty, D. G.; Meyer, T. J. *Chem. Rev.* **2012**, *112*, 4016–4093.

(32) Elgrishi, N.; McCarthy, B. D.; Rountree, E. S.; Dempsey, J. L. *ACS Catal.* **2016**, *6*, 3644–3659.

(33) Tyburski, R.; Liu, T.; Glover, S. D.; Hammarström, L. *J. Am. Chem. Soc.* **2021**, *143*, 560–576.

(34) Brown, S. E.; Shakib, F. A. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2535–2556.

(35) Gillet, N.; Elstner, M.; Kubař, T. *J. Chem. Phys.* **2018**, *149*, 072328.

(36) Metcalfe, C.; Cresswell, P.; Ciaccia, L.; Thomas, B.; Barclay, A. N. *Open Biol.* **2011**, *1*, 110010.

(37) Chiu, J.; Hogg, P. J. *J. Biol. Chem.* **2019**, *294*, 2949–2960.

(38) Holmgren, A. *J. Biol. Chem.* **1989**, *264*, 13963–13966.

(39) Kozlov, G.; Määttänen, P.; Thomas, D. Y.; Gehring, K. *FEBS J.* **2010**, *277*, 3924–3936.

(40) Hamlin, T. A.; Swart, M.; Bickelhaupt, F. M. *ChemPhysChem* **2018**, *19*, 1315–1330.

(41) Bell, G. *Science* **1978**, *200*, 618–627.

(42) Li, W.; Gräter, F. *J. Am. Chem. Soc.* **2010**, *132*, 16790–16795.

(43) Singh, R.; Whitesides, G. M. *J. Am. Chem. Soc.* **1990**, *112*, 1190–1197.

(44) Bach, R. D.; Dmitrenko, O.; Thorpe, C. *J. Org. Chem,* **2008**, *73*, 12–21.

(45) Fernandes, P. A.; Ramos, M. J. *Chem. Eur. J.* **2004**, *10*, 257–266.

(46) Jensen, F., *Introduction to Computational Chemistry*, Third edition; Wiley: 2017.

(47) Lewars, E. G., *Computational Chemistry; Introduction to the Theory and Applications of Molecular and Quantum Mechanics*; Springer International Publishing: 2018.

(48) Hastie, T.; Tibshirani, R.; Friedman, J. H., *The elements of Statistical Learning: Data Mining, Inference, and prediction*; Springer: 2017.

(49) Bishop, C. M., *Pattern Recognition and Machine Learning*; Springer: New York, NY, 2006.

(50) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947–12957.

(51) Seifert, G.; Porezag, D.; Frauenheim, T. *Int. J. Quantum Chem.* **1996**, *58*, 185–192.

(52) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(53) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861–10873.

(54) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

(55) Witek, H. A.; Irle, S.; Morokuma, K. *J. Phys. Chem.* **2004**, *121*, 5163–5170.

(56) Hourahine, B. DFTB+: general package for performing fast atomistic calculations, `https://github.com/bhourahine/dftbplus/tree/perturbation`, Accessed: 2021-12-30.

(57) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.

(58) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(59) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

(60) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(61) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(62) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13749–13754.

(63) Wang, L.; Friesner, R. A.; Berne, B. J. *J. Phys. Chem. B* **2011**, *115*, 9431–9438.

(64) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(65) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. *International Journal of Quantum Chemistry* **2015**, *115*, 1094–1101.

(66) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. *The Journal of Physical Chemistry Letters* **2015**, *6*, 2326–2331.

(67) Behler, J.; Parrinello, M. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(68) Bartók, A. P.; Kondor, R.; Csányi, G. *Phys. Rev. B* **2013**, *87*, 184115.

(69) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. *Chem. Phys.* **2019**, *150*, 064105.

(70) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. *Chem. Phys.* **2020**, *152*, 044107.

(71) Maag, D.; Mast, T.; Elstner, M.; Cui, Q.; Kubař, T. *Proc. Natl. Acad. Sci. USA* **2021**, *118*.

(72) Mast, T. Simulation of Long-Range Proton Transfer - Development and Application, Ph.D. Thesis, 2019.

(73) Nango, E. et al. *Science* **2016**, *354*, 1552–1557.

(74) Wolf, S.; Freier, E.; Gerwert, K. *Biophys. J.* **2014**, *107*, 174–184.

(75) Tripathi, R.; Forbert, H.; Marx, D. *J. Phys. Chem. B* **2019**, *123*, 9598–9608.

(76) Luecke, H.; Schobert, B.; Richter, H. T.; Cartailler, J. P.; Lanyi, J. K. *J. Mol. Biol.* **1999**, *291*, 899–911.

(77) Furutani, Y.; Shibata, M.; Kandori, H. *Photochem. Photobiol. Sci.* **2005**, *4*, 661–666.

(78) Edman, K.; Nollert, P.; Royant, A.; Belrhali, H.; Pebay-Peyroula, E.; Hajdu, J.; Neutze, R.; Landau, E. M. *Nature* **1999**, *401*, 822–826.

(79) Schobert, B.; Cupp-Vickery, J.; Hornak, V.; Smith, S. O.; Lanyi, J. K. *J. Mol. Biol.* **2002**, *321*, 715–726.

(80) Bondar, A.-N.; Fischer, S.; Suhai, S.; Smith, J. C. *J. Phys. Chem. B* **2005**, *109*, 14786–14788.

(81) Mak-Jurkauskas, M. L.; Bajaj, V. S.; Hornstein, M. K.; Belenky, M.; Griffin, R. G.; Herzfeld, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 883–888.

(82) Wolter, T.; Elstner, M.; Fischer, S.; Smith, J. C.; Bondar, A.-N. *J. Phys. Chem. B* **2015**, *119*, 2229–2240.

(83) Bondar, A. N.; Elstner, M.; Suhai, S.; Smith, J. C.; Fischer, S. *Structure* **2004**, *12*, 1281–1288.

(84) Bondar, A.-N.; Baudry, J.; Suhai, S.; Fischer, S.; Smith, J. C. *J. Phys. Chem. B* **2008**, *112*, 14729–14741.

(85) Lanyi, J. K.; Schobert, B. *J. Mol. Biol.* **2007**, *365*, 1379–1392.

(86) Clemens, M.; Phatak, P.; Cui, Q.; Bondar, A. N.; Elstner, M. *J. Phys. Chem. B* **2011**, *115*, 7129–7135.

(87) Zimanyi, L.; Cao, Y.; Needleman, R.; Ottolenghi, M.; Lanyi, J. K. *Biochemistry* **1993**, *32*, 7669–7678.

(88) Lanyi, J. K. *Biochim. Biophys. Acta, Bioenergetics* **2006**, *1757*, Proton Transfer Reactions in Biological Systems, 1012–1018.

(89) Lorenz-Fonfria, V. A.; Saita, M.; Lazarova, T.; Schlesinger, R.; Heberle, J. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E10909–E10918.

(90) Freier, E.; Wolf, S.; Gerwert, K. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 11435–11439.

(91) Wolf, S.; Freier, E.; Cui, Q.; Gerwert, K. *Chem. Phys.* **2014**, *141*, 22D524.

(92) Dioumaev, A. K.; Brown, L. S.; Needleman, R.; Lanyi, J. K. *Biochemistry* **2001**, *40*, PMID: 11560478, 11308–11317.

(93) Brown, L. S.; Lanyi, J. K. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 1731–1734.

(94) Smith, S. O.; Curry, B.; Mathies, R.; Pardoen, J. A.; Mulder, P. P.; Lugtenburg, J. *Biochemistry* **1983**, *22*, 6141–6148.

(95) Cao, Y.; Váró, G.; Klinger, A. L.; Czajkowsky, D. M.; Braiman, M. S.; Needleman, R.; Lanyi, J. K. *Biochemistry* **1993**, *32*, 1981–1990.

(96) Balashov, S. P.; Lu, M.; Imasheva, E. S.; Govindjee, R.; Ebrey, T. G.; Othersen, B.; Chen, Y.; Crouch, R. K.; Menick, D. R. *Biochemistry* **1999**, *38*, 2026–2039.

(97) Ames, J. B.; Mathies, R. A. *Biochemistry* **1990**, *29*, 7181–7190.

(98) Richter, H.-T.; Needleman, R.; Kandori, H.; Maeda, A.; Lanyi, J. K. *Biochemistry* **1996**, *35*, 15461–15466.

(99) van Stokkum, I. H. M.; Lozier, R. H. *J. Phys. Chem. B* **2002**, *106*, 3477–3485.

(100) Zscherp, C.; Heberle, J. *J. Phys. Chem. B* **1997**, *101*, 10542–10547.

(101) Dioumaev, A. K.; Brown, L. S.; Needleman, R.; Lanyi, J. K. *Biochemistry* **1999**, *38*, 10070–10078.

(102) Zscherp, C.; Schlesinger, R.; Heberle, J. *Biochem. Biophys. Res. Commun.* **2001**, *283*, 57–63.

(103) Ge, X.; Gunner, M. R. *Proteins Struct. Funct. Bioinf.* **2016**, *84*, 639–654.

(104) Phatak, P.; Frähmcke, J. S.; Wanko, M.; Hoffmann, M.; Strodel, P.; Smith, J. C.; Suhai, S.; Bondar, A. N.; Elstner, M. *J. Am. Chem. Soc.* **2009**, *131*, 7064–7078.

(105) Balashov, S. P. *Biochim. Biophys. Acta, Bioenerg.* **2000**, *1460*, 75–94.

(106) Ghosh, N.; Prat-Resina, X.; Gunner, M. R.; Cui, Q. *Biochemistry* **2009**, *48*, 2468–2485.

(107) Phatak, P. V. Investigation of Proton Transfer Pathways in Bacteriorhodopsin with Multi-Length-Scale Simulations, Ph.D. Thesis, TU Braunschweig, 2009.

(108) Imasheva, E. S.; Balashov, S. P.; Ebrey, T. G.; Chen, N.; Crouch, R. K.; Menick, D. R. *Biophys. J.* **1999**, *77*, 2750–2763.

(109) Otto, H.; Marti, T.; Holz, M.; Mogi, T.; Stern, L. J.; Engel, F.; Khorana, H. G.; Heyn, M. P. *Biophysics* **1990**, *87*, 1018–1022.

(110) Riccardi, D.; König, P.; Prat-Resina, X.; Yu, H.; Elstner, M.; Frauenheim, T.; Cui, Q. *J. Am. Chem. Soc.* **2006**, *128*, 16302–16311.

(111) Riccardi, D.; König, P.; Guo, H.; Cui, Q. *Biochemistry* **2008**, *47*, 2369–2378.

(112) König, P. H.; Ghosh, N.; Hoffmann, M.; Elstner, M.; Tajkhorshid, E.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. A* **2006**, *110*, 548–563.

(113) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(114) Nina, M.; Roux, B.; Smith, J. C. *Biophys. J.* **1995**, *68*, 25–39.

(115) Baudry, J.; Crouzy, S.; Roux, B.; Smith, J. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1018–1024.

(116) Tajkhorshid, E.; Paizs, B.; Suhai, S. *J. Phys. Chem. B* **1997**, *101*, 8021–8028.

(117) Tajkhorshid, E.; Suhai, S. *J. Phys. Chem. B* **1999**, *103*, 5581–5590.

(118) Tajkhorshid, E.; Baudry, J.; Schulten, K.; Suhai, S. *Biophys. J.* **2000**, *78*, 683–693.

(119) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(120) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(121) Lindahl, E.; Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B. *J. Chem. Theory Comput.* **2010**, *6*, 459–466.

(122) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1–2*, 19–25.

(123) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(124) Kandt, C.; Ash, W. L.; Tieleman, D. P. *Methods* **2007**, *41*, 475–488.

(125) Tieleman, D. P. *Biochemistry* **1998**, *37*, 17554–17561.

(126) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(127) Bussi, G. *Mol. Phys.* **2014**, *112*, 379–384.

(128) Goyal, P.; Qian, H. J.; Irle, S.; Lu, X.; Roston, D.; Mori, T.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2014**, *118*, 11007–11027.

(129) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(130) Kubař, T.; Welke, K.; Groenhof, G. *J. Comput. Chem.* **2015**, *36*, 1978–1989.

(131) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. *J. Phys. Chem. B* **2006**, *110*, 3533–3539.

(132) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(133) Hourahine, B. et al. *J. Chem. Phys.* **2020**, *152*, 124101.

(134) Bonomi, M. et al. *Nat. Methods* **2019**, *16*, 670–673.

(135) Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

(136) Gunner, M. R.; Amin, M.; Zhu, X.; Lu, J. *Biochim. Biophys. Acta, Bioenerg.* **2013**, *1827*, 892–913.

(137) Misra, S.; Martin, C.; Kwon, O.-H.; Ebrey, T. G.; Chen, N.; Crouch, R. K.; Menick, D. R. *Photochem. Photobiol.* **1997**, *66*, 774–783.

(138) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.

(139) Lu, X.; Fang, D.; Ito, S.; Okamoto, Y.; Ovchinnikov, V.; Cui, Q. *Mol. Simul.* **2016**, *42*, 1056–1078.

(140) Phatak, P.; Ghosh, N.; Yu, H.; Cui, Q.; Elstner, M. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 19672–19677.

(141) Goyal, P.; Ghosh, N.; Phatak, P.; Clemens, M.; Gaus, M.; Elstner, M.; Cui, Q. *J. Am. Chem. Soc.* **2011**, *133*, 14981–14997.

(142) Mogi, T.; Stern, L. J.; Marti, T.; Chao, B. H.; Khorana, H. G. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 4148–4152.

(143) Fitch, C. A.; Platzer, G.; Okon, M.; Garcia-Moreno E., B.; McIntosh, L. P. *Protein Sci.* **2015**, *24*, 752–761.

(144) Herzfeld, K. F. *Ann. Phys.* **1919**, *364*, 635–667.

(145) Karton, A.; O'Reilly, R. J.; Radom, L. *J. Phys. Chem. A* **2012**, *116*, 4211–4221.

(146) Hammes-Schiffer, S. *Acc. Chem. Res.* **2001**, *34*, 273–281.

(147) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.

(148) Goyal, P.; Lu, J.; Yang, S.; Gunner, M. R.; Cui, Q. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18886–18891.

(149) Son, C. Y.; Yethiraj, A.; Cui, Q. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E8830–E8836.

(150) Liang, R. B.; Swanson, J. M. J.; Peng, Y. X.; Wikström, M.; Voth, G. A. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 7420–7425.

(151) Mühlbauer, M. E.; Saura, P.; Nuber, F.; Di Luca, A.; Friedrich, T.; Kaila, V. R. I. *J. Am. Chem. Soc.* **2020**, *142*, 13718–13728.

(152) Armstrong, C. T.; Mason, P. E.; Anderson, J. L. R.; Dempsey, C. E. *Sci. Rep.* **2016**, *6*, 21759.

(153) Rouhani, S.; Cartailler, J. P.; Facciotti, M. T.; Walian, P.; Needleman, R.; Lanyi, J. K.; Glaeser, R. M.; Luecke, H. *J. Mol. Biol.* **2001**, *313*, 615–628.

(154) Song, Y.; Gunner, M. R. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 16377–16382.

(155) Sasaki, J.; Brown, L.; Chon, Y.; Kandori, H.; Maeda, A.; Needleman, R.; Lanyi, J. *Science* **1995**, *269*, 73–75.

(156) Ganea, C.; Tittor, J.; Bamberg, E.; Oesterhelt, D. *Biochim. Biophys. Acta, Biomembranes* **1998**, *1368*, 84–96.

(157) Cui, Q. *J. Chem. Phys.* **2016**, *145*, 140901.

(158) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. *Chem. Rev.* **2016**, *116*, 5301–5337.

(159) Faham, S.; Yang, D.; Bare, E.; Yohannan, S.; Whitelegge, J. P.; Bowie, J. U. *J. Mol. Biol.* **2004**, *335*, 297–305.

(160) Nakanishi, T.; Kanada, S.; Murakami, M.; Ihara, K.; Kouyama, T. *Biophys. J.* **2013**, *104*, 377–85.

(161) Vogeley, L.; Sineshchekov, O. A.; Trivedi, V. D.; Sasaki, J.; Spudich, J. L.; Luecke, H. *Science* **2004**, *306*, 1390–3.

(162) Gautier, A.; Mott, H. R.; Bostock, M. J.; Kirkpatrick, J. P.; Nietlispach, D. *Nat. Struct. Mol. Biol.* **2010**, *17*, 768–74.

(163) Furuse, M. et al. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, *71*, 2203–16.

(164) Kato, H. E.; Zhang, F.; Yizhar, O.; Ramakrishnan, C.; Nishizawa, T.; Hirata, K.; Ito, J.; Aita, Y.; Tsukazaki, T.; Hayashi, S.; Hegemann, P.; Maturana, A. D.; Ishitani, R.; Deisseroth, K.; Nureki, O. *Nature* **2012**, *482*, 369–374.

(165) Guo, Y.; Beyle, F. E.; Bold, B. M.; Watanabe, H. C.; Koslowski, A.; Thiel, W.; Hegemann, P.; Marazzi, M.; Elstner, M. *Chem. Sci.* **2016**, *7*, 3879–3891.

(166) Weber, W.; Thiel, W. *Theo. Chem. Acc.* **2000**, *103*, 495–506.

(167) Koslowski, A.; Beck, M. E.; Thiel, W. *J. Comput. Chem.* **2003**, *6*, 714–726.

(168) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.11, last accessed on 04 March 2022.

(169) Böser, J. Calculations of Proton-Coupled Electron Transfers, MA thesis, Karlsruhe Institute of Technology, 2022.

(170) Maag, D. QM/MM Studies of Proton-Coupled Electron Transfer in Ribonucleotide Reductase, MA thesis, Karlsruhe Institute of Technology, 2018.

(171) Uhlin, U.; Eklund, H. *Nature* **1994**, *370*, 533–539.

(172) Maag, D.; Putzu, M.; Gómez-Flores, C. L.; Gräter, F.; Elstner, M.; Kubař, T. *Phys. Chem. Chem. Phys.* **2021**, *23*, 26366–26375.

(173) Alegre-Cebollada, J.; Kosuri, P.; Rivas-Pardo, J. A.; Fernández, J. M. *Nat. Chem.* **2011**, *3*, 882–887.

(174) Wiita, A. P.; Ainavarapu, S. R. K.; Huang, H. H.; Fernandez, J. M. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 7222–7227.

(175) Grandbois, M.; Beyer, M.; Rief, M.; Clausen-Schaumann, H.; Gaub, H. E. *Science* **1999**, *283*, 1727–1730.

(176) Stacklies, W.; Vega, M. C.; Wilmanns, M.; Gräter, F. *PLoS Comput. Biol.* **2009**, *5*, e1000306.

(177) Kolšek, K.; Aponte-Santamaría, C.; Gräter, F. *Sci. Rep.* **2017**, *7*, 9858.

(178) Nagy, P. *Antioxid. Redox Signaling* **2013**, *18*, 1623–1641.

(179) Bach, R. D.; Dmitrenko, O.; Thorpe, C. *J. Org. Chem* **2008**, *73*, 12–21.

(180) Wu, C.; Belenda, C.; Leroux, J.-C.; Gauthier, M. A. *Chem. Eur. J.* **2011**, *17*, 10064–10070.

(181) Snyder, G. H.; Cennerazzo, M. J.; Karalis, A. J.; Locey, D. *Biochemistry* **1981**, *20*, 6509–6519.

(182) Snyder, G. H.; Reddy, M. K.; Cennerazzo, M. J.; Field, D. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **1983**, *749*, 219–226.

(183) Britto, P. J.; Knipling, L.; Wolff, J. *J. Biol. Chem.* **2002**, *277*, 29018–29027.

(184) Hansen, R. E.; Østergaard, H.; Winther, J. R. *Biochemistry* **2005**, *44*, 5899–5906.

(185) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950–1958.

(186) Nosé, S.; Klein, M. L. *Mol. Phys.* **1983**, *50*, 1055–1076.

(187) Leontyev, I.; Stuchebrukhov, A. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2613–2626.

(188) Kirby, B. J.; Jungwirth, P. *J. Phys. Chem. Lett.* **2019**, *10*, 7531–7536.

(189) Putzu, M.; Gräter, F.; Elstner, M.; Kubař, T. *Phys. Chem. Chem. Phys.* **2018**, *20*, 16222–16230.

(190) Wilson, J. M.; Bayer, R. J.; Hupe, D. J. *J. Am. Chem. Soc.* **1977**, *99*, 7922–7926.

(191) Pappas, J. A. *J. Chem. Soc., Perkin Trans. 2* **1979**, 67–70.

(192) Gómez-Flores, C. L.; Maag, D.; Kansari, M.; Vuong, V.-Q.; Irle, S.; Gräter, F.; Kubař, T.; Elstner, M. *J. Chem. Theory Comput.* **2022**.

(193) Neves, R. P. P.; Fernandes, P. A.; Varandas, A. J. C.; Ramos, M. J. *J. Chem. Theory Comput.* **2014**, *10*, 4842–4856.

(194) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(195) Neves, R. P. P.; Fernandes, P. A.; Ramos, M. J. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E4724–E4733.

(196) Gaus, M.; Lu, X.; Elstner, M.; Cui, Q. *J. Chem. Theory Comput.* **2014**, *10*, 1518–1537.

(197) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. *J. Chem. Phys.* **1995**, *103*, 4129–4137.

(198) Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J. *Chem. Sci.* **2018**, *9*, 2261–2269.

(199) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. *J. Chem. Phys.* **2018**, *148*, 241722.

(200) Shen, L.; Yang, W. *J. Chem. Theory Comput.* **2018**, *14*, 1442–1455.

(201) Böselt, L.; Thürlemann, M.; Riniker, S. *J. Chem. Theory Comput.* **2021**, *17*, 2641–2658.

(202) Gastegger, M.; Schütt, K. T.; Müller, K.-R. *Chem. Sci.* **2021**, *12*, 11473–11483.

(203) Zeng, J.; Giese, T. J.; Ekesan, Ş.; York, D. M. *J. Chem. Theory Comput.* **2021**.

(204) Zhang, L.; Han, J.; Wang, H.; Saidi, W. A.; Car, R.; Weinan, E. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates Inc.: Montréal, Canada, 2018, pp 4441–4451.

(205) Pan, X.; Yang, J.; Van, R.; Epifanovsky, E.; Ho, J.; Huang, J.; Pu, J.; Mei, Y.; Nam, K.; Shao, Y. *J. Chem. Theory Comput.* **2021**, *17*, 5745–5758.

(206) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(207) Zhu, J.; Vuong, V. Q.; Sumpter, B. G.; Irle, S. *MRS Commun.* **2019**, *9*, 867–873.

(208) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. *J. Chem. Phys.* **2013**, *139*, 134101.

(209) Neese, F. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(210) Neese, F. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1327.

(211) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(212) Xu, E.-R.; von Bülow, S.; Chen, P.-C.; Lenting, P. J.; Kolšek, K.; Aponte-Santamaría, C.; Simon, B.; Foot, J.; Obser, T.; Schneppenheim, R.; Gräter, F.; Denis, C. V.; Wilmanns, M.; Hennig, J. *Blood* **2019**, *133*, 366–376.

(213) Kutzki, F.; Butera, D.; Lay, A. J.; Maag, D.; Chiu, J.; Woon, H.-G.; Kubař, T.; Elstner, M.; Aponte-Santamaría, C.; Hogg, P. J.; Gräter, F. *manuscript* **2022**.

(214) Schmidt, B.; Ho, L.; Hogg, P. J. *Biochem.* **2006**, *45*, 7429–7433.

(215) Reddy, K. K. A.; Jayashree, M.; Govindu, P. C. V.; Gowd, K. H. *Proteins Struct. Funct. Bioinf.* **2021**, *89*, 599–613.

(216) Sadler, J. E. *Annu. Rev. Biochem.* **1998**, *67*, 395–424.

(217) Ruggeri, Z. M.; Mendolicchio, G. L. *Circ. Res.* **2007**, *100*, 1673–1685.

(218) Bryckaert, M.; Rosa, J.-P.; Denis, C. V.; Lenting, P. J. *Cell. Mol. Life Sci.* **2015**, *72*, 307–326.

(219) Springer, T. A. *Blood* **2014**, *124*, 1412–1425.

(220) Rauch, A.; Wohner, N.; Christophe, O. D.; Denis, C. V.; Susen, S.; Lenting, P. J. *Mediterr. J. Hematol. Infect. Dis.* **2013**, *5*, e2013046.

(221) Choi, H.; Aboulfatova, K.; Pownall, H. J.; Cook, R.; Dong, J.-F. *J. Biol. Chem.* **2007**, *282*, 35604–35611.

(222) Butera, D. et al. *Sci. Adv.* **2018**, *4*, eaaq1477.

Part IV.

Appendix

A. Bacteriorhodpsin

Interaction of the R134 sidechain with the PRG

R134 is close to E194 though, but the available X-ray structures (PDB ID 5B6V and 1C3W) never feature a salt bridge between them, rather just a hydrogen bridge between the sidechain of R134 and the backbone C=O of E194. In these structures, E194 forms a strong hydrogen bridge with E204, with the R134 guanidinium group ca. 7 Å away from the E194 carboxyl group. PRG in an open configuration (**O** state, large E194–E204 distance) makes E194 approach R134 somewhat more. Still, our HREX simulations showed that E194 never rotates towards R134 fully, and a water molecule resides between these residues most of the time. Instead, the carboxyl group of E194 tends to form a hydrogen bond with the Y83 sidechain.

The distance of the R134 guanidinium from the E194 carboxyl group was measured and correlated with the distance E194–E204 along the QM/MM metadynamics simulation of the **O**→**bR** transition, see Fig. A.1: The guanidinium of R134 slightly approaches E194 in the protonation states that have open PRG. The distance of the guanidinium from the backbone C=O of E194 was also measured, see Fig. A.2: The nearly permanent hydrogen bridge between these moieties is evident. These structural features are apparent in the representative snapshots from the QM/MM metadynamics shown in Fig. A.3.

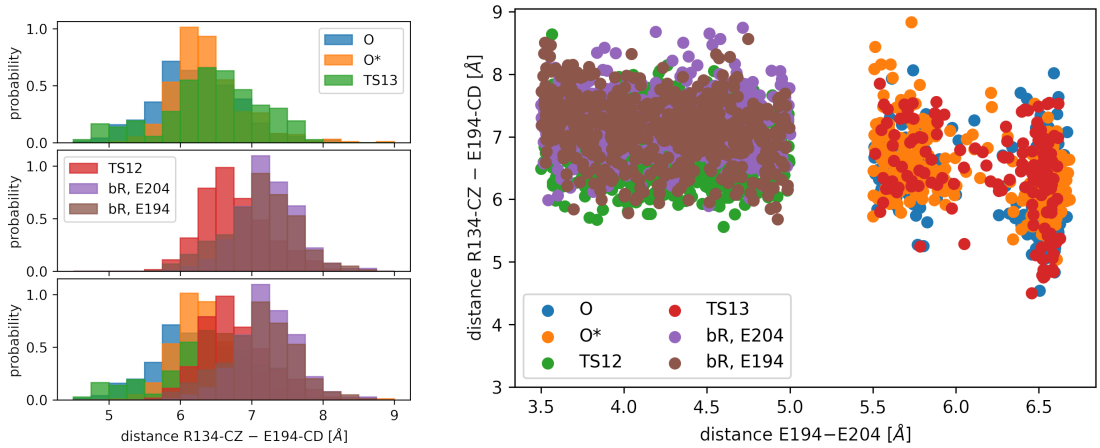

Figure A.1.: Left: histogram of R134–E194 distances for different protonation states of bR; right: PRG distance vs R134–E194 distance.

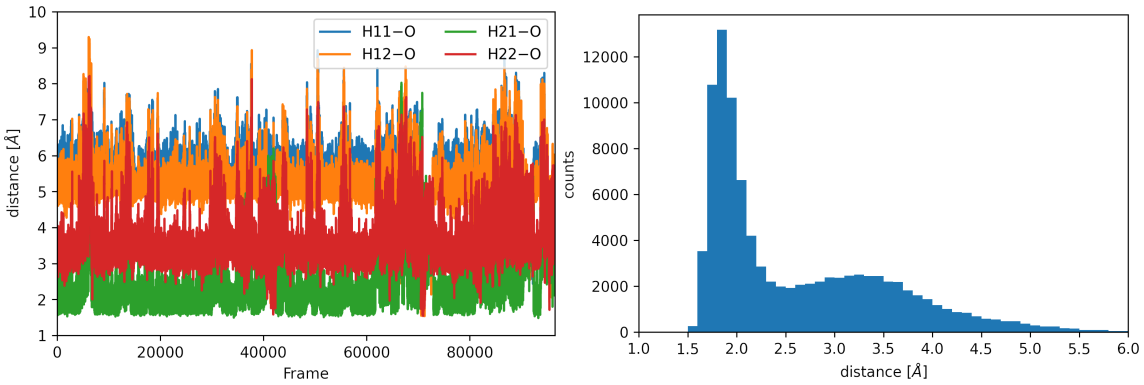

Figure A.2.: Distances between the guanidinium hydrogen atoms of R134 and the backbone carbonyl oxygen atom of E194 – time series (left) for all four hydrogen atoms and histogram (right) only for the closest hydrogen in each frame.

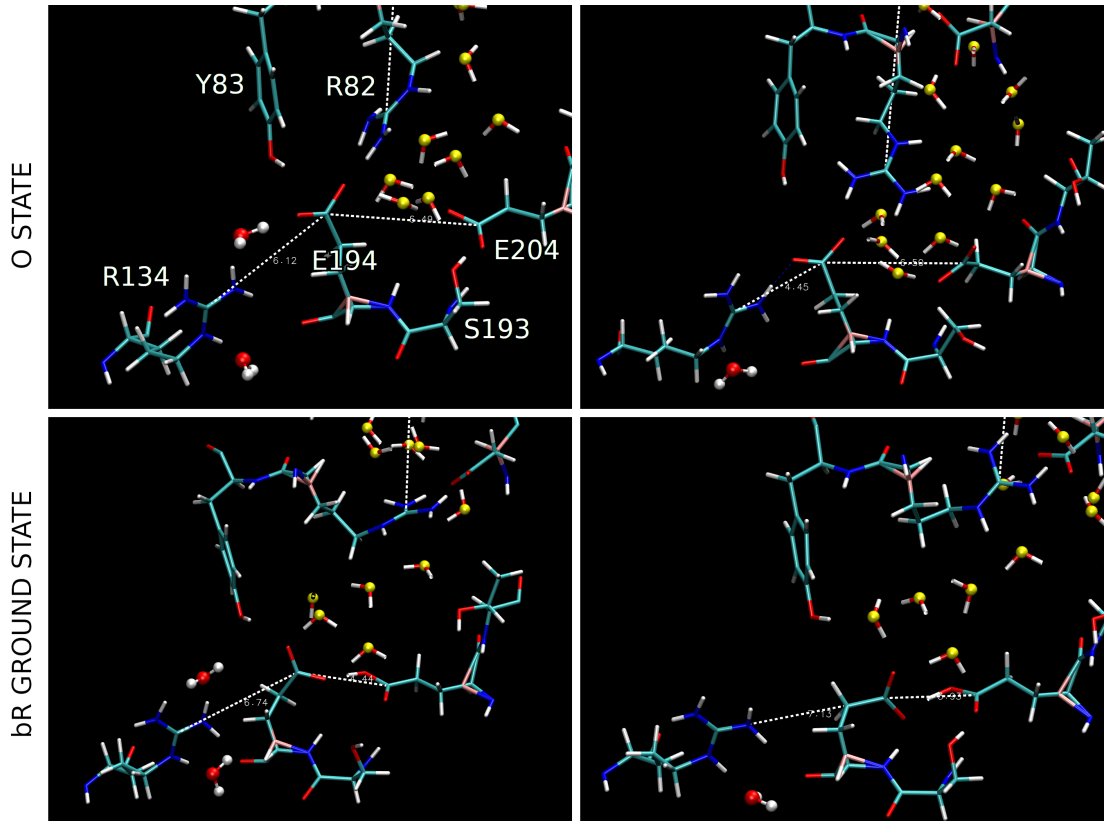

Figure A.3.: Selected snapshots from QM/MM simulations showing the orientation of the R134 sidechain.

B. Histidine Kinase Rhodopsin

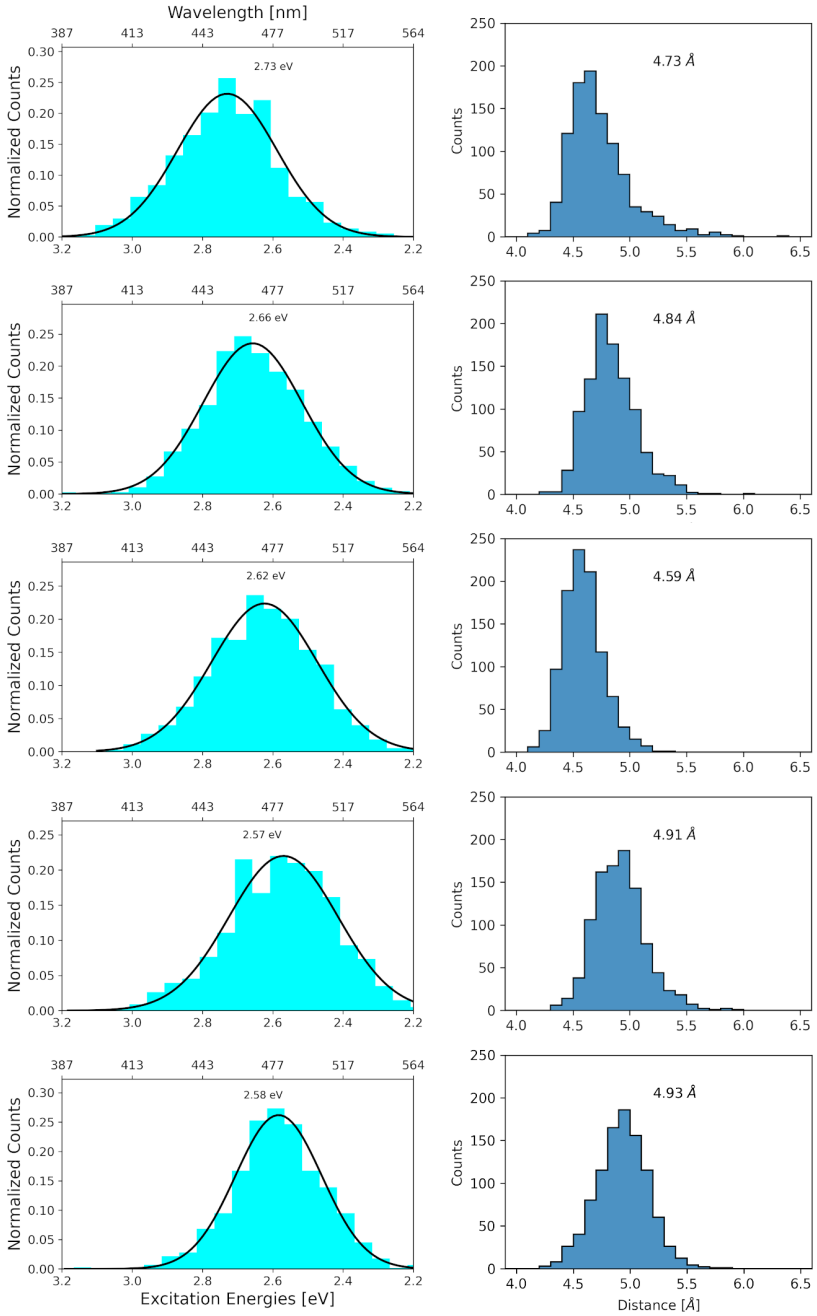

Figure B.1.: Excitation energies and N–Cγ distance histograms between the Schiff base and D239, based on 1 ns QM/MM simulations for each replica.

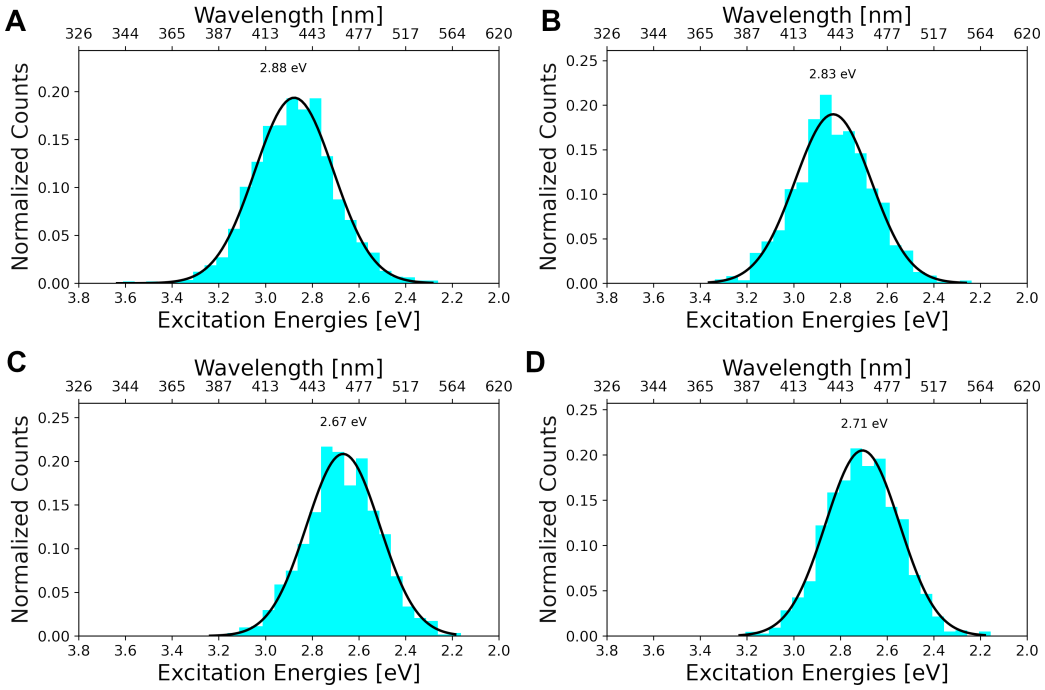

Figure B.2.: Excitation energies of the P570 state based for replica 1 with (A) r = 1Å and (B) r = 2Å, and for replica 5 with (C) r = 1Å and (D) r = 2Å.

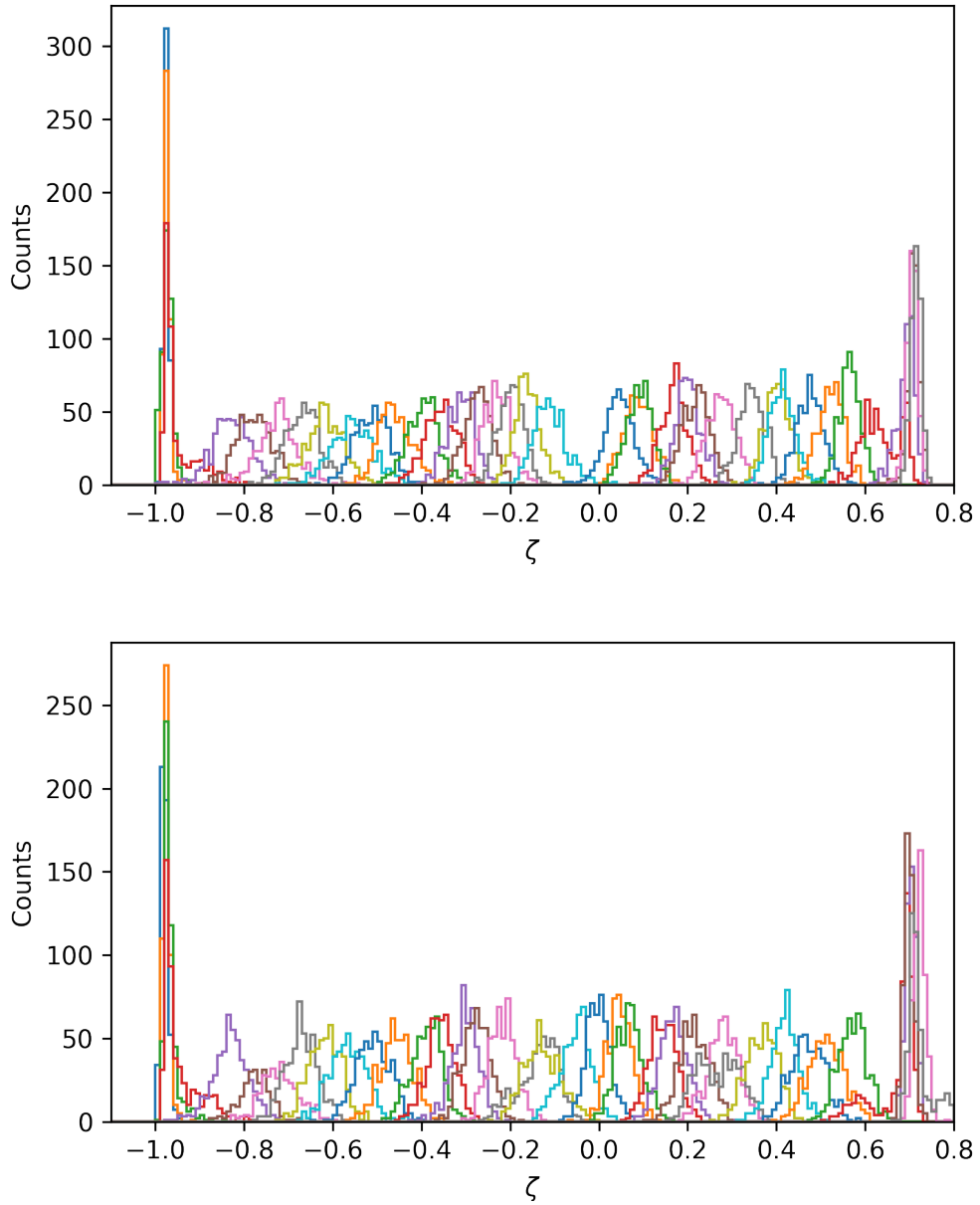

Figure B.3.: Histograms of Umbrella Sampling of the PT from the Schiff base to D239 for $r_{1\,\text{Å}}$ (restr.) and $r_{1\,\text{Å}}$ (free).

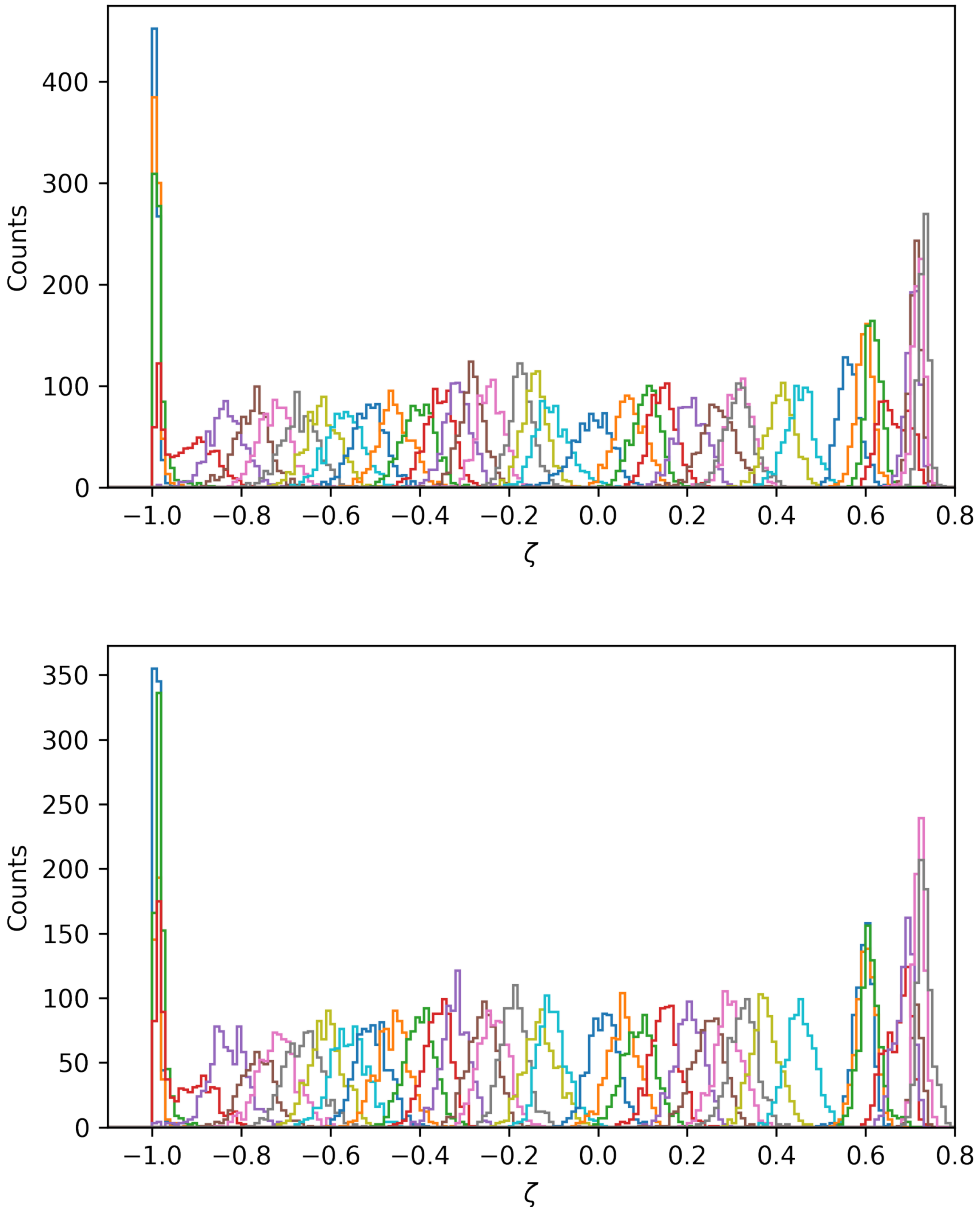

Figure B.4.: Histograms of Umbrella Sampling of the PT from the Schiff base to D239 for $r_{2\,\text{Å}}$ (restr.) and $r_{2\,\text{Å}}$ (free).

C. Immunoglobulin I27*

Distance histogram and free energy profile of the distances S32–S24 and S32–S55.

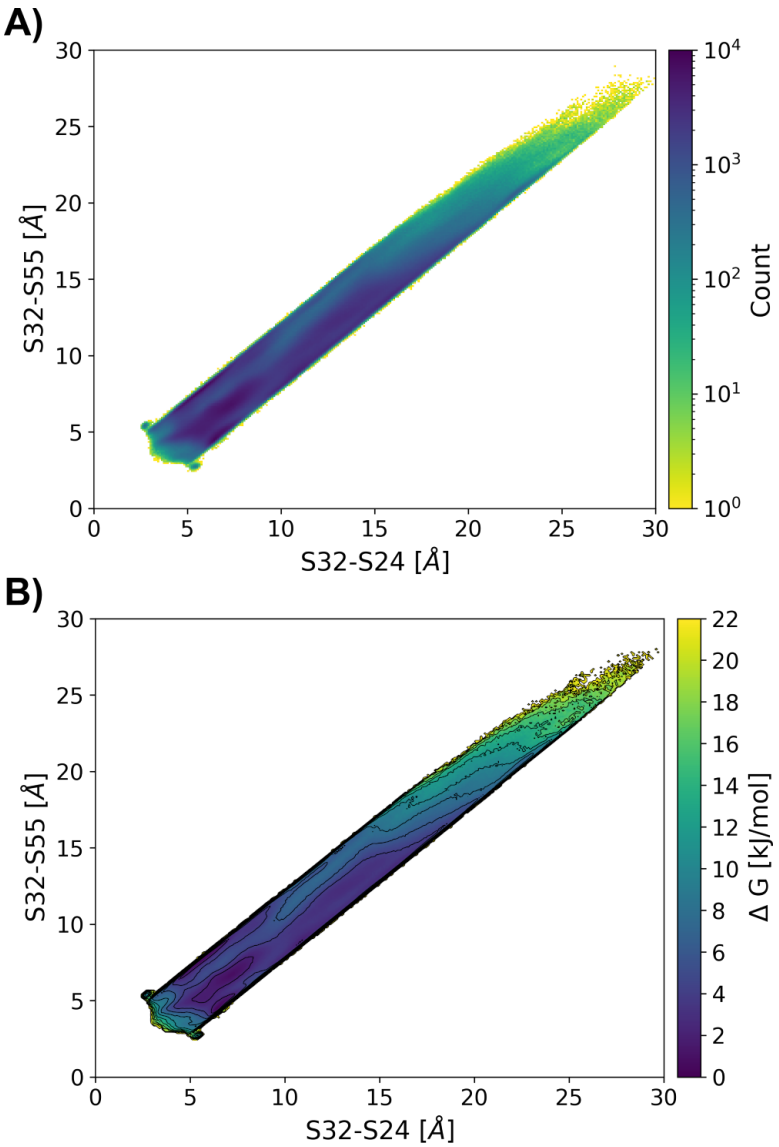

Figure C.1.: Histogram of distances S32–S24 and S32–S55 (A), converted to a free energy profile (B). Contour lines are drawn every 2 kJ/mol.

To see how the distances between S32 and the disulfide bond correlate with the length of that bond, histograms of the distances S32–S24 and S32–S55 were generated for different S24–S55 bond lengths observed. Specifically, the snapshots collected along the QM/MM simulations were classified into nine different bins with the length of bond S24–S55 ranging from 1.95 to 2.40 Å with a bin width of 0.05 Å. A separate couple of 2D histograms of the distances S32–S24 and S32–S55 were obtained over the snapshots in every bin and were converted to free energy values. Just like above, the probabilities in the "upper" and "lower" regions were summed up, and their ratios were recalculated to free energy differences, which are shown in Fig. 10.5. Also, all 2D histograms obtained are shown in Figs. C.2–C.4.

In all bins except the first (1.95–2.00 Å), S32 is closer to S55 on average, and even for large S24–S55 distances, hundreds to thousands occurrences are counted. The energy difference $G_{\text{lower}} - G_{\text{upper}}$ is positive and increases linearly with increasing S24–S55 distance. The maximum distribution of the "upper" region is found at |S24–S55| of 2.05–2.10 Å whereas the maximum of the "lower" region is found at larger bond lengths of 2.10–2.15 Å.

These findings indicate that whenever S32 is closer to S55, a longer S24–S55 bond is favored. Consequently, it may be easier for the system to stretch the bond S24–S55 further to pass to a transition state. By contrast, whenever S32 is closer to S24, a shorter bond is favored and thus a transition state is less likely to form. Still, it should be mentioned that it is difficult to interpret our findings since we are looking for very small energy differences of only a few kJ/mol.

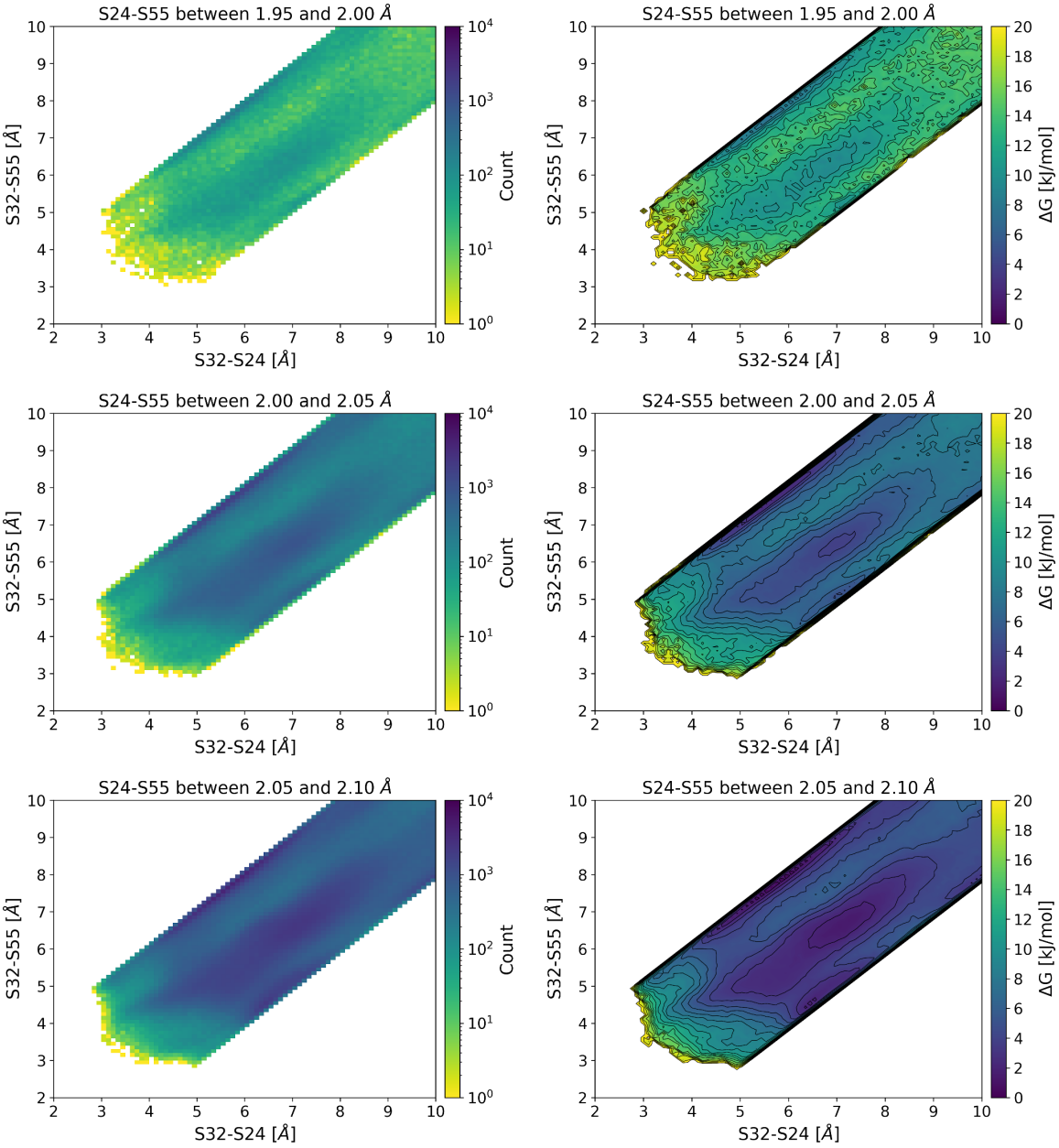

Figure C.2.: Histograms (left) and the resulting free energy profile (right) of the S32-S24 and S32-S55 distances with S24-S55 distances between 1.95 and 2.00 Å (top), 2.00 and 2.05 Å (middle), 2.05 and 2.10 Å (bottom). Contour lines are drawn every 1 kJ/mol.

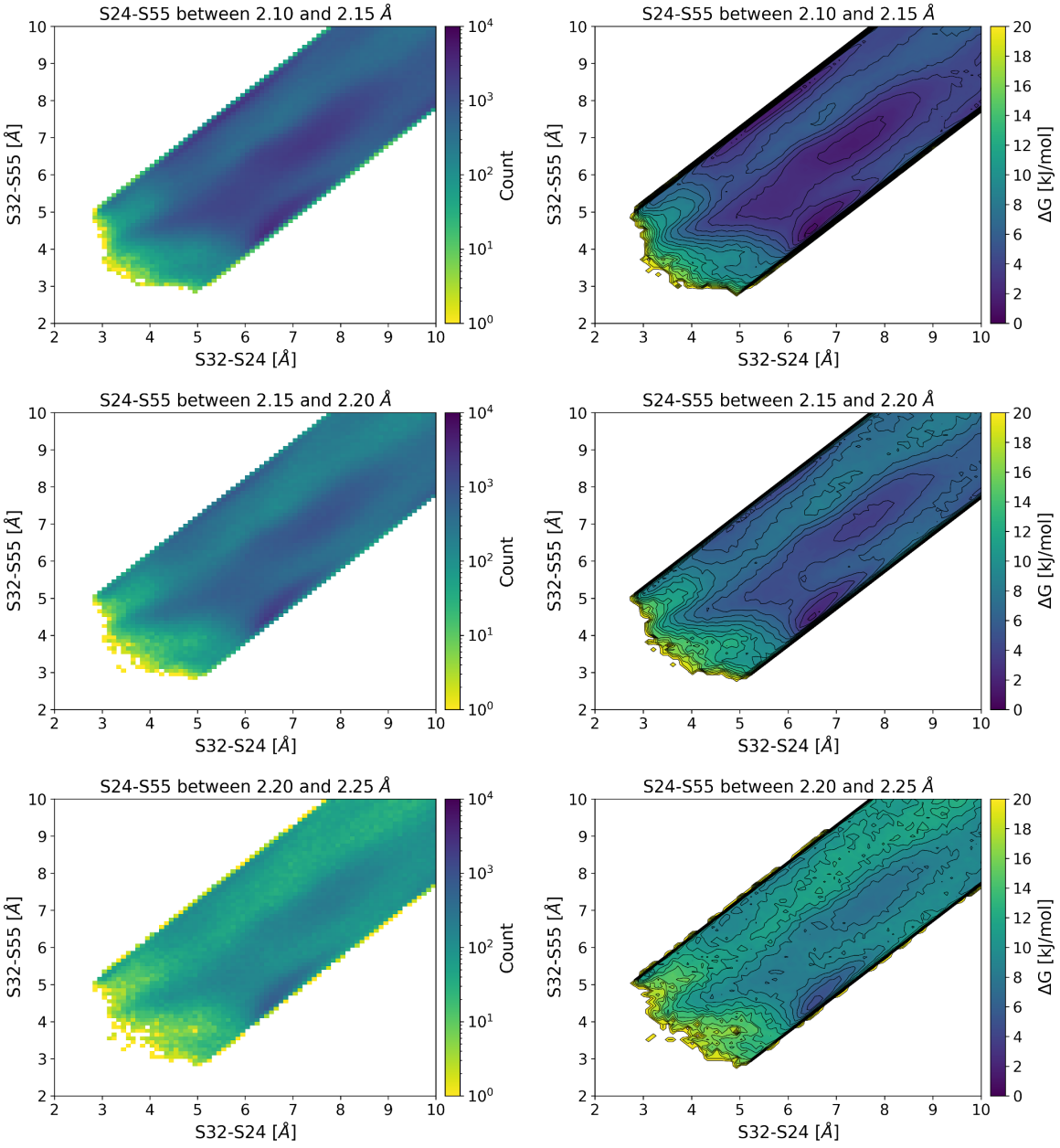

Figure C.3.: Histograms (left) and the resulting free energy profile (right) of the S32-S24 and S32-S55 distances with S24-S55 distances between 2.10 and 2.15 Å (top), 2.15 and 2.20 Å (middle), 2.20 and 2.25 Å (bottom). Contour lines are drawn every 1 kJ/mol.

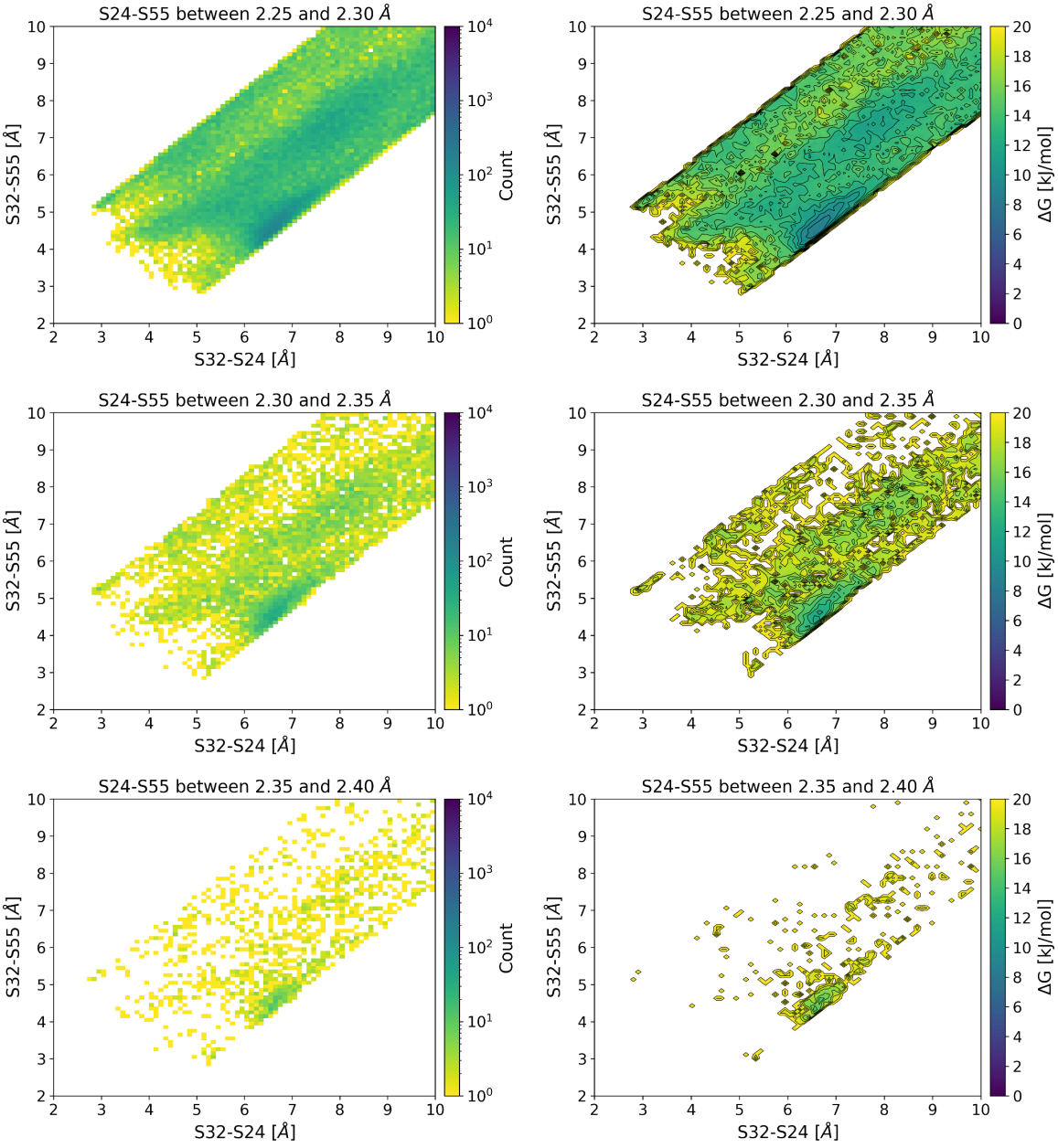

Figure C.4.: Histograms (left) and the resulting free energy profile (right) of the S32-S24 and S32-S55 distances with S24-S55 distances between 2.25 and 2.30 Å (top), 2.30 and 2.35 Å (middle), 2.35 and 2.40 Å (bottom). Contour lines are drawn every 1 kJ/mol.

Metadynamics simulation of disulfide shuffling in I27*

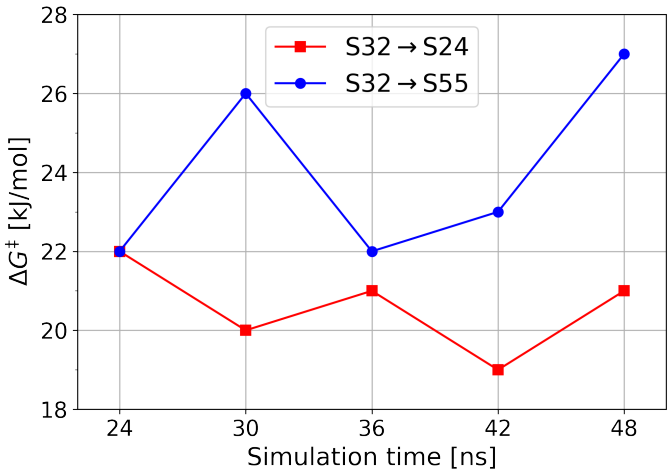

Figure C.5.: Barrier heights of the 2D metadynamics simulations of S32→S24 and S32→S55 after a total simulation time of 24 ns, 30 ns, 36 ns, 40 ns, and 48 ns.

Example of distances, charges and ESP on the sulfurs before and after a reaction.

Fig. C.6 depicts the temporal course of sulfur–sulfur distances, sulfur charges, and ESP on the sulfurs before, during and after a reaction between S32–S24 (left side) and S32–S55 (right side) in two selected simulations. When a transition state is formed, the S_{nuc}–S_{ctr} increases to ca. 2.7 Å and the S_{ctr}–S_{lg} decreases to ca. 2.7 Å, too. The S_{nuc}–S_{lg} distance is slightly smaller than the sum of $|S_{nuc}–S_{ctr}|$ and $|S_{ctr}–S_{lg}|$, thus the angle S_{nuc}–S_{ctr}–S_{lg} < 180 °. During the reaction, the negative charge of S32 ($Q(S_{nuc}) \sim -1.1$ e) is transferred to the leaving sulfur S_{lg} ($Q(S_{lg}) \sim -0.1$ e) without any charge accumulation on S_{ctr} ($Q(S_{ctr}) \sim 0$ e). Analogously, the ESP on S_{nuc} and S_{lg} interchange during a reaction whereas the ESP on S_{ctr} does not change much.

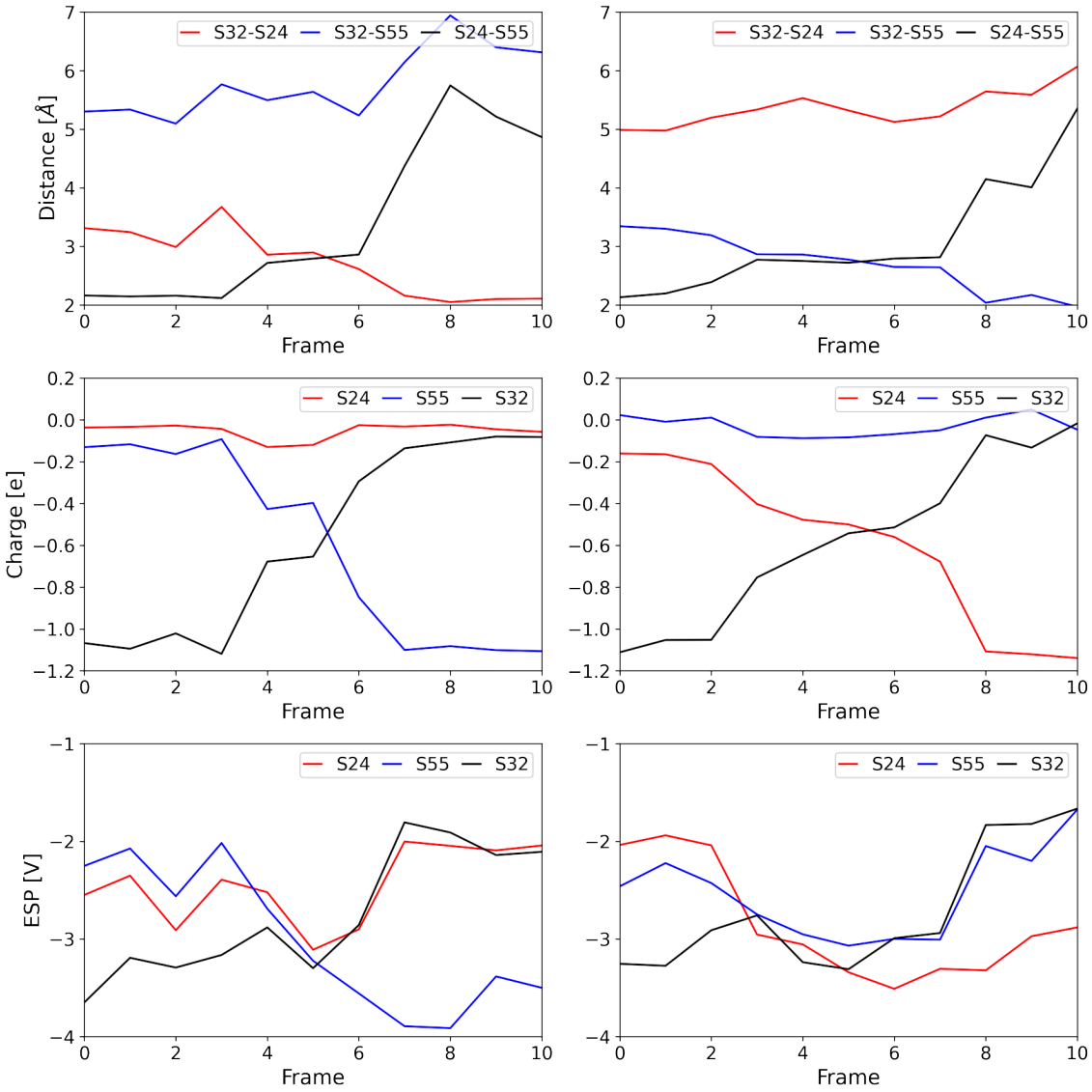

Figure C.6.: Sulfur–sulfur distances (top), sulfur charges (middle), and ESP on the sulfur atoms (bottom) before, during and after a reaction between S32–S24 (left) and S32–S55 (right) over 11 frames (5.5 ps).

Contributions to the ESP on the sulfur atoms caused by the MM environment and by the QM atoms.

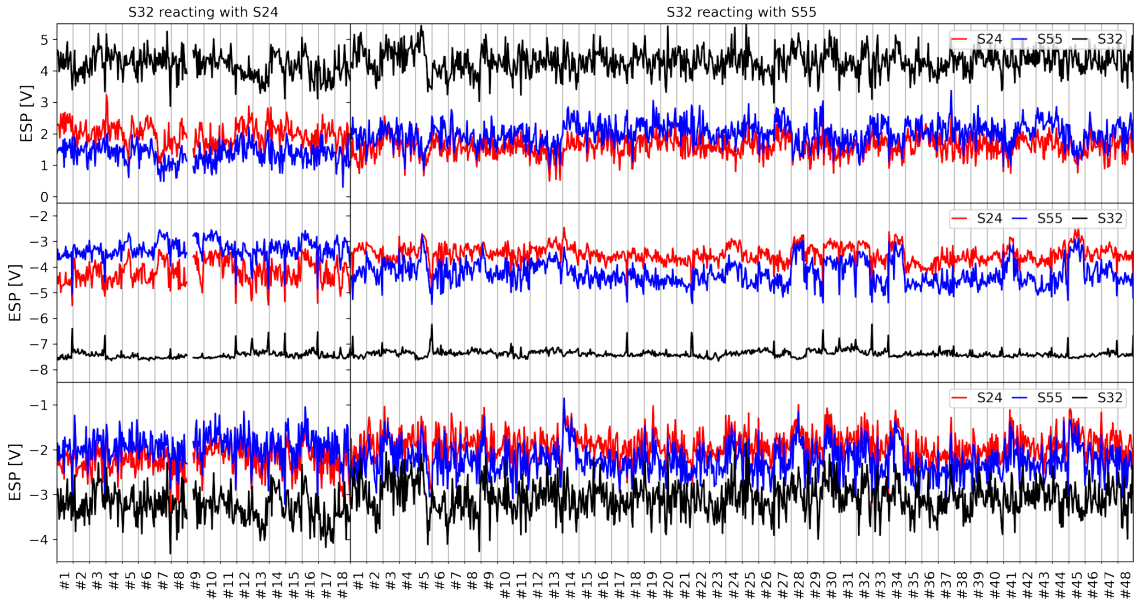

Figure C.7.: ESP on the three sulfur atoms arising from all MM atoms (top), the QM atoms (middle), and the sum of both (bottom).

Simulations of the model system

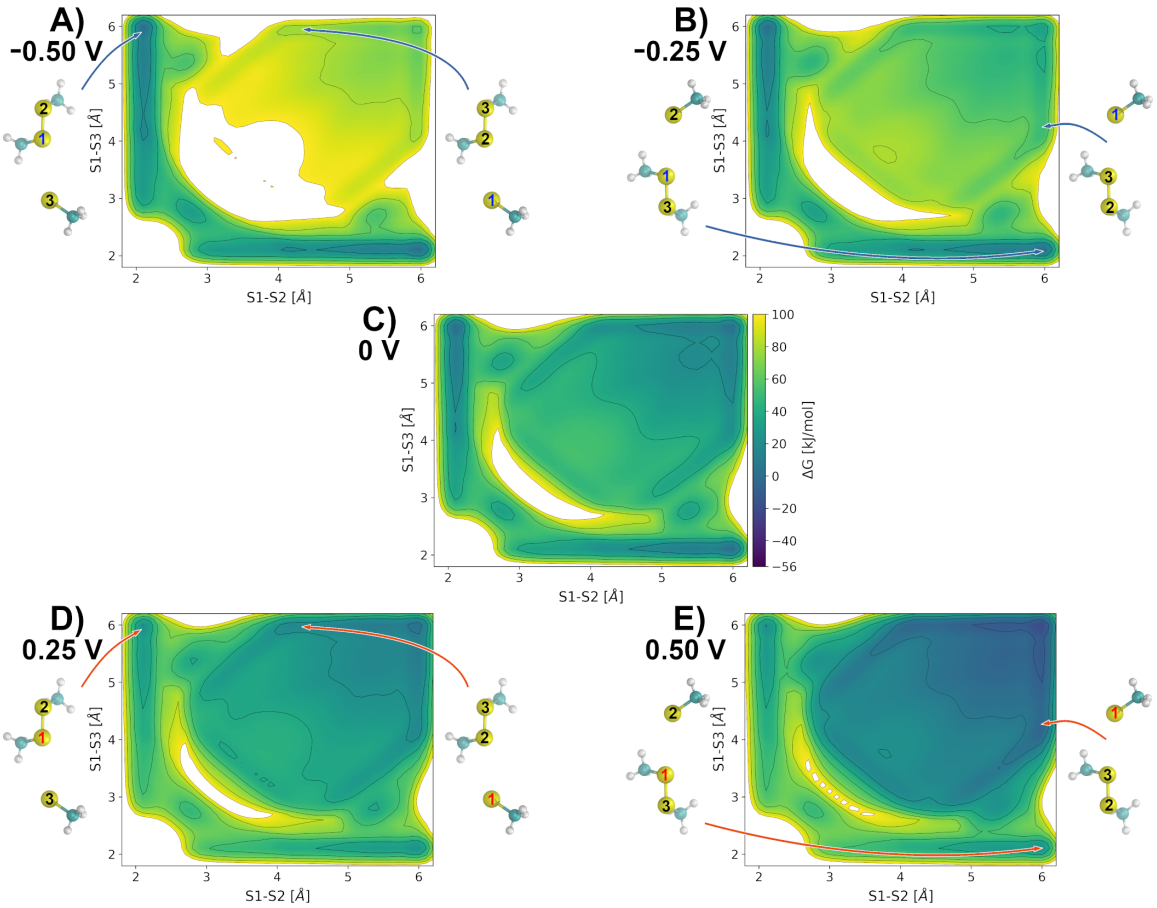

Figure C.8.: Free energy profiles of the solvated anionic trisulfide system with an additional electric potential imposed onto the atom S_1. Energies of disulfide bonds S_1–S_2 and S_1–S_3 set to zero; contour lines drawn every 20 kJ/mol.

Table C.1.: Reaction barriers in kJ/mol. *Forward and backward reactions correspond to the same process, so the barriers should be the same.

ESP_{ext}	reaction	
on S_1	S_1–$S_2 \rightleftharpoons S_1$–$S_3$*	S_1–$S_2 \rightleftharpoons S_2$–$S_3$
−0.50 V	47/47	95/36
−0.25 V	50/50	72/40
±0 V	52/52	51/49
+0.25 V	54/54	47/74
+0.50 V	56/56	40/96

D. Neural network corrected DFTB/MM methodology for thiol-disulfide exchange reactions.

PMF differences of the benchmark system.

To asses the overall accuracy of the obtained PMFs, their difference from the reference PMF, which was obtained with the CC machine learned correction (V), was calculated. For this, each PMF shown in Fig. 7 in the main text was subtracted from the ML/CC PMF (Fig. 7D), see Fig. D.1. Only energy differences between -6 and +6 kcal/mol were considered. As expected, the energy difference plots are symmetrical. The difference between the PMF obtained with DFTB/3OB (I) and the reference (Fig. D.1A) lies beyond the set limit around the transition state areas because the transition state appears at too long S–S bonds and too deep energies with 3OB parameters. Other parts of the surface along the reaction pathway agree with the reference within ca. -3 to +3 kcal/mol. In other relevant area where the sulfurs are in a triangular configuration, e.g. S1–S2 and S1–S3 ca. 4.5 Å, the maximum difference is -4 kcal/mol. The comparison of the PMF obtained with ML/B3LYP (III) with the reference PMF in Fig. D.1C shows rather small ΔG values between ca. -2 and +1 kcal/mol along the reaction pathway. For other relevant S–S–S configurations the difference becomes larger, up to -5 kcal/mol. The differences become even larger when the free energy profile obtained with method (III), the reparameterized S–S potential fitted to B3LYP data, is compared to the reference in D.1D. Along the reaction pathway ΔG lies between ca. -4 and +2 kcal/mol and goes up to -6 kcal/mol for other S–S–S configurations. The PMF obtained with the reparamterized S–S potential fitted to G3B3 data (V) lies higher in energy than the reference PMF and thus their energy differences along the exemplary pathway range from 0 to -5 kcal/mol and up to -6 kcal/mol for other conformations (Fig. D.1D).

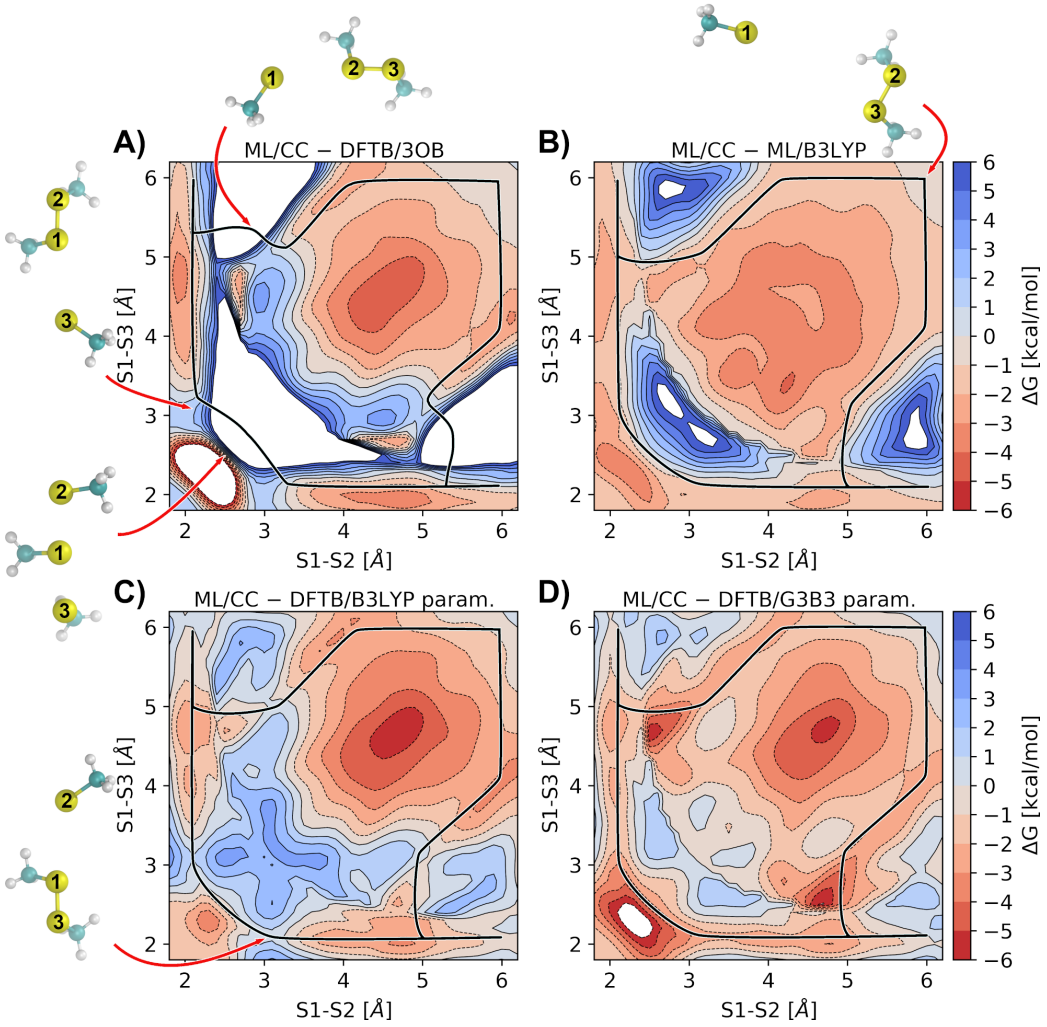

Figure D.1.: PMF of disulfide shuffling between a dimethyl disulfide and a methylthiolate in aqueous solution obtained with Coupled Cluster machine learned correction minus the PMF obtained with – (A) DFTB with 3OB parameters, (B) DFTB with 3OB parameters and a machine learned energy correction based on B3LYP, (C) DFTB with a reparemterized S–S repulsive potential fitted to B3LYP data, (D) DFTB with a reparemterized S–S repulsive potential fitted to G3B3 data. Contour lines are drawn every 1 kcal/mol. Only differences between -6 and +6 kcal/mol were considered. In white areas the differences are larger than |6 kcal/mol|. Exemplary pathways based on the respective free energy profiles in Fig. 7 are drawn as black lines.

Von Willebrand Factor's C4 domain

Table D.1.: Reaction barriers of disulfide shuffling in C4 between S36, S78 and S79. Barriers are calculated w.r.t. the global minimum of the respective bond. Energies are in kcal/mol.

	b1 ⇌ b2	b2 ⇌ b3	b3 ⇌ b1
method	$S79^-$ + S36–S78 ⇌ S79–S36 + $S78^-$	$S78^-$ + S79–S36 ⇌ S78–S79 + $S36^-$	$S36^-$ + S78–S79 ⇌ S36–S78 + $S79^-$
DFTB/3OB	4.5 / 3.5	13.3 / 3.1	2.2 / 9.4
ML/B3LYP	9.1 / 8.6	16.0 / 6.8	9.0 / 13.9
DFTB/B3LYP	8.7 / 8.4	14.8 / 7.8	8.0 / 14.3
ML/CC	9.9 / 9.3	17.8 / 7.4	10.0 / 13.6
DFTB/G3B3	11.9 / 10.3	19.8 / 9.5	10.1 / 17.5

Thiol-disulfide exchange b1⇌b2: The S36–S78 bond (b1) is the wild-type disulfide bond which bridges the two subdomains of the protein, compare Fig. D.2A. S79 can perform a nucleophilic attack on S36 when an almost linear S–S–S angle is formed which results in a disulfide bond between S79–S36 (b2) and a S78 anion. The barrier obtained with the reference method ML/CC is $\Delta G^{\ddagger}_{\mathrm{b1}\to\mathrm{b2}}$ = 9.9 kcal/mol. With ML/B3LYP the barrier is 0.8 kcal/mol smaller, with DFTB/B3LYP it is 1.2 kcal/mol smaller and with DFTB/G3B3 it is 2 kcal/mol higher and thus agree well with the reference, whereas DFTB/3OB underestimates the barrier by 5.4 kcal/mol. Since the newly formed S79–S36 lies higher in energy, 0.3 to 1.6 kcal/mol depending on the used method, the barriers for the backward reaction b2→b1 are consequently slightly smaller. The reference value with ML/CC is $\Delta G^{\ddagger}_{\mathrm{b2}\to\mathrm{b1}}$ = 9.3 kcal/mol and all methods except DFTB/3OB ($\Delta G^{\ddagger}_{\mathrm{b2}\to\mathrm{b1}}$ = 3.5 kcal/mol) agree within ±1 kcal/mol, compare Tab. D.1. Compared to thiol-disulfide exchange between the methylthiolate and dimethyldisulfide in aqueous solution, the reaction barriers in the protein are significantly smaller by a factor of ca. 2.5. Thus, the protein environment is catalytic for the b1⇌b2 reaction caused by electrostatic and steric interactions: (i) there are several hydrophobic residues in direct proximity of the reactive sulfurs, (ii) the sulfurs are less exposed to water, compare Fig. D.5. This means that the flexible loops form a small hydrophobic pocket for the b1⇌b2 exchange which leads to a more delocalized charge of the sulfurs and thus a smaller reaction barrier.

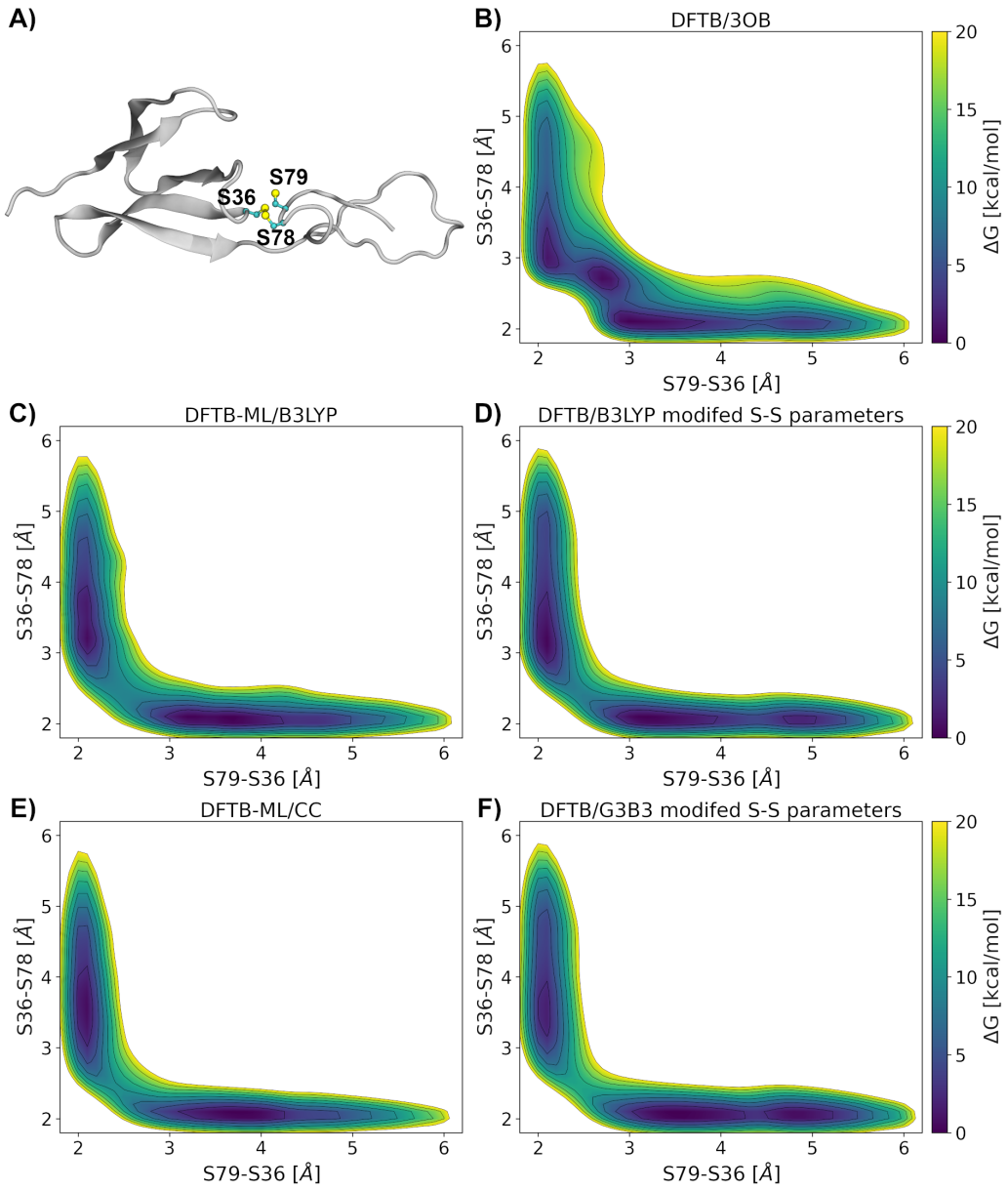

Figure D.2.: (A) Exemplary structure of the Von Willebrand factor C4 domain with a disulfide bond between S36–S78 and S79 in an anionic state. Free energy profiles of the disulfide exchange reaction between S36, S78, and S79 computed with (B) DFTB with 3OB parameters, (C) DFTB with 3OB parameters and a machine learned energy correction based on B3LYP, (D) DFTB with a reparemterized S–S repulsive potential fitted to B3LYP data, (E) DFTB with 3OB parameters and a machine learned energy correction based on CC, (F) DFTB with a reparemterized S–S repulsive potential fitted to G3B3 data. Contour lines are drawn every 2.5 kcal/mol.

Thiol-disulfide exchange b2⇌b3: When the S78 anion approaches the S79–S36 bond (b2) for a nucleophilic attack on S79, a vicinal disulfide between the two sequence adjacent cysteines S78–S79 (b3) is formed which lies ca. 10 kcal/mol higher in energy. Vicinal disulfides are the smallest possible intramolcular disulfide loop and experience a a high dihedral strain due to their small ring size and are therefore unfavorable in terms of energy.[214, 215] The reaction barrier is also rather high with $\Delta G^{\ddagger}_{\mathrm{b2}\rightarrow\mathrm{b3}}$ = 17.8 kcal/mol, obtained with ML/CC. The obtained barrier with ML/B3LYP is 1.8 kcal/mol smaller, with DFTB/B3LYP 3 kcal/mol smaller, with DFTB/3OB 4.5 kcal/moll smaller and with DFTB/G3B3 the barrier is 2 kcal/mol higher. Since b3 lies higher in energy than b2, the barriers for the backward reaction are smaller, $G^{\ddagger}_{\mathrm{b3}\rightarrow\mathrm{b2}}$ = 7.4 kcal/mol with ML/CC. With ML/B3LYP the barrier is 0.6 kcal/mol smaller, with DFTB/B3LYP 0.4 kcal/mol higher, with DFT3/OB 4.3 kcal/mol smaller and with DFTB/G3B3 the barrier lies 2.1 kcal/mol higher. In summary, the S79–S36 bond is thermodynamically favored over the vicinal S78–S79. Again this can be explained by different steric and electrostatic interactions: (i) vicinal bonds experience a high dihedral strain, (ii) the S78 anion might be more stabilized than the S36 anion; in the vicinity of S78 there is a positively charged lysine Lys74, while in vicinity of S36 there is a negatively charged Glu33. Moreover, the sulfurs are more solvent exposed than in the previous reaction which might contribute to a more localized charge on S78, compare RDF in Fig. D.5.

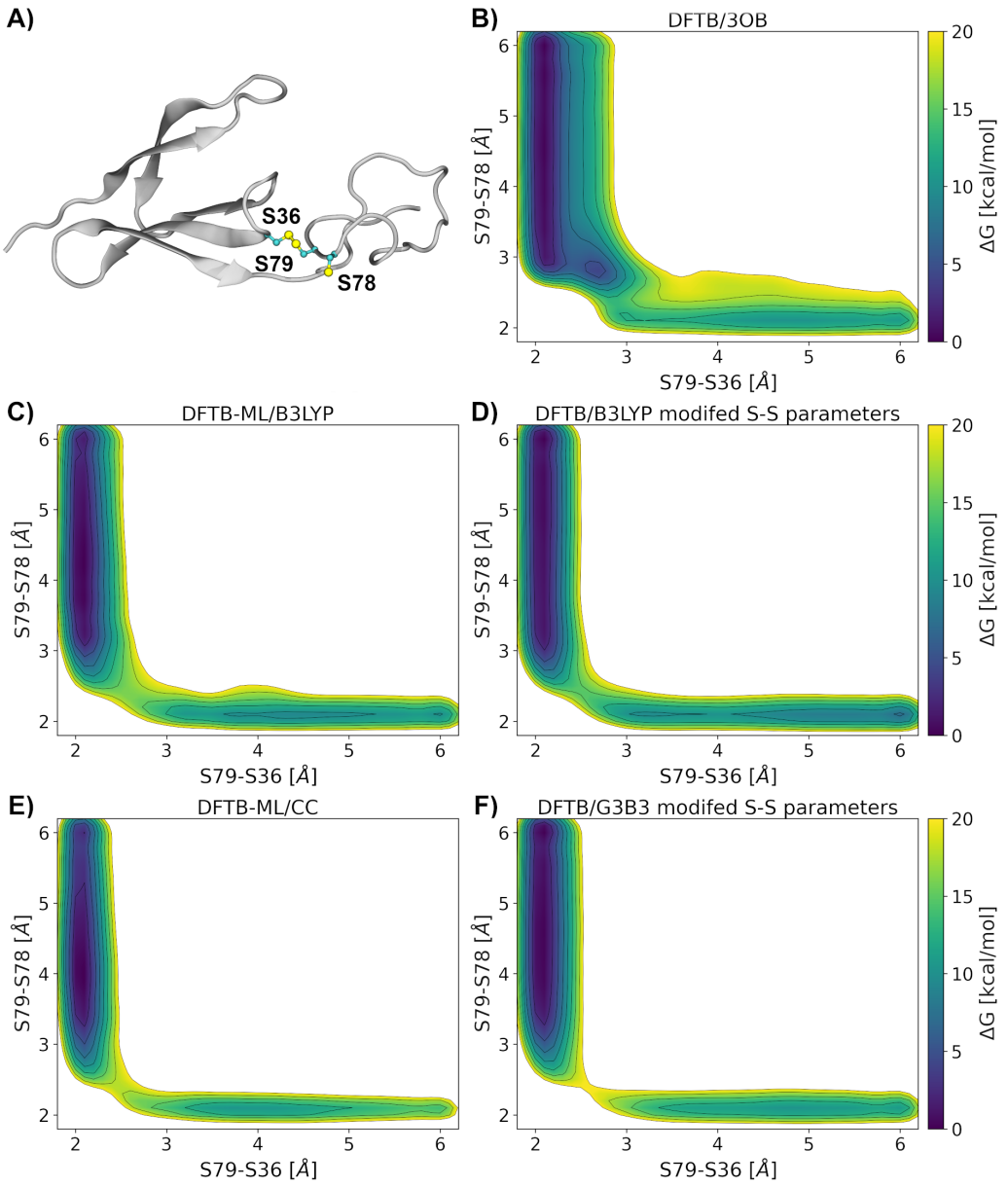

Figure D.3.: (A) Exemplary structure of the Von Willebrand factor C4 domain with a disulfide bond between S36–S79 and S78 in an anionic state. Free energy profiles of the disulfide exchange reaction between S36, S78, and S79 computed with (B) DFTB with 3OB parameters, (C) DFTB with 3OB parameters and a machine learned energy correction based on B3LYP, (D) DFTB with a reparemterized S–S repulsive potential fitted to B3LYP data, (E) DFTB with 3OB parameters and a machine learned energy correction based on CC, (F) DFTB with a reparemterized S–S repulsive potential fitted to G3B3 data. Contour lines are drawn every 2.5 kcal/mol.

Thiol-disulfide exchange b3⇌b1: When the S36 anion performs a nucleophilic attack on S78 the wild-type disulfide bond S36–S78 (b1) is recovered. the vicinal S78–S79 (b3) lies 3.6 to 7.4 kcal/mol higher in energy than S36–S78 (b1). With ML/CC the reaction barrier $G^{\ddagger}_{\mathrm{b3}\to\mathrm{b1}}$ = 10.0 kcal/mol. The barrier obtained with ML/B3LYP is 1 kcal/mol smaller, with DFTB/B3LYP it is 2 kcal/mol smaller, with DFTB/3OB 7.8 kcal/mol smaller and with DFTB/G3B3 the barrier is 0.1 kcal/mol higher. For the backward reaction the barriers are higher, the ML/CC reference yields $G^{\ddagger}_{\mathrm{b1}\to\mathrm{b3}}$ = 13.6 kcal/mol. With ML/B3LYP the backward reaction barrier is 0.7 kcal/mol higher, with DFTB/B3LYP 0.3 kcal/mol higher, with DFTB/3OB 4.2 kcal/mol smaller and with DFTB/G3B3 the barrier is 3.9 kcal/mol higher. Analogously to the b2⇌b3 disulfide exchange, the vicinal S78–S79 bond is less stable than S36–S78 due to a higher dihedral strain energy. However, the reaction energy is smaller and so are the barrier heights. This might be explained in terms of solvent accessibility, compare Fig. D.5. In the b3⇌b1 reaction the sulfurs are less exposed to water than in the b2⇌b3 reaction which could lead to a more delocalized charge and thus smaller barriers.

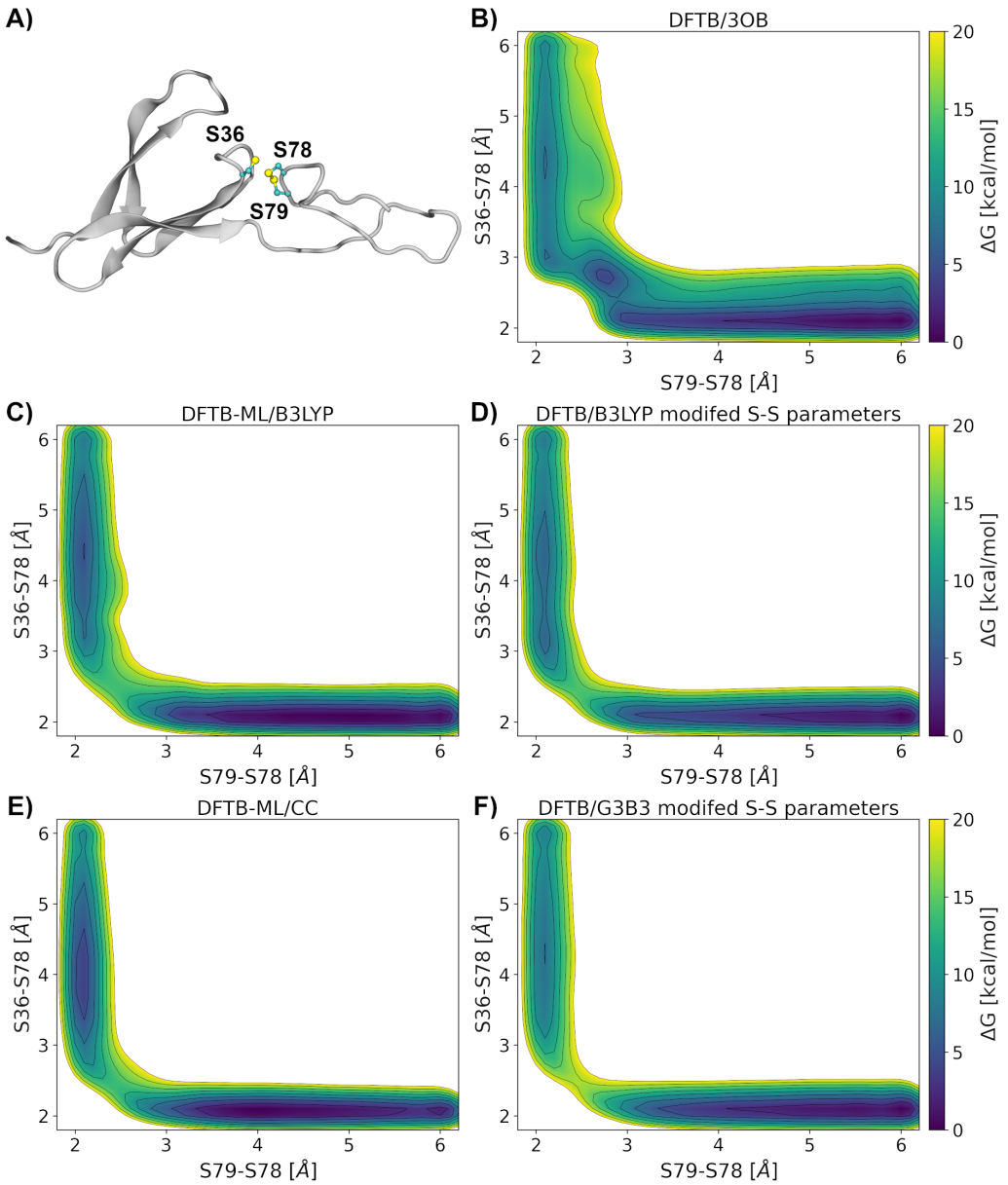

Figure D.4.: (A) Exemplary structure of the Von Willebrand factor C4 domain with a disulfide bond between S78–S79 and S36 in an anionic state. Free energy profiles of the disulfide exchange reaction between S36, S78, and S79 computed with (B) DFTB with 3OB parameters, (C) DFTB with 3OB parameters and a machine learned energy correction based on B3LYP, (D) DFTB with a reparemterized S–S repulsive potential fitted to B3LYP data, (E) DFTB with 3OB parameters and a machine learned energy correction based on CC, (F) DFTB with a reparemterized S–S repulsive potential fitted to G3B3 data. Contour lines are drawn every 2.5 kcal/mol.

Radial distribution function of water. In each setup the radial distribution function of water w.r.t. the central sulfur atoms of each considered reaction was calculated, see Fig. D.5.

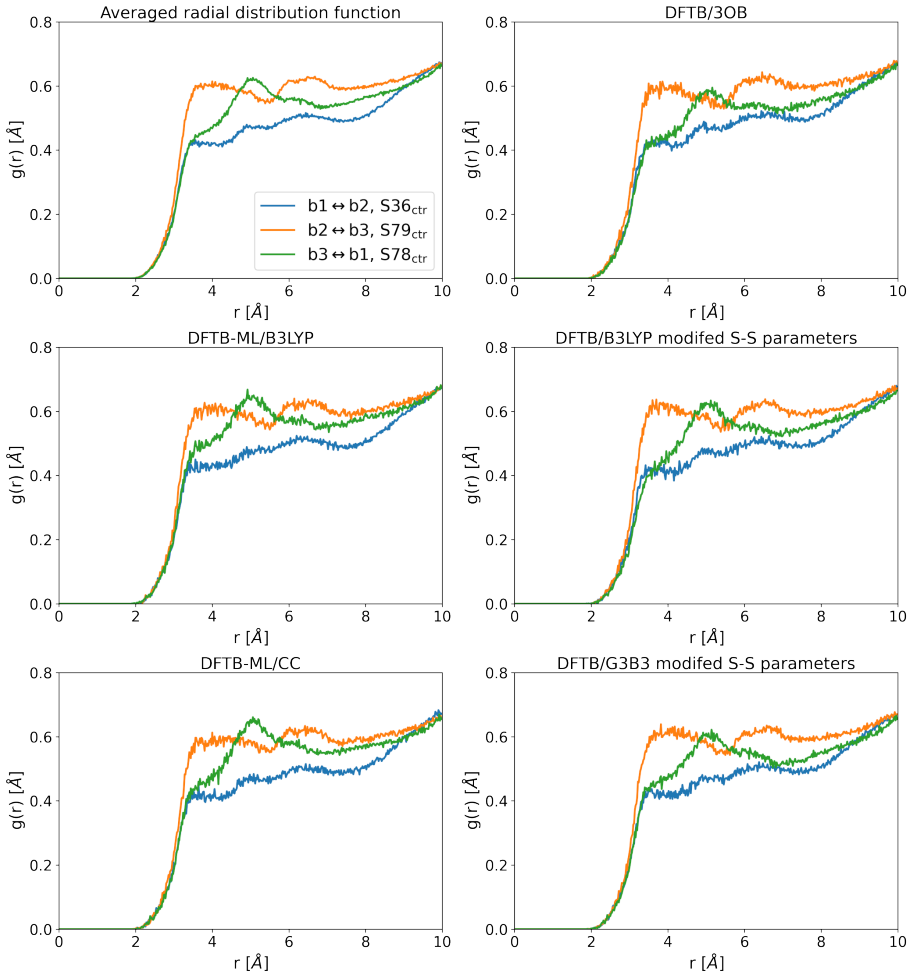

Figure D.5.: RDFs of water w.r.t. the central sulfur S_{ctr}: (A) Averaged over all five setups and the respective considered reaction performed with (B) DFTB with 3OB parameters, (C) DFTB with 3OB parameters and a machine learned energy correction based on B3LYP, (D) DFTB with a reparemterized S–S repulsive potential fitted to B3LYP data, (E) DFTB with 3OB parameters and a machine learned energy correction based on CC, (F) DFTB with a reparemterized S–S repulsive potential fitted to G3B3 data.

E. Force clamp simulations of vWf's C4 domain

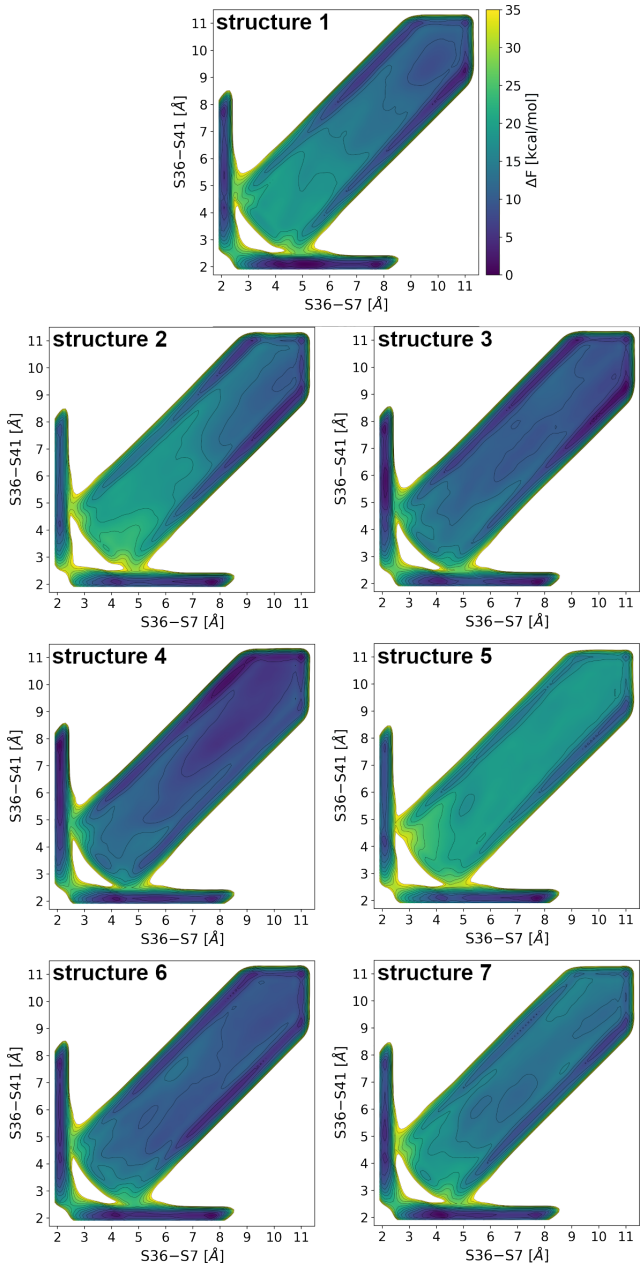

Figure E.1.: PMFs as function of the S36–S7 and S36–S41 distances, S7-S41 distances are integrated out, obtained from metadynamics with different starting structures.

Previous publications and Manuscripts

- **Chapter 7**:
  Maag, D.; Mast, T.; Elstner, M.; Cui, Q.; Kubař, T.
  O to bR transition in bacteriorhodopsin occurs through a proton hole mechanism.
  Proc. Natl. Acad. Sci. USA 2021, 118 (39).

- **Chapter 10**:
  Maag, D.; Putzu, M.; Gómez-Flores, C. L.; Gräter, F.; Elstner, M.; Kubař, T.
  Electrostatic interactions contribute to the control of intramolecular thiol–disulfide isomerization in a protein.
  Phys. Chem. Chem. Phys. 2021.

- **Chapter 11**:
  Gómez-Flores, C. L.[1]; Maag, D.[1]; Kansari, M.; Vuong, V. Q.; Irle, S.; Gräter, F.; Kubař, T.; Elstner, M.
  Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural Network Corrected DFTB/MM Methodology.
  J. Chem. Theory Comput. 2022.

- **Chapter 12**:
  Kutzki, F.; Butera, D.; Lay, A. J.; Maag, D.; Chiu, J.; Wook, H.-G.; Kubař, T.; Elstner, M.; Aponte-Santamaría, C.; Hogg, P. J.; Gräter, F.
  Force-Propelled Reduction and Exchange of Disulfide Bonds in Von Willebrand Factor's C4 Domain.
  manuscript.

[1] Authors contributed equally

Acknowledgement

I would like to express my deepest appreciation to Prof. Marcus Elstner, who gave me the opportunity to work on many interesting projects in his group and welcomed me in the Graduiertenkolleg 2450. He encouraged me to direct my research according to my own interests, was always available for discussions and gave me valuable scientific advice many times.

I am deeply indebted to Dr. Tomáš Kubař for his guidance and all the helpful advice, from which I have benefited greatly. His (virtual) door was always open for discussion and this thesis would not have been possible without him.

I would also like to extend my deepest gratitude to my collaborators Prof. Qiang Cui, Thilo Mast, Marina Putzu, Claudia Leticia Goméz-Flores and Mayukh Kansari for their invaluable contributions to this work.

I am extremely grateful to the Graduiertenkolleg 2450 for the financial support and to all the members of the Graduiertenkolleg 2450 for organizing many interesting hands-on tutorials, seminar days and workshops.

I also wish to show my appreciation to the Heidelberg P4 members of the GRK2450, Prof. Frauke Gräter, Dr. Camilo Aponte-Santamaría and Fabian Kutzki for the fruitful discussions and meetings.

I would like to extend my sincere thanks to all colleagues in the TCB Group for the nice working atmosphere and the wonderful rooftop barbecues.

Finally, I would like to thank my girlfriend Lisa Sattel and my family for their relentless support. They were always there for me and without them I could not have written this thesis.