Collective Dynamics in Allosteric Transitions: A Molecular Dynamics Study

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm IMPRS-pbcs
der Georg-August University School of Science (GAUSS)

vorgelegt von
Martin David Vesper
aus Berlin

Göttingen, 2012

- Betreuungsausschuss[1]

  Prof. Dr. Bert de Groot (Betreuer)
  Computational and Biomolecular Dynamics Group[1]

  Prof. Dr. Marcus Müller
  Institute for Theoretical Physics[2]

  Prof. Dr. Ralf Ficner
  Institute for Structural Biology[2]

- Prüfungskommission

  Prof. Dr. Bert de Groot (Referent)

  Prof. Dr. Ralf Ficner (Koreferent)

  Prof. Dr. Marcus Müller

  Prof. Dr. Helmut Grubmüller
  Department of Theoretical and Computational Biophysics[1]

  Prof. Dr. Marina Bennati
  Electron Spin Resonance Spectroscopy Group[1]

  Dr. Adam Lange
  Solid-State NMR Spectroscopy Group[1]

[1]Max Planck Institute for Biophysical Chemistry
[2]Georg-August University Göttingen

Tag der mündlichen Prüfung: 18.12.2012

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

Göttingen, den 12.11.2012

Martin Vesper

"Eine weitere experimentelle Verfolgung dieses Gegenstandes blieb für Physiologie und Pharmakodynamik wünschenswerth und für das Studium der Natur der Blutmolekeln auch erspriesslich."
Hünefeld, 1840 [1]

Contents

1 Introduction

In all living cells, proteins are performing a vast amount of functions. These functions are often controlled by a special mechanism: allosteric regulation. This thesis is focusing on the dynamics of proteins that are underlying the allosteric regulation.

In this chapter, I will give a short introduction on proteins in general and how collective motions are related to them, and on allosteric interactions in proteins[1]. It is a common hypothesis that these collective motions are underlying the allosteric regulation.

1.1 Proteins, Conformations & Collective Dynamics

From Amino Acids to Protein Structures In all cells, proteins perform a large variety of tasks, but are composed of a small set of building blocks only: 21[2] amino acids. In a protein, amino acids are connected by peptide bonds forming a chain, usually containing more than 100 amino acids. The main line is called the backbone or main chain. Every amino acid has a side chain that is characteristic for it. Side chains vary in shape, charge, and atom composition. For example, while arginine has a long side chain with a positive charge at the end, aspartate's negatively charged side chain is only half as long. And while proline's side chain is restrained by a ring structure, making it an element that locally reduces main chain mobility, glycine has no side chain at all, allowing more flexibility for the neighbouring side chains. An example of amino acids in a protein structure is illustrated on the left in Fig. 1.1.

Under physiological conditions, most proteins fold into certain configurations, which are energetically favoured by interactions between main and side chain atoms. When folded, proteins are found in specific states (see e.g. Fig. 1.1, middle), where most of them perform their function.

The structure of a protein can be described on different levels, from the primary to the quaternary structure, each level containing the information of the lower levels. The primary structure of each protein is the sequence of amino acids. Hydrogen bonds between backbone atoms give the protein fundamental structural features like α-helices and β-sheets: the secondary structure. Side chain interactions fold this into the three-dimensional tertiary structure. If two or more protein chains arrange in a protein complex, the form what is called the quaternary structure.

[1] A detailed description of these topics can be found in protein textbooks, like *Proteins: Structures and Molecular Properties* [2].

[2] in eukaryotes

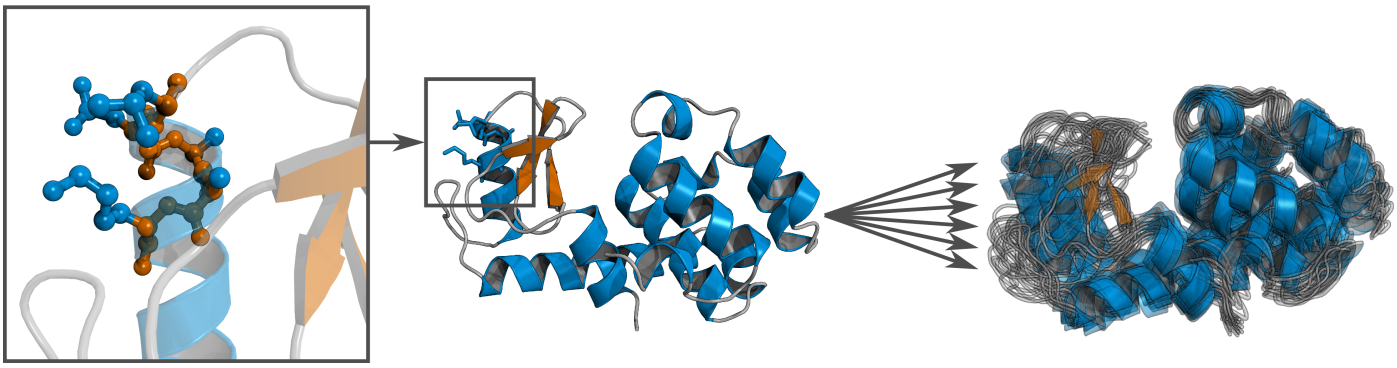

Figure 1.1: **Proteins: From Amino Acids to Structure Ensembles.** Schematic representation of different levels of protein hierarchy on the example of T4 lysozyme (PDB id: 2LZM [3]). Left: Close up on an α-helical LNAAK motif with main chain shown in orange and side chains in blue. Middle: Location of the motif within the whole protein, which is shown in cartoon representation, α-helices in blue, β-sheets in orange and loops in grey. Right: An ensemble of structures representing the flexibility of the system (ensemble generated with CONCOORD [4]).

From Protein Structures to Structure Ensembles From a more physical point of view, proteins are many-particle systems with a large number of restraints. Covalent bonds, sterical interactions, hydrogen bonds, etc. are limiting the configurational space accessible to a protein. Nevertheless, proteins are still flexible and can adopt different conformations while fulfilling the restraints.

When performing their function, proteins adopt different conformations. For example in case of ligand binding (see next section), different conformations of binding pockets can have different accessibilities for ligands to bind and, once bound, they offer different chemical environments for the ligands. Hence, specific conformations are crucial for specific tasks, and the understanding of conformational changes in proteins is necessary to understand biochemical processes on an atomic level.

Figure 1.1 summarizes a simplified view of how a protein ensemble inherits its dynamic properties from the amino acid sequence.

Imagining each state as a rigid structure is a crude simplification. In reality, each state consists of a continuous ensemble of structures with different probabilities. Important changes in function may be happening when these weights shift within the ensemble.

Mathematically, the pathway connecting two conformations can be very complicated when looking at it in Cartesian coordinates. With the right coordinate transformation, these conformational changes can be described in an easier way by suitable collective coordinates. In Chap. 2 methods will be explained that focus on collective dynamics.

1.2 Allostery

Proteins can bind various substances ranging from small molecules – so-called ligands – to other proteins. Binding is essential for a vast number of molecular processes in our body. Once the substances are bound, proteins can fulfil various tasks: The serine protease trypsin, for example, binds proteins and cleaves the peptide bonds next to lysine and arginine residues [5]. Myoglobin, on the other hand, transports and stores oxygen inside muscle cells [6]. These special functions require specialized interactions of the proteins.

Since "Power is nothing without control" is true also for ligand binding, multiple ways of altering binding behaviour have evolved. Molecules that affect the binding in one way or another are called *effectors*. While *activators* increase the binding affinity or catalysis rate, *inhibitors* do the opposite.

If an effector binds in the same site as the ligand it is affecting, the regulation is called orthosteric[3]. The effector can directly act on the ligand, or it can directly manipulate the binding site. If now an effector binds at a site *distant* from the site whose binding affinity it is changing, the regulation is called *allosteric*[4]. The effector is called an allosteric effector and the whole phenomenon is referred to as *allostery*.

To better understand the characteristics of allostery some explanations on non-allosteric systems will follow.

1.2.1 Non-Allosteric Systems

For a protein P and a ligand L in solution there exists and equilibrium between the complex $P \cdot L$ and the not complexed species P. The number of proteins in complex with the ligand depends on the binding affinity of L to P and the concentrations of both. It is described by the association constant K_a:

$$P + L \overset{K_a}{\rightleftharpoons} P \cdot L \tag{1.1}$$

and calculated by measuring the respective protein, ligand and complex concentrations:

$$K_a = \frac{[P \cdot L]}{[P][L]}, \tag{1.2}$$

For a fixed protein concentration one would assume a higher fraction of complexes for a higher concentration of ligands and saturation at sufficiently high $[L]$. Both is true, as can be seen from the fraction of complexed proteins:

$$\frac{[P \cdot L]}{[P \cdot L] + [P]} = \frac{K_a[L]}{1 + K_a[L]} = 1 - \frac{1}{1 + K_a[L]}. \tag{1.3}$$

[3] from Greek "at the right place"
[4] from Greek "at another place"

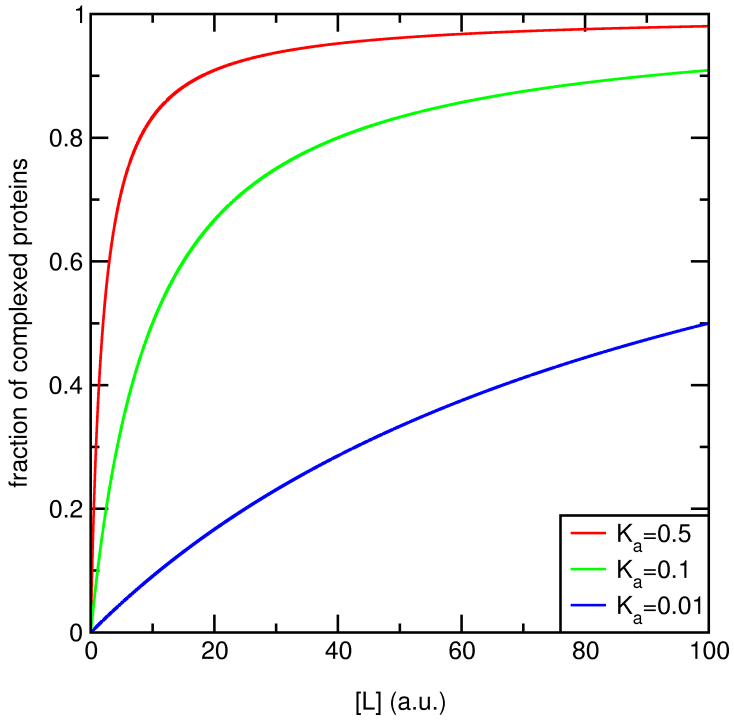

Figure 1.2: **Model Association Curves** according to equation 1.3 for different constants K_a.

This is one branch of a hyperbola with the asymptote of 1. K_a can easily be extracted from the curve since the fraction of ligated proteins is one half for $[L] = 1/K_a$ (see Fig. 1.2). A higher binding affinity means that a lower ligand concentration suffices to bind the same amount of ligands to a certain protein.

In terms of energetics, the free energy difference of the uncomplexed and the complexed state is calculated by the difference in phase space volume of both states (or the probability to find the system in one of the states respectively):

$$\Delta G = -RT \ln \frac{[P \cdot L]}{[P][L]} + RT \ln \frac{[P]}{[P][L]} = -RT \ln \frac{[P \cdot L]}{[P]} = -RT \ln(K_a[L]) \tag{1.4}$$

1.2.2 Allosteric Systems

Allosteric regulation or allostery describes an interaction at a binding site interfering with a distant binding site. An example is the case of an enzyme, in which the binding affinity for a substrate in the substrate binding site is affected by the binding of an effector molecule in the allosteric site. If the effector binding site would coincide with the substrate binding site the regulation would be *in situ*. An allosteric effector could

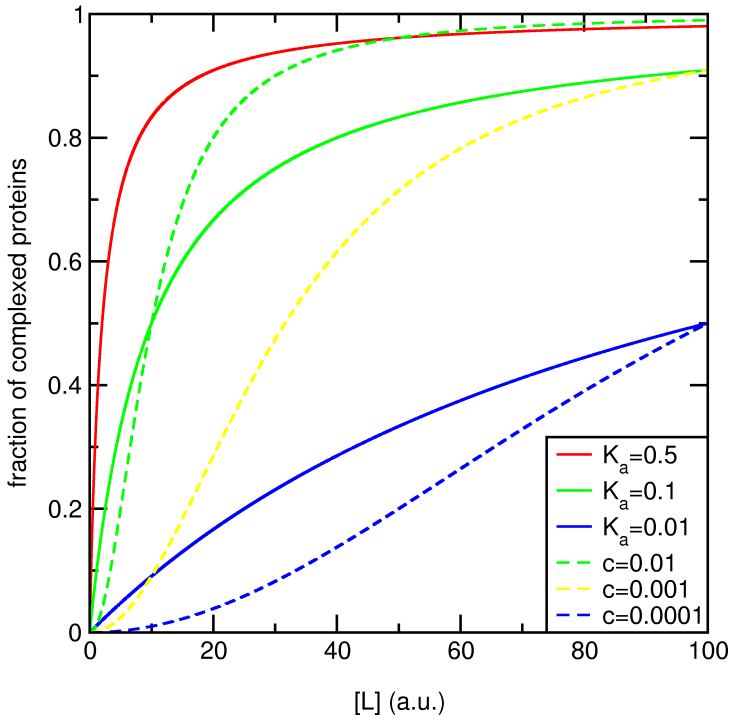

Figure 1.3: **Comparison Between Positive Non-Allosteric And Allosteric Association Curves:** Non-allosteric (solid lines) and allosteric (dashed lines) model binding association curves are shown. For the allosteric case, the functional property curves is given in [2].

increase the substrate binding affinity (allosteric activator) or decrease it (allosteric inhibitor).

Between the binding sites information must be transferred. The intriguing part about allosteric interactions is that there is no general answer to how this information flow is manifested in different systems. In Figure 1.4 a typical allosteric regulation is sketched.

Example An example of how complex and manifold allosteric interactions can be, is the γ-aminobutyric acid receptor A ($GABA_A$). It is a membrane channel in neurons that, upon opening, conducts chloride ions leading to a hyperpolarization of the neuron. The central conducting pore is opened by binding of the main agonist γ-aminobutyric acid (GABA). In addition to the orthosteric GABA binding site, $GABA_A$ has a number of allosteric binding sites. The positive allosteric effectors include the widely investigated binding sites for benzodiazepines and ethanol [7]. Further, the channel can be regulated with additional positive or negative allosteric effectors [8].

Compared to a simple open/closed mechanism, the allosteric regulation allows for a powerful level of control that cannot be realized with orthosteric binding alone.

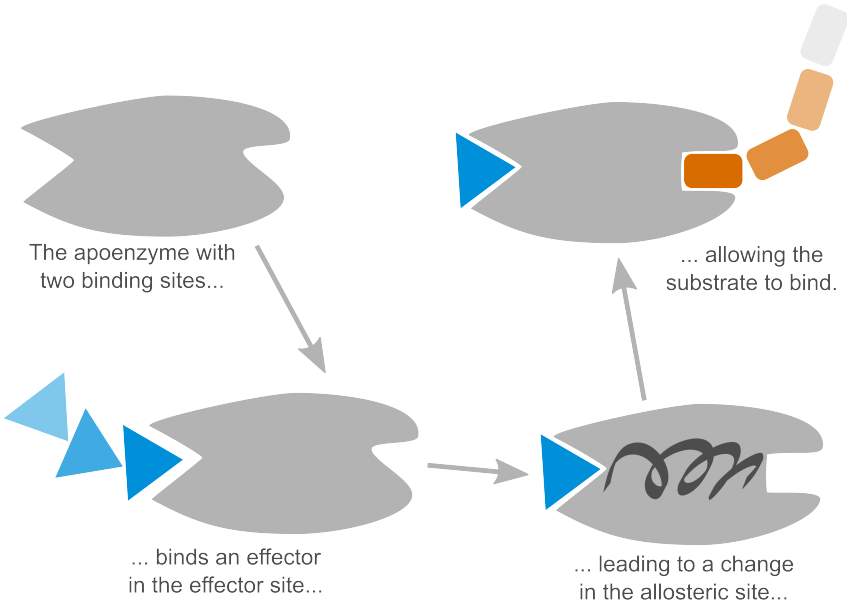

Figure 1.4: **Schematic Representation of a Typical Allosteric Interaction:** An enzyme (grey) with binding sites on left and right binds an effector molecule (blue) which changes the binding affinity at the distant binding site, facilitating the binding of a substrate (orange).

1.2.3 Cooperativity

If a ligand can bind to multiple sites of a protein and these sites affect each other allosterically, the binding process is called cooperative. In positive cooperativity each binding site occupied by the ligand increases the binding affinity in the remaining binding sites. In negative cooperativity the first ligand bound has the highest binding affinity, which decreases with every ligand bound.

In comparison with the un-cooperative binding discussed at the beginning of this chapter, cooperativity changes the shape of the association curve from a hyperbola to a sigmoidal curve (see examples in Fig. 1.3).

At low concentrations the curve looks like un-cooperative binding with a lower binding constant. For higher ligand concentrations the already bound ligands increase the binding affinity for the remaining sites, resulting in a curve resembling a higher binding affinity curve for high concentrations.

In the case of negative cooperativity, bound ligands hinder further binding events. This results in a steep affinity curve at low concentrations and the transition to a low binding affinity at high concentrations.

This smooth switch between low and high affinity allows for a level of regulations not possible in un-cooperative binding. E.g. positive cooperativity allows the effective transport of ligands as the carrying

protein is either empty or fully loaded, with a low population of partially loaded states.

1.3 Outline of This Thesis

The present thesis aims at a fundamental understanding of allostery and the underlying collective dynamics.

In the first project, we investigated the allosteric regulation leading to the cooperative binding of molecular oxygen to hemoglobin's four protein chains. Hemoglobin's structures from X-ray crystallography revealed two dominant states: the low binding affinity T-state (e.g. [9]) and the high binding affinity R-state (e.g. [10]). From experiments it is known that without the conformational dynamics between the states, hemoglobin lacks the cooperativity [11, 12]. This led to the notion that this collective motion is responsible for the allosteric communication of the four subunits.

In chapter 3, we are addressing the question how collective motions within the chains couple to the global collective T-to-R transition and thereby communicate with each other. Therefore, we analyse Molecular Dynamics (see Chap. 2) simulation trajectories that show a transition between T and R. Decomposing the motions into local and global parts allows us to identify collective motions responsible for the coupling and their underlying molecular interactions.

In the second project, we focused on allosteric interactions within the protein ABCE1. ABCE1 is a member of the large ABC proteins family that is associated with multi drug resistance in cancer cells, hindering chemotherapy [13, 14]. In contrast to most ABC proteins, ABCE1 lacks a transmembrane domain and therefore is not a transporter protein. ABCE1 contains two nucleotide binding domains (NBD) that are common to all ABC proteins. Both NBDs can catalyse ATP to ADP and this hydrolysis is associated with a conformational change from an "open" to a "closed" state. Despite their structural similarity, the NBDs are asymmetric in function [15]: While a characteristic mutation in one NBD decreases the overall affinity, the symmetric mutation in the other site increases the affinity.

In chapter 4, we take first steps in solving this riddle: While a structure from the open state is available [15], a closed structure is still missing. We aim at characterizing the collective motion from common structural features of ABC proteins, allowing us to drive the open structure towards a closed state. From this were able to suggest suitable mutations that are predicted to stabilize the closed conformation. Structural information of the closed state is crucial for future understanding of the allosteric interaction between the two ATP binding sites.

2 Theory and Methods

In this chapter I will give a short introduction into Molecular Dynamics (MD) simulations, the central method that was applied in this thesis to obtain atomistic trajectories for Hemoglobin (see Chap. 3) and ABCE1 (see Chap. 4). More information regarding MD can be found in many books and reviews [16–18]. Further, I will explain the methods with which the multidimensional simulation data were analysed.

2.1 Molecular Dynamics Simulations

Molecular Dynamics (MD) simulation is a method in the field of computational biophysics that provides insight into many biochemical processes on an atomistic level. A typical atomistic model of a protein can be obtained by X-ray crystallography. Although this static picture provides very valuable information, in many cases a dynamic picture is required, be it for configurational entropy estimates or for insight into conformational changes of the protein. In the case of proteins, MD can be successfully applied for example to

- calculate thermostability of different amino acid mutations [19],
- understand conduction and selectivity of potassium channels [20,21],
- study pathogenic peptide aggregation [22]
- understand the molecular basis for atomic force microscopy experiments [23],
- shed light on the protein folding process [24],
- estimate rates for tRNA translocation in the ribosome [25].
- identify the collective mode responsible for molecular recognition in ubiquitin [26,27]

In MD simulations each atom of a system (e.g. a protein solvated by water) is represented by its position and momentum. The interactions between the atoms are defined in a force field consisting of terms for the bonded and non-bonded interactions. By solving Newton's equations of motion, the atoms move in the potentials given by the force field. Thus, the system evolves over time. With evolving computer power nowadays systems with more than a million atoms can be simulated. When simulating small systems, time scales of up to 1 ms can be reached with specialized hardware [28].

2.1.1 Basic Approximations of MD

MD is a classic description of systems in regimes of time- and length scales where quantum-mechanical effects play a role. The following approximations are the basis of MD:

1. Born-Oppenheimer approximation: Nuclei and electrons can be treated separately.
2. Classic nuclei motions: The motions of nuclei are described by Newton's equations of motion.
3. Application of force fields: The interactions can be defined by classical force fields.

Born-Oppenheimer Approximation The time evolution of a quantum mechanical system is given by the time-dependent Schrödinger equation:

$$\mathcal{H}\psi = i\hbar\frac{\partial}{\partial t}\psi, \tag{2.1}$$

where $\mathcal{H}$ denotes the Hamiltonian of the system and ψ the wave-function. The latter depends on the position of nuclei and electrons. In the Born-Oppenheimer approximation this dependence can be decoupled for both [29]. The assumption behind this is that the heavy nuclei move much slower than the light electrons. Hence, at every moment the electrons *feel* a static potential from the nuclei. In Eq. 2.1 ψ is described by the product of the decoupled wave functions for the nuclei and electrons

$$\psi(\mathbf{R}_{nuc}, \mathbf{r}_{el}) = \psi_{nuc}(\mathbf{R}_{nuc}) \cdot \psi_{el}(\mathbf{R}_{nuc}; \mathbf{r}_{el}).$$

Hereby $\mathbf{R}_{nuc}$ and $\mathbf{r}_{el}$ denote the vectors of nuclei and electron positions respectively.

Classic Nuclei Motions Based on the Born-Oppenheimer approximation the nuclei motions in MD are treated classically by Newtonian dynamics. This assumption follows Ehrenfest's theorem that states that (under certain conditions) the time evolution of expectation values of operators can be described classically. This is true especially for the position operator. Together with the assumption $\langle F(x)\rangle \approx F(\langle x\rangle)$ this leads to the Newtonian equation of motion:

$$m\frac{d^2}{dt^2}\langle x\rangle = F(\langle x\rangle).$$

In other terms, the nuclei are moving in a potential V according to

$$m_i\frac{d^2}{dt^2}R^i_{nuc} = -\nabla_{R^i_{nuc}} V(\mathbf{R}_{nuc}). \tag{2.2}$$

Application of Force Fields In classical MD interactions between atoms are defined by classical functions. For example, the interaction of two atoms due to a covalent bond is described by a harmonic potential $V = k/2(l - l^{(0)})^2$. The force constant k and the equilibrium bond length $l^{(0)}$ depend on the atom types and the type of bond between them. Potential terms for other bonded interactions like angles, dihedrals and improper dihedrals are defined in a similar way. Non-bonded interactions including Van der Waals and Coulomb interactions are calculated for all atom pairs close enough to each other. That said, no chemistry can occur in normal MD, since the bonded interaction partners are constant, and bonds cannot form or break.

A typical MD potential for Eq. 2.2 as used in the GROMACS software package [30, 31] looks like this:

$$
\begin{aligned}
V = & \sum_{\text{bonds } i} \frac{k_i}{2} \left(l_i - l_i^{(0)} \right)^2 && \text{(bonds, bonded)} \\
& + \sum_{\text{angles } i} \frac{f_i}{2} \left(\rho_i - \rho_i^{(0)} \right)^2 && \text{(angles, bonded)} \\
& + \sum_{\text{dihed. } i} \frac{d_i}{2} \left(1 + \cos \left(n\phi_i - \phi_i^{(0)} \right) \right) && \text{(diheadrals, bonded)} \\
& + \sum_{\text{imp. dih. } i} \frac{m_i}{2} \left(\xi_i - \xi_i^{(0)} \right)^2 && \text{(improper diheadrals, bonded)} \\
& + \sum_{\text{atoms } i,j} 4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) && \text{(Van der Waals, non-bonded)} \\
& + \sum_{\text{atoms } i,j} \frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{q_i q_j}{r_{ij}} && \text{(Coulomb, non-bonded)}
\end{aligned}
$$

The parameters used to calculate the potential are defined in so-called force fields. For MD simulations of proteins several force fields are established. They differ in the derivation of the parameters but mostly derive force constants from quantum-mechanical calculations and from experimental measurements like vibrational bond-spectra or melting points of solvents. Some of the most commonly used force field for bio-molecular MD are AMBER [32], GROMOS [33, 34], OPLS [35, 36], CHARMM [37, 38], of which the GROMOS 43a2 and AMBER 99sb-ildn were used for projects in this thesis.

2.1.2 Time Integration

With a given set of coordinates for all atoms and assigned velocities, Newton's equations are integrated stepwise. The integrator used is known as the *leap-frog* algorithm, named after the shifted calculation of positions

and velocities:

$$\mathbf{v}(t + \Delta t/2) = \mathbf{v}(t - \Delta t/2) + \mathbf{F}(t)\Delta t/m$$
$$\mathbf{r}(t) = \mathbf{r}(t - \Delta t) + \mathbf{v}(t + \Delta t/2)\Delta t$$

For numerical stability the time step Δt has to be chosen so that it is sufficiently shorter than the fastest motion of the system, as vibrations of bonds between carbon and hydrogen atoms. Thus, commonly a time step of $\Delta t = 2\,fs$[1] is chosen for those force fields. By freezing the fastest degrees of freedom a larger time step can be chosen. This is the underlying idea of the so-called virtual sites [41], where amongst others the angles between hydrogen atoms in CH_3-groups are kept.

2.1.3 Temperature, Pressure and Periodic Images

To reproduce the conditions in cells, MD simulations of biological systems are usually simulated as a canonical ensemble, i.e. with constant temperature and pressure. Numerical errors of the time integration and force calculation may lead to slow drifts in the temperature. To reduce this effect, algorithms have been developed, called thermostats [42, 43]. Likewise, barostats are used to keep pressure constant [44, 45].

When simulating small systems, the boundaries of the simulation box can be critical and finite size effects have to be taken into account. To this end, the system can be simulated including the interaction with its periodic images. This *de facto* renders an infinite system, for which boundaries do not play a role. This allows for an effective calculation of the long-range electrostatic interaction: the particle-mesh Ewald summation [46,47]. The basic idea is to split up the interactions into a short range part, which can be calculated normally, and a long range part, for which the periodicity facilitates the calculation in the Fourier transformed reciprocal space.

2.1.4 Data Structure

In order to explore atomic coordinates of an MD simulation or properties related to the coordinates, a suitable frame for describing the coordinates must be chosen. A given atomic structure of a system containing N atoms can be thought of as N points in a 3-dimensional space. Alternatively, each structure can be described as one point in a $3N$-dimensional space,

[1] This already implies that water molecules are constrained with the Settle algorithm [39], and other bond lengths are constrained with LINCS [40].

known as the configuration space. A structure then can be rewritten as

$$\mathbf{x} = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{pmatrix} \rightarrow \mathbf{x} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ \vdots \\ y_N \\ z_N \end{pmatrix}$$

Whereas the first alternative is more intuitive as it describes the atomic structure in space atom by atom, the second one is favoured in mathematical handling of trajectories and descriptions of collective motions; it will be used below. In this framework trajectories are paths in the configuration space. A trajectory of M snapshots can be written in the following $3N \times M$ matrix form

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}(t_1) & \mathbf{x}(t_2) & \dots & \mathbf{x}(t_m) \end{pmatrix}.$$

2.2 Analysis Methods for Multidimensional Data

Dealing with multidimensional data such as trajectories from Molecular Dynamics simulations often requires the application of special techniques for either finding a lower dimensional representation of the data set or extracting collective coordinates corresponding to certain observables. In this section several of these techniques are described where Principal Component Analysis belongs to the first category and Partial Least Squares to the second. The methods will be discussed in reference to the analysis of MD data, but can be applied more generally to tasks in multivariate analysis and data mining [48].

2.2.1 Principal Component Analysis

A common technique to find a lower dimensional representation of a data set is Principal Component Analysis (PCA). The basic idea is to find a coordinate transformation that describes the majority of structural fluctuations with just a small number of new, collective coordinates.

This is achieved with PCA by expressing the fluctuations in terms of covariance and the coordinate transformation with a matrix diagonalization.

The diagonalization of the covariance matrix yields a new set of orthonormal vectors given by the eigenvectors. The corresponding eigenvalues describe how much the motion along each eigenvector fluctuates. If the set of eigenvectors is ordered by decreasing eigenvalues, the first eigenvectors (the *principal components*) describe the majority of covariance and thereby structural fluctuations of the system.

Originally, PCA was developed by Pearson to least-square fit planes to multi-dimensional point clouds [49]. In 1933 Hotelling introduced PCA for analysing correlations within multi-dimensional data [50]. As an example application one case of Hotelling's original work will be discussed shortly.

> For a number of school children a test measured the ability and speed in reading and the ability and speed in calculus. With this each of the k individuals was scored in 4 items. Hence, the raw data consisted of k variables in $n = 4$ dimensions and can be written in as matrix elements x_p^q with $p \in \{1, 2, 3, 4\}$ and $q \in \{1, .., k\}$.
>
> The question at hand was if there were correlations within the data that permit to describe most of the measured data in a lower dimensional ($n < 4$) representation, and what these relations looked like. To answer the question, the covariance matrix was calculated which describes how strongly the different dimensions vary simultaneously. If the variables are centered, the covariance of two coordinates i and j reads
>
> $$C_{ij} = \left\langle x_i^k \cdot \left(x_j^k\right)^{\mathrm{T}} \right\rangle_k$$
>
> Diagonalizing $\mathbf{C}$ yields four eigenvectors (of collective features) and the corresponding eigenvalues indicating the contribution of that eigenvector to the total covariance.
>
> Since $\mathbf{C}$ is symmetric, it can be written:
>
> $$\mathbf{C} = \mathbf{Y}\mathbf{\Lambda}\mathbf{Y}^{\mathrm{T}} \text{ with } \mathbf{Y} = (\mathbf{y_1}\ \mathbf{y_2}\ \mathbf{y_3}\ \mathbf{y_4}),\ \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$$
>
> With the $\mathbf{y_i}$ written in the Cartesian space of the raw data, each component marks the contribution of that specific feature to the total direction y_i. In Hotelling's example the eigenvector with the highest eigenvalue was mainly composed of the general ability to read and calculate, showing a positive correlation between the two item scores. Along the second eigenvector reading and calculus showed a negative correlation. The PCA revealed that the test results could be described in two dimensions instead of four: The first one measuring the combined ability in calculus and reading, and the second separating between the children's better subject.

Mathematics of PCA Mathematically speaking, PCA is a linear transformation of the Cartesian coordinate system to a coordinate system with variances maximized along the coordinate axes (see Fig. 2.1).

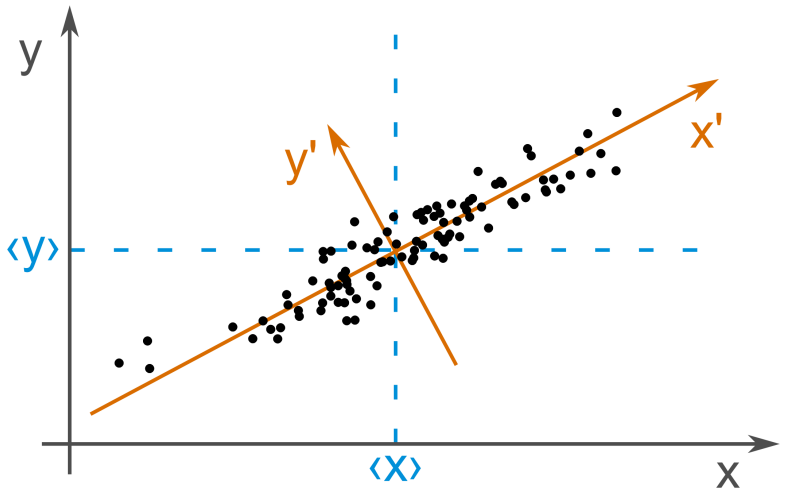

Figure 2.1: **Schematic Representation of PCA:** The two-dimensional data in the original coordinates ($x - y$, grey) show a linear dependence. The PCA shifts the center of the coordinate system to the average (blue) and rotates the axes to maximize variances along the first one ($x' - y'$, orange).

For a given trajectory $\mathbf{x}(t)$[2] the covariance matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \left\langle \left(\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle_t\right) \cdot \left(\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle_t\right)^T \right\rangle_t .$$

Diagonalization yields

$$\mathbf{\Lambda} = \mathbf{Y}^{\mathrm{T}} \mathbf{C} \mathbf{Y},$$

with the diagonal matrix of eigenvalues $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_i, \ldots, \lambda_{3N})$ and the matrix of corresponding eigenvectors $\mathbf{Y} = (\mathbf{y}_1 \ \ldots \ \mathbf{y}_i \ \ldots \ \mathbf{y}_{3N})$. Usually the matrices are ordered with decreasing λ_i. Since the covariance matrix is positive semi-definite[3], all $\lambda_i \geq 0$. Each λ_i describes the variance along the corresponding eigenvector.

A reduction in dimensionality can be obtained by projecting each structure $\mathbf{x}(t)$ onto a smaller subspace $\mathbf{x}(t) \mapsto \mathbf{z}(t)$:

$$z_i(t) = \mathbf{y}_i \cdot \left(\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle_t\right), \text{ with } i \in \{1, \ldots, M\}, \ M < 3N \tag{2.3}$$

[2] In MD simulations, the parameter t will typically be the time, but any ensemble index is possible.

[3] Positive semi-definite means that $\mathbf{v}^T \mathbf{C} \mathbf{v} \geq 0$ holds for all non-zero $\mathbf{v}$. This is true because

$$\begin{aligned} \mathbf{v}^T \mathbf{C} \mathbf{v} &= \mathbf{v}^T \langle (\mathbf{x} - \langle \mathbf{x} \rangle) (\mathbf{x} - \langle \mathbf{x} \rangle)^T \rangle \mathbf{v} \\ &= \langle \left(\mathbf{v}^T (\mathbf{x} - \langle \mathbf{x} \rangle)\right) \left(\mathbf{v}^T (\mathbf{x} - \langle \mathbf{x} \rangle)\right)^T \rangle, \text{ with time-independent } \mathbf{v} \\ &= \langle s^2 \rangle \geq 0, \text{ with } s = \mathbf{v}^T (\mathbf{x} - \langle \mathbf{x} \rangle) . \end{aligned}$$

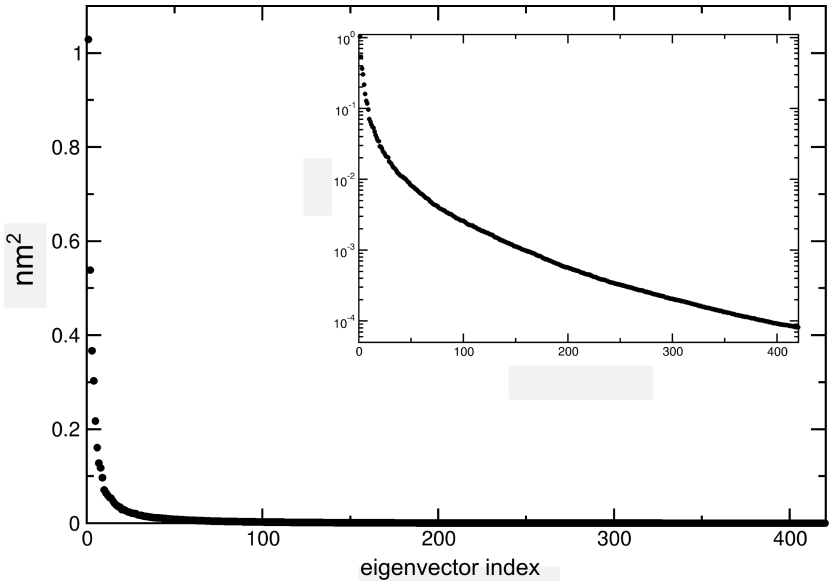

Figure 2.2: **Typical PCA Eigenvalue Distribution:** In this example of a T4 lysozyme MD simulation a fast decrease of the eigenvalues can be seen. The logarithmic plot shows that the eigenvalues decrease faster than exponential.

Application in MD A PCA on a given MD trajectory can yield insight into the collective dynamics of the system, revealing the dominant global motions like e.g. a conformational change upon ligand binding or the motion of two domains connected by a hinge region.

In MD simulations the system, e.g. a protein, can diffuse through the solvent. This motion often contributes the most to the coordinate changes, but is not of an interest if focusing on internal dynamics. To not detect these degrees of freedom each structure of the simulation can be structurally aligned to a reference structure prior to calculating the covariance matrix. That way, the six global degrees of freedom are removed from the system. Nevertheless, this fitting procedure can be ambiguous in flexible systems and thus produce artifacts [51, 52].

The large number of constraints including bonds, angles and sterical restrictions greatly reduces the degrees of freedom actually available for a protein. Usually, in MD simulations the PCA eigenvalues ordered by decreasing variances decay fast. In Fig. 2.2 the PCA eigenvalues for a T4 lysozyme MD simulation are shown. Like in this example, in protein dynamics the first few eigenvectors of a PCA describe anharmonic large-scale motions and together form the so called *essential subspace* [53].

The new coordinates given by the PCA eigenvectors are often called *collective* in the sense that in general all atoms contribute to each individual eigenvector. When encountering PCA on MD data for the first time, it may at seem unusual that, even though the motion along any eigenvector usually involves all atoms, it is still a one-dimensional motion. Thereby,

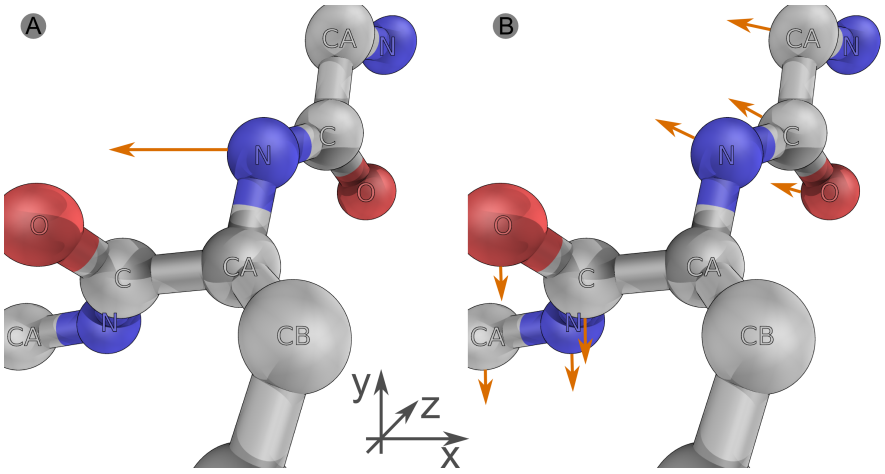

Figure 2.3: **Comparison of Cartesian and Collective Coordinates:** (A) a motion along a Cartesian coordinate: x-coordinate of a specific nitrogen atom; (B) a motion along a collective coordinate: rotation along the CA-CB bond involving all atoms.

affecting this motion only affects one degree of freedom of the system. One could argue that the PCA coordinates are a linear approach to a more natural coordinate system: In Fig. 2.3 a schematic comparison between a coordinate in the Cartesian space (A) and a coordinate in a collective space (B) is shown. The system will not move far along the Cartesian coordinate since this would involve breaking of bonds, but the collective coordinate can be realized by a rotation along the CA-CB bond.

Strengths and Weaknesses PCA gives a new set of vectors which are complete in the same way as the Cartesian coordinates are. A lower dimensionality of the system is gained by only focussing on the first PCA eigenvectors. This reduction of dimensionality always means a loss of information for the sake of simplification; and, a priori, it is not granted that the desired features of the system are not lost when focusing on the first PCA eigenvectors only. For example, in the case of a ligand binding to a protein, the fluctuation of a side chain in the binding pocket may be uncorrelated with the large global motions of the whole protein. Nevertheless, PCA has proven extremely useful to identify large motions that are often related with the protein's function.

In PCA, a linear coordinate transformation is performed. If the relation between the individual components is not linear – think of a curved point cloud –, PCA will not be able to fully detect the underlying relation and will result in a higher dimensionality than required in case of non-linearity. Several techniques have been adopted to non-linear cases, e.g. Kernel-PCA [54].

2.2.2 Linear Regression

Principal Component Analysis focusses on the motions showing large internal fluctuations. If an additional external observable is given, the dependence of this observable on the atomic coordinates can be investigated. For given atomic coordinates $\mathbf{x}(t)$ and the observable $f(t)$, a linear model (linear regression) can be constructed for the relation $\mathbf{x}(t) \mapsto f(t)$:

$$(f(t_1) \ \dots \ f(t_m)) = \underbrace{(\alpha \ \dots \ \alpha)}_{m} + (\beta_1 \ \dots \ \beta_{3N}) \cdot \mathbf{X} + (\epsilon_1 \ \dots \ \epsilon_m)$$

where α accounts for the offset from the origin of f and $\boldsymbol{\beta}$ is the actual model. The ϵ_i are the residuals of the fit to be minimized. In detail, $\boldsymbol{\beta}$ contains the factors, with which each atomic coordinate is weighted to yield the new model coordinate. Fluctuations along this model coordinate maximize fluctuations in $f(t)$.

The number of coordinates (here: 3N) should be small relative to the number data points (here: m). Otherwise, the model can suffer due to overfitting. In section 7.1, it will be explained by a simple example what overfitting is and how cross-validation can be used to test if a model is free of overfitting. For typical protein simulations, the number of atomic coordinates is rather high and it is difficult to construct a model by linear regression that does not suffer due to overfitting.

2.2.3 Functional Mode Analysis Based On Partial Least Squares

Functional Mode Analysis (FMA) [55] is a method to assign a collective coordinate within the atomic coordinates that correlates best with an arbitrary functional property f. It addresses the overfitting problem that arises when using a simple linear regression by reducing the original data sets dimensionality with a PCA: The data are projected onto the first few components of the coordinate system given by the PCA and the regression is calculated in this new, smaller space.

The numbers of PCA eigenvectors used to form the reduced subspace depends on the functional property but may need to be rather high to capture the details of the motion. Still, for example, even a reduction to 100 dimensions for a system of $N = 5000$ atoms (and 3N coordinates) lowers the parameters down to 1%, which can have a significant effect in avoiding overfitting. Nevertheless, the PCA coordinates in general may have nothing to do with the functional property, and therefore a problem remains: A very large number of PCA components may have to be chosen to cover motions correlated with f and that – in the worst case – could lead to overfitting.

To resolve this, a second approach was developed – the one used in the hemoglobin project of this thesis – based on the main idea of FMA, but replacing the PCA dimensionality reduction with a partial least squares

(PLS) regression [56]. The PLS regression takes account of the functional property while reducing the dimensionality and thereby ensures that the important motions are part of the new reduced subspace.

Mathematics of PLS-Based FMA For a given functional property f and a trajectory $\mathbf{X}$[4] in PLS [56] the regression is not used on the full coordinates as in a simple linear regression but on new coordinates $\mathbf{V}$

$$\mathbf{f} = \mathbf{b}^T \cdot \mathbf{V} + \boldsymbol{\varepsilon}.$$

The coordinates $\mathbf{V}_i$ are defined iteratively so that each is a linear combination of the original coordinates $\mathbf{X}$ with maximal covariance to f and no correlation to the previous coordinates. Since each new coordinate is correlated with f to some extent, the number of components can be drastically reduced in comparison with the PCA-based FMA. The number n of the so-called latent vectors $\mathbf{V}_i$ is a parameter that has to be controlled by cross-validation. For example n can be increased until predictive power of the model in the independent cross-validation set decreases. Details about the different implementations of PLS can be found in the works of Denham and Helland [57, 58]. The work of Krivobokova and Briones focuses on application with relation to FMA [56].

2.3 Essential Dynamics

Essential Dynamics (ED) is a technique that applies the knowledge of principal components to enhance, or more generally, to alter sampling along these collective modes. The first implementation was based on sampling the principal components only, not considering the other degrees of freedom [53]. This approach suffered from interference of the degrees of freedom with small eigenvalues. Thus, the next implementations included all degrees of freedom with increased sampling in the essential subspace.

The present implementation in the tool `make_edi` from the GROMACS software package [30, 31] allows for different algorithms to control the dynamics during MD. This renders it possible to move the system stepwise in a specific direction in the essential space, to forbid the system to evolve in a given direction, or just keep it in a given position. In this thesis in the ABCE1 project (see Chap. 4) an option was used to move the protein towards a target structure given in ED space by only allowing MD steps that move the system closer to the target along a predefined eigenvector. Keep in mind that with the target defined in a small-dimensional subspace, the system is free in the vast majority of degrees of freedom. Another application is to increase the probability of leaving a deep energetic minimum by growing a repulsive potential at that point – the conformational flooding [59, 60].

[4] both centered for simplicity

The number of possibilities of ED increased further, since not only PCA components can be given as the ED space, but any suitable collective coordinates. More information about possible applications can be found in the works of de Groot, Grubmüller and Lange [59–62].

Mathematics of Essential Dynamics The coordinate system of the ED space is usually at rest in the center of the protein. With the protein diffusing during simulations, before applying ED modifications, the internal and external coordinate system have to be aligned. This is achieved by least squares fitting the simulation structure to the reference structure and can be written as subtraction of the center of mass ($\mathbf{x}_{COM}$) and subsequent rotation ($\mathbf{R}$):

$$\mathbf{x}' = \mathbf{R}(\mathbf{x} - \mathbf{x}_{COM}). \tag{2.4}$$

At this point changes to the structure are applied in the form of $\mathbf{x}' \mapsto \tilde{\mathbf{x}}'$. For example, if it is desired to move the system by 0.1 nm in direction of the second ED dimension, this change would be $\tilde{\mathbf{x}}' = \mathbf{x}' + (0\ 0.1\,nm\ 0\ \ldots\ 0)^{\mathrm{T}}$. After that, the transformation is reversed:

$$\tilde{\mathbf{x}} = \mathbf{R}^{-1}\tilde{\mathbf{x}}' + \mathbf{x}_{COM}.$$

If the specific ED algorithm involves force calculation as in conformational flooding, the potential is given in the collective coordinates of the ED space (here m-dimensional): $V = V(y_1, \ldots, y_m)$. Instead of transforming the forces to the ED space, the transformation from the Cartesian to the collective coordinates can be included with the chain rule of differentiation:

$$\mathbf{F} = \frac{\partial V}{\partial \mathbf{x}} = \frac{\partial V}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}.$$

The first factor is quick to compute, and the second is given by the following: Each simulation structure is transformed to the reference frame by eq. 2.4 and within the reference frame the collective coordinates are defined as in eq. 2.3.

Strengths and Weaknesses Where other methods like pulling in MD add forces on single atoms, ED directly addresses sampling in a collective subspace. This can be, as mentioned earlier, a more natural way to influence the dynamics of the system. As seen, a large variety of sampling related problems can be tackled that way.

If the ED dimensions are derived from a PCA, the coordinates are orthogonal but not necessary uncoupled. By moving a system along a PCA eigenvector with a high velocity, the other degrees of freedom will most likely not be in equilibrium. This should be considered when extracting structural information from ED simulations.

3 Collective Dynamics Underlying Allosteric Transitions in Hemoglobin

Abstract Hemoglobin (Hb) is the prototype of an allosteric protein. Still, its molecular allosteric mechanism is not fully understood. To elucidate the mechanism of cooperativity on an atomistic level, we developed a novel computational technique to analyse the coupling of tertiary and quaternary motions.

From Molecular Dynamics (MD) simulations showing spontaneous quaternary transitions, we separated the transition trajectories into two orthogonal sets of motions: one consisting of intra-chain motions only (referred to as *tertiary-only*) and one consisting of global inter-chain motions only (referred to as *quaternary-only*). The two underlying subspaces are orthogonal by construction and their direct sum is the space of full motions.

Using Functional Mode Analysis (FMA), we were able to identify a collective coordinate within the tertiary-only subspace that is correlated to the most dominant motion within the quaternary-only motions, hence providing direct insight into the allosteric coupling mechanism between tertiary and quaternary conformation changes. This coupling-motion is substantially different from tertiary structure changes between the crystallographic structures of the T- and R-state. We found that hemoglobin's allosteric mechanism of communication between subunits is equally based on hydrogen bonds and steric interactions. In addition, we were able to affect the T-to-R transition rates by choosing different histidine protonation states, thereby providing a possible atomistic explanation for the Bohr effect.

3.1 Introduction

3.1.1 Hemoglobin in the Human Body

Human red blood cells bind dioxygen molecules in the lungs and transport them through the blood vessels. In the capillaries of peripheral body tissues, they release the oxygen. The body cells use the oxygen as an oxidizing agent, e.g. to phosphorylate ADP to ATP[1]. Within the red blood cells, the protein *hemoglobin* (Hb) is responsible for binding and releasing the oxygen. Hemoglobin constitutes the largest part of each red blood cell – around 97% of the mass of its dry part [64].

[1] ATP, with its "energy-rich phosphate bond" [63], is an essential source of energy in our body, and therefore often called the "molecular energy currency".

In his book "Der Chemismus in der thierischen Organisation", published in 1840, Hünefeld describes his experimental findings about Hb binding oxygen [1]. Later, in 1866, Hoppe-Seyler reported about the reversibility of that process [65]. Still more than a century ago, in 1904, Bohr measured O_2 dissociation curves[2] that showed an unexpected sigmoidal shape and by that indicated a cooperative binding of O_2 to Hb (see Sec. 1.2).

Regulation of Hemoglobin The oxygen binding affinity of Hb is affected by several effectors. In 1967, Benesch & Benesch discovered that the small molecule 2,3-bisphosphoglycerate (BPG) plays an important role in oxygen transport [66, 67]. By stabilizing the T-state (definition below), BPG binds to Hb and thereby helps oxygen to unbind.

A second regulatory effect is the so-called Bohr effect. The oxygen binding affinity has a peculiar dependence on the carbon dioxide concentration in the blood. The dissociation curves measured by Bohr in 1904 are shifted to the right for increasing CO_2 concentration [68]. This is the so-called Bohr effect. CO_2 in the blood will partially be hydrated to H_2CO_3, which in turn partially reacts to HCO_3^- and H^+. Thereby, a high CO_2 concentration also decreases the pH. For that reason, the Bohr effect is nowadays extended to the effect of the pH value on the oxygen dissociation curves. A decreased pH shifts the equilibrium of histidine side chain protonations towards the doubly protonated side chains. Thereby the induced positive charge is thought to affect the structural ensemble [69, 70].

3.1.2 The Structure of Hemoglobin

In 1959, just one year after Kendrew resolved the first three-dimensional protein structure model with myoglobin [71], Perutz followed with the structure of horse hemoglobin [72]. Hemoglobin is a heterotetramer consisting of four α-helix rich protein chains: two α- and two β-chains. The arrangement of the chains is shown in Fig. 3.1. The alpha and beta chains are structurally rather similar (see Figure 3.2).

Within each chain a porphyrine with a central iron atom is located: the heme group. Each heme group is sandwiched by two histidine residues. One is directly bound to the iron, the other – on the opposite site of the heme plane – points towards the iron, leaving enough space for dioxygen to bind in between (see Fig. 3.3).

The crystal structures revealed several different conformations. The dominant two are the deoxy *T state* (tense) with a low binding affinity (e.g. PDB id 2HHB) and the oxy *R state* (relaxed) with a high binding affinity (e.g. PDB id 1IRD). Today, many structures are available, including oxy states, deoxy states, carbon monoxide bound structures and structures

[2] dependence of the O_2 affinity on its partial pressure

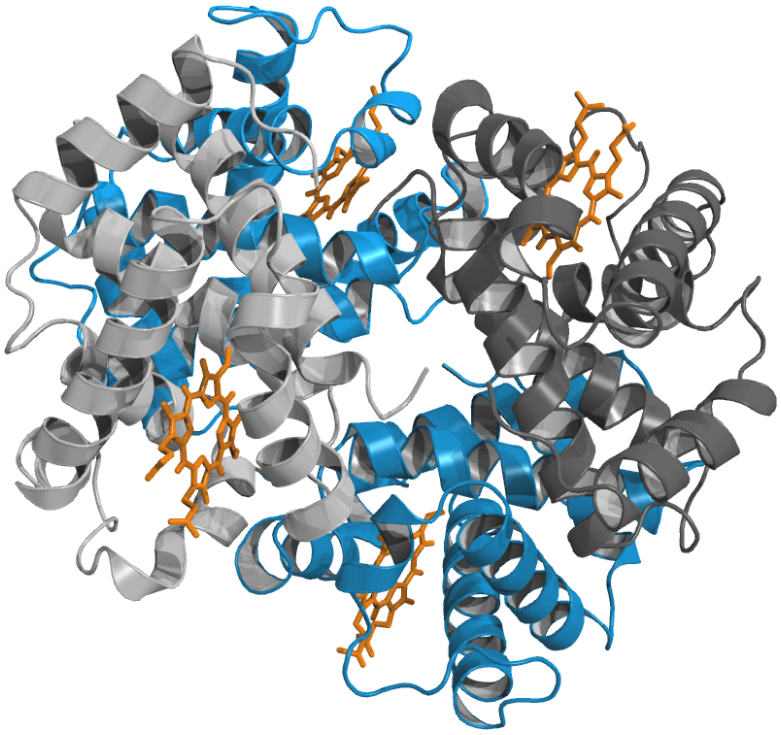

Figure 3.1: **Cartoon Representation of Hemoglobin in the T-State:** The two α chains are shown in light and dark blue and the two β chains are shown in light and dark grey. The heme groups are shown in a stick representation in orange. (PDB id 2HHB)

of a large number of point mutations [73]. Dioxygen dissociation curves are recorded for many of the mutants, making it possible to see the effect of individual residues on the cooperativity. Dynamical information is obtained from e.g. spectroscopic studies, observing transition states in the oxy to deoxy transition, and analysing specific bonds during CO dissociation [74].

3.1.3 The Cooperativity of Hemoglobin

Hemoglobin's binding affinity is dependent on the oxygen partial pressure. This dependence shows a characteristic behaviour, deviating from a typical hyperbolic shape as it would be observed for other binding processes [68]. This sigmoidal dissociation curve indicates cooperative binding behaviour (see Sec. 1.2). Hence, in comparison with a non-cooperative binding behaviour, Hb favours oxygen at high oxygen partial pressure and disfavours it at low oxygen partial pressure. This results in a more effective O_2 uptake in the lungs and an efficient release in the body tissues.

From hemoglobin's dissociation curve it can be deduced that it is an oligomer. This can be seen when comparing to the dissociation curve of Hb to myoglobin, which resembles just one chain of Hb. Myoglobin

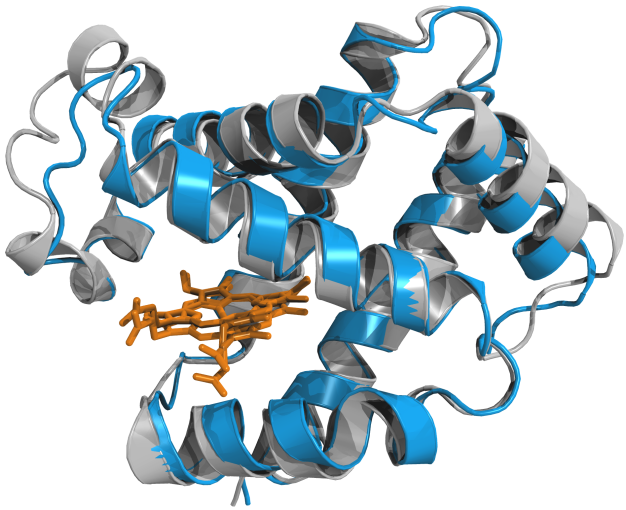

Figure 3.2: **Cartoon Representation of the Superpositioning of Hemoglobin's α- and β-Chain.** The α chain is shown in blue, the β chain in grey, the heme groups are shown in orange sticks.

does not show sigmoidal characteristics but rather a non-cooperative hyperbola [75]. Indeed, the structure of Hb confirmed its multimeric conformation.

If dioxygen binding in one of Hb's chains changes the binding affinity of the other chains, there must be an information flow between the distant (allosteric) binding sites. When dioxygen binds in one of Hb's chains, the chemical environment in that binding site is changed. This change may lead to a local rearrangement of the amino acids and the heme group. Since the binding site completely within one chain, this conformational change is tertiary. To connect structural changes (induced by binding) in two chains, a rearrangement may occur of the chains with respect to each other, i.e. a quaternary change.

The necessity of this quaternary change for this allosteric coupling was elegantly shown by slowing down Hb's conformation transitions by trapping it in silica gels, resulting in non-cooperative binding characteristics [11, 12].

Models for Cooperativity Theoretical models have been developed to explain the allosteric coupling of the four Hb chains including the pioneering works of Monod, Wyman and Changeux (MWC model [76]) and Koshland, Nemethy and Filmer (KNF model [77]). The ground-breaking MWC model describes the whole Hb by two crystallographic states: a low binding affinity tense state (T) and a high binding affinity relaxed state (R). Both states occur in equilibrium with certain probabilities. Each dioxygen binding to Hb shifts this equilibrium towards the R state. The KNF model extends the MWC model to T- and R-states for the individual

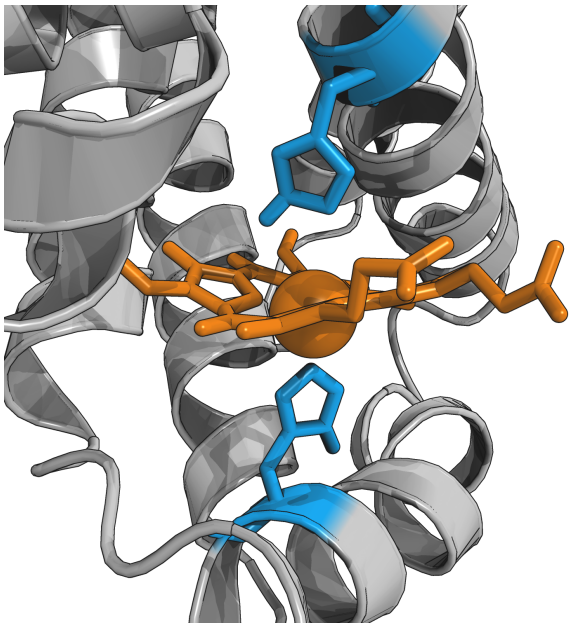

Figure 3.3: **The Heme Group in Hemoglobin:** The Hb chain β in the T state (grey) is shown together with the heme group including the iron atom (orange), the proximal (blue, bottom) and the distal histidine (blue, top). Dioxygen binds in the gap between the distal histidine and the iron atom.

protein chains (overview over different models in [78, 79]). Here, the binding event in each chain increases the binding affinity in the other chains. As both models have their strengths and weaknesses, new models were developed to incorporate more details of Hb's cooperativity, e.g. the Tertiary Two State (TTS) model of Henry et al. [80]. According to this model, the chains each have two conformations and Hb as a whole has two. A possible sequence of events, leading from one chain binding oxygen to oxygen binding in all chains, is depicted in Fig. 3.4.

The previous models have described the allosteric mechanism by states containing only one structure each. Because of the ensemble nature of proteins, this description bares limitations. Cooper & Dryden showed in their theoretical work, that allostery could – in principle – also be manifested without a change in the average states, but only a change in the width of the conformational distribution [81]. This entropic contribution (e.g. changing the stiffness of a binding site without changing the average conformation) is included in recent studies like the ensemble allosteric model by Hilser et al. [82].

3.1.4 Molecular Dynamics Simulations of Hemoglobin

In addition to experiments, Molecular Dynamics (MD, see Sec. 2.1) simulations have provided insight for shorter timescales from μs down to ps

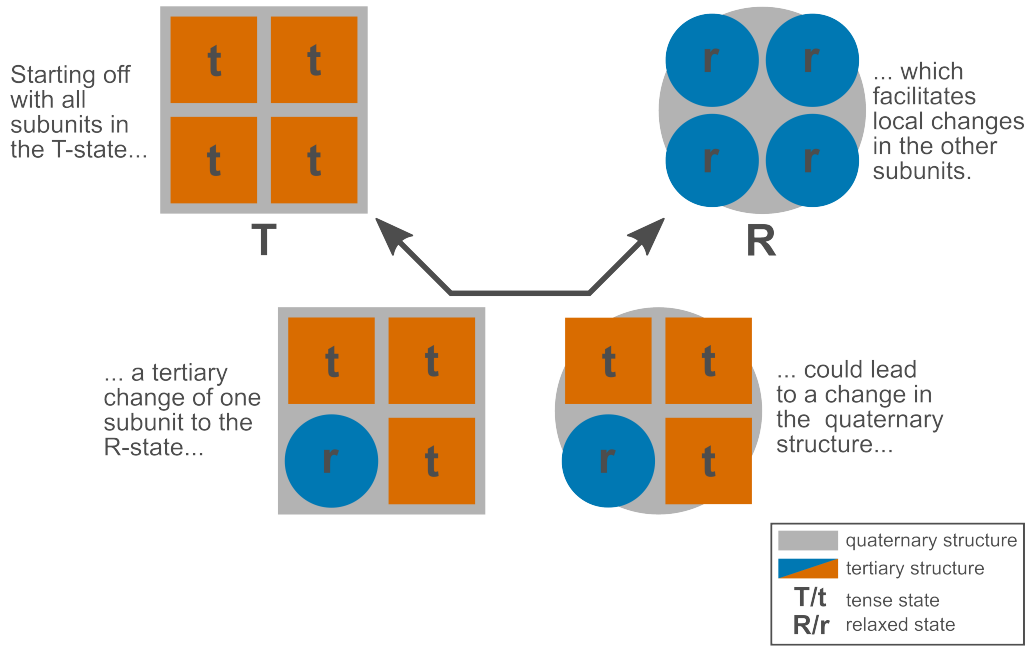

Figure 3.4: **Schematic Representation of a Possible Pathway from the T- to the R-State.**

while maintaining the full atomistic picture of Hb. The works of Shadrina et al. and of Lepeshkevich et al. focused on O_2 diffusion in Hb studied with MD [83,84]. Ramadas and Rifkind simulated conformational changes due to perturbations of the heme pocket for methemoglobin dimers [85]. In the work by Mouawad et al., the Hb T-to-R transition was enforced by restraining the Hb coordinates with decreasing structural distance to the R-state structure [86]. The study from Yusuff et al. focused on 100 ns simulations from different crystallographic structure models [87]. Recently, J. Hub and co-workers observed for the first time spontaneous reproducible transitions from the T- to the R-state during MD simulations [88], matching best to the TTS model of Henry et al. [80]. They described a tendency for the β-chains to couple more strongly to the quaternary motion than the α-chains.

3.1.5 Scope of This Study

In the present study we investigated how the local intra-chain motions couple to the global inter-chain motions on a molecular level. To this end, we enhanced the statistical basis for the transition trajectories with respect to the original set [88], and developed a method which allows to characterize the coupling between global and local motions. For this purpose, we first separated global from local motions and then identified the

coupling mechanism between them. We analysed the resulting coupling collective coordinate on the level of molecular contacts, shedding light on the molecular allosteric mechanism of hemoglobin. In addition, by using a different set of histidine protonations with a higher fraction of doubly protonated side chains, we were able to show a reduction in the number of transition trajectories. This constitutes a possible explanation for the Bohr effect in Hb.

3.2 Results

3.2.1 Molecular Dynamics Simulations

From the Hb simulations carried out by Hub and co-workers [88] the ones starting from the T-state with doubly protonated and thus positively charged His(β)146 (all other histidines neutral) showed transitions to the R-state in all three runs. In our study, we extended these simulations to improve the statistical basis of the T-to-R transitions. From a total of 50 simulations (200 ns long each; 10 μs total simulation time) 22 showed a spontaneous transition (see paragraph 3.5.1). The specific simulation setup is described in Table 3.5.1. For the further steps, only transition trajectories were taken into account.

3.2.2 Coupling of Quaternary and Tertiary Motions

Starting from our T-to-R transition trajectories, to analyse the interplay of local and global motions, we separated local from global motions as the first step. Here, the MD transition trajectories were decomposed into two trajectories: quaternary-only (Q) and tertiary-only (T). The first consists of inter-chain motions with the Hb chains translating and rotating as rigid bodies and the second contains intra-chain motions, omitting the global movements. The combination of the Q and T trajectories yields the full MD trajectories. For a visual explanation of the basic idea of the decomposition see Figure 3.5, for a detailed description see section 3.5.2.

The two corresponding subspaces for Q and T are orthogonal by construction, but the actual motions along them may still be correlated, thus reflecting the underlying allosteric mechanism in hemoglobin. We therefore investigated if there was a coupling between local (T) and global (Q) motions. In other words: can we construct a linear combination of the T coordinates that is correlated to the Q motion?

We simplified the Q trajectory by only considering the first eigenvector of a Principal Component Analysis (PCA) as the most dominant motion (referred to as cQ , details in 3.5.2). For obtaining a collective coordinate within T that maximally correlates to cQ, we applied Functional Mode Analysis ((FMA), [55]) based on Partial Least Squares [56]. We assessed the risk of overfitting, arising from the high dimensionality (d=13644) of the

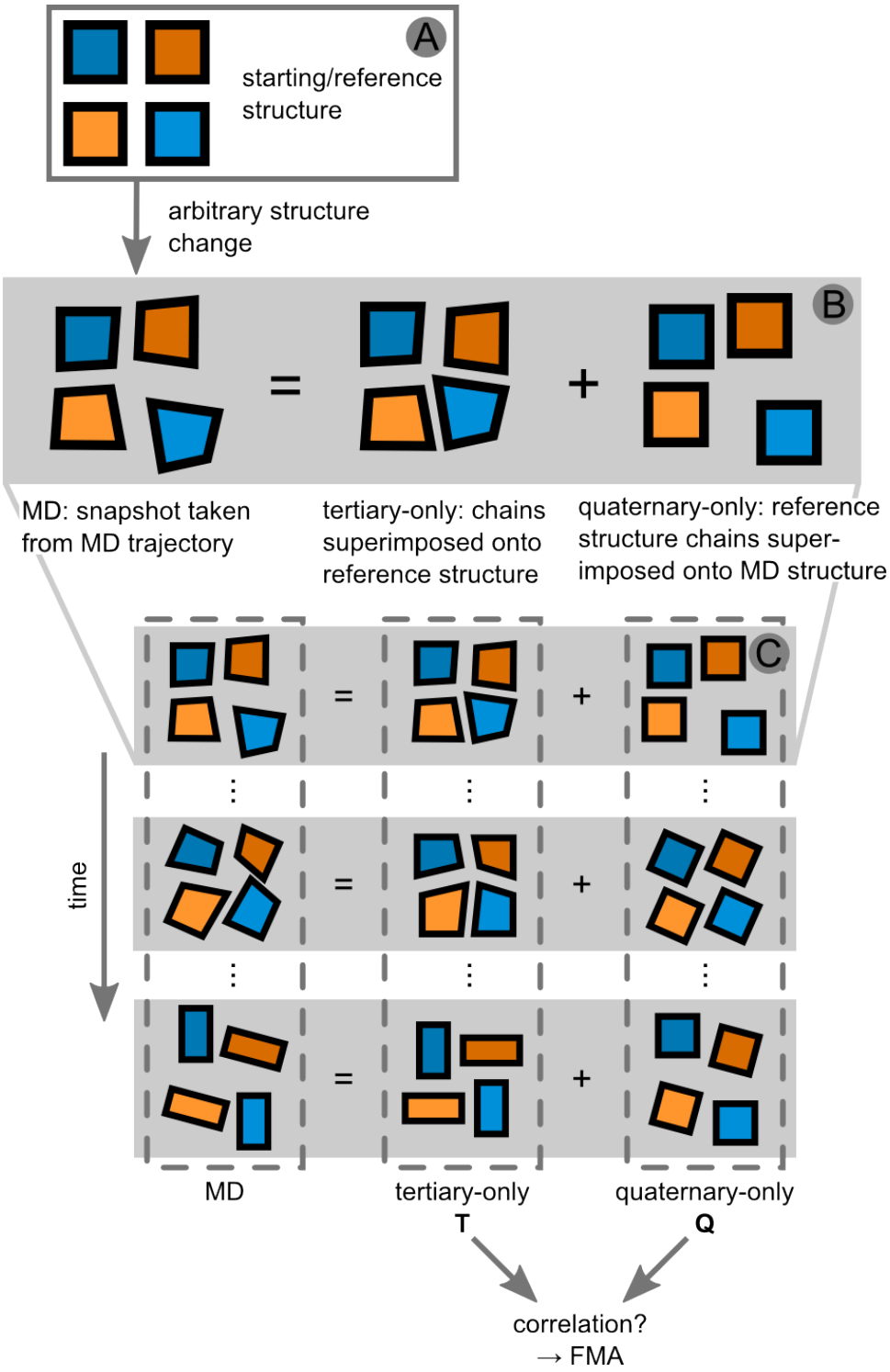

Figure 3.5: **Illustration of the Separation of the MD Trajectories into *T* and *Q* Trajectories:** On top is shown how a single MD snapshot is decomposed (B) with respect to the reference structure (A). This procedure is applied to all snapshots yielding the two desired trajectories of intra- and inter-chain motions (C). The schematic system was chosen to resemble Hb with its four chains.

T space, by cross-validation. A detailed description of the application of FMA can be found in 3.5.2. The resulting tertiary FMA model correlated to the cQ motion with a Pearson correlation coefficient of $R = 0.98$ (see Fig. 3.6) This means that despite the orthogonal nature of the two underlying subspaces, T and Q, we found a coordinate within T that is strongly coupled to Q. This allows us to predict the quaternary state from the internal subunit coordinates alone, thereby providing insight into the allosteric mechanism of subunit communication through quaternary conformation changes.

Within the FMA framework it can be desirable to reweigh the individual latent vectors with their contribution to the overall variance (see [55]): While the FMA mode *cT* is the maximally correlated motion, which may actually be restricted within the protein and not happening in simulations, the ensemble-weighted mode *cTew* is the most probable motion that correlates with the functional property. For further analysis, we used this ensemble-weighted motion (*cTew*).

3.2.3 Molecular Coupling Mechanism

Now that we found a collective motion (*cTew*) within the local intra-chain motions (T) that correlates strongly to the global inter-chain motions (Q), we can investigate the underlying molecular mechanism for this allosteric coupling. For a closer look, we reassembled the coupled tertiary and quaternary motions in the following way: Starting from the T-state we moved stepwise along *cQ* and independently along *cTew*. The step size was chosen to be 1/20 of the distance between the extreme projections of the simulations. This provided a grid of 20 x 20 structures in the plane spanned by *cQ* and *cTew*. This is the smallest subspace showing the coupling of local and global motions as derived from our simulations. For the subsequent contact analysis, we picked a specific pathway in this plane. Starting from the T-state, the first part of the path is along *cQ*, and the second along *cTew*. In Fig. 3.7 Hb structures along both parts of the path are shown. The path is also marked in white in Figure 3.8 and will be referred to as *cQ-cTew*. This pathway artificially separates motions that are occurring simultaneously in the simulations. This allows us to classify the contacts according to their decomposition into global and local motions.

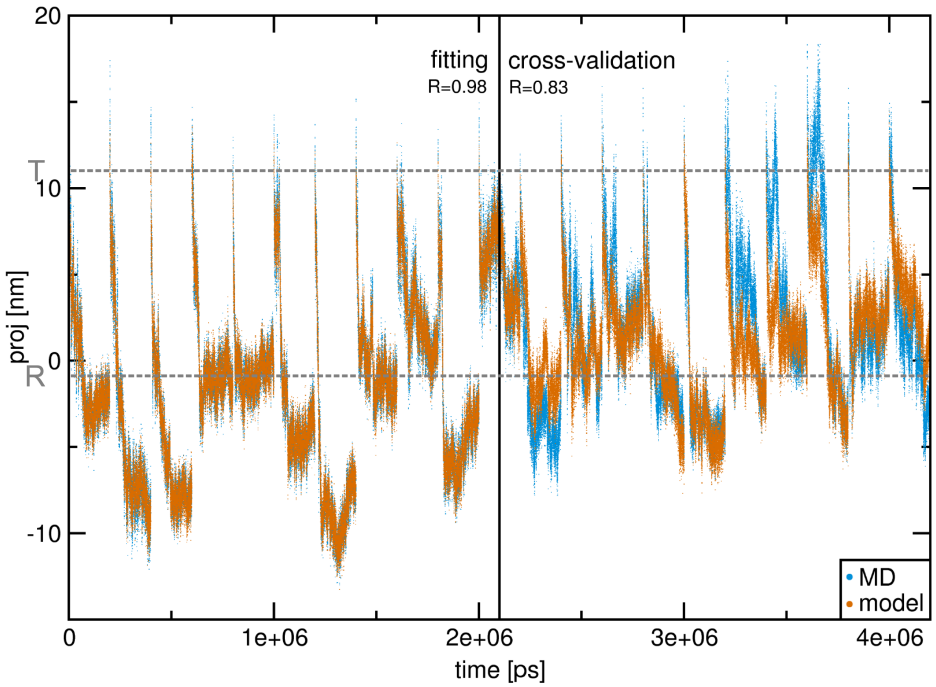

Figure 3.6: **Functional Mode Analysis Input Data and Fit Results:** Projections of the concatenated MD trajectories onto cQ (blue) and onto the constructed model cT (orange). The first half of the data has been used for constructing the model and the second half for cross-validation. Pearson correlation coefficients comparing MD data and FMA model for both parts are shown on top. The x-axis is the consecutive time in ps and the y-axis the projection onto the principal quaternary eigenvector in nm. The projections for the T- and R-state X-ray structures are marked (dotted grey).

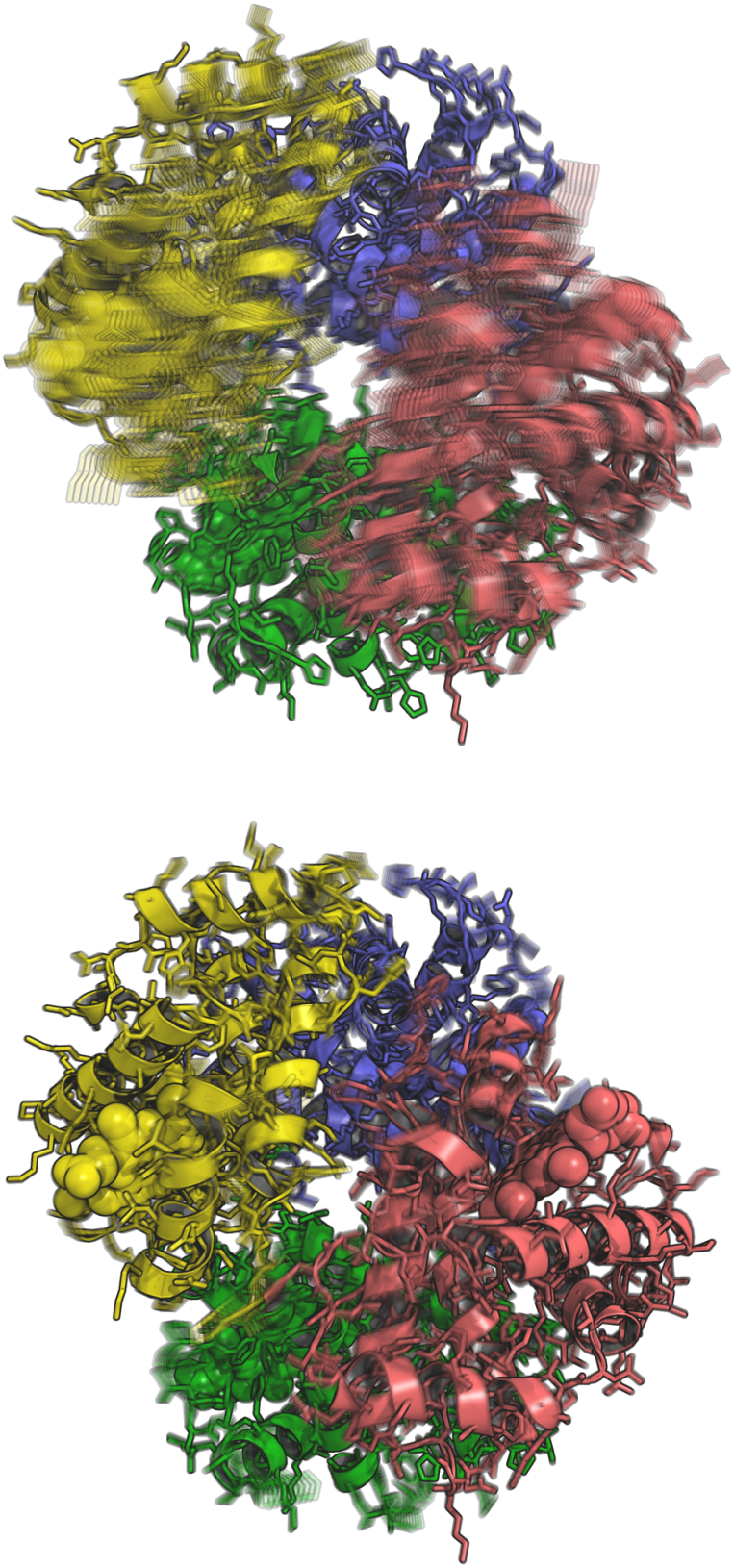

Figure 3.7: **Collective Coordinates *cQ* & *cTew*:** An overlay of the structures along the *cQ*-*cTew* path is shown. The first half along *cQ* (top) and the second half along *cTew* (bottom).

In order for the information of a local conformation to flow from one protein chain to another, it has to cross the corresponding interface. It is therefore of interest to investigate interactions at the subunit interfaces. Different interaction mechanisms for the coupling of local and global motions were considered. Inter-chain attractive interactions could pull at the protein chains, re-arranging them locally while being moved globally. Alternatively, repulsive interactions between protein chains could push subunits towards a different global conformation. For that reason, we focused on inter-chain van der Waals overlaps and general distance based contacts to investigate the interactions underlying the allosteric coupling mechanism.

Van der Waals Overlaps To estimate the influence of interatomic repulsion due to van der Waals (vdW) overlaps as a driving force for the allosteric coupling, we calculated the overlap of atomic vdW spheres (more details in 3.5.3). We did this for the 20×20 structures in the plane spanned by cQ and $cTew$, and took into account only overlaps between atoms from different protein chains. The higher the overlap in a specific structure, the more energetically unfavourable it is. As can be seen in Figure 3.8, the overlaps are minimal along the main diagonal while increasing when moving orthogonally. Projecting the structures from the MD simulations (white dots) onto this plane shows that they coincide with the low vdW overlap region.

Hydrogen Bonds Hydrogen bonds are crucial for secondary and tertiary protein structure formation. Since hydrogen bonds between Hb's subunits are also important for its quaternary structure, we analysed inter-subunit hydrogen bonds along the quaternary T-to-R transition. Therefore the structures of the transition trajectories were ordered according to the projection onto cQ , and hydrogen bond energies were estimated using the Espinosa formula [89] for each of the structures. With our focus on the subunit interface, only inter-chain hydrogen bonds were considered, resulting in 36 hydrogen bonds.

In the top plot in Fig. 3.9 the energies for the specific observed hydrogen bonds are shown. The hydrogen bond energies fluctuate strongly, showing mostly on/off patterns. To focus on the larger trend in energies of the individual hydrogen bonds, the energies have been averaged over neighbouring structures with the result shown in the lower plot in Fig. 3.9. For most hydrogen bonds no clear trend is apparent. An exception is the interaction between the side chains of Arg$(\alpha_2)31$ and Gln$(\beta_2)127$ (marked as "30" in Fig. 3.9), which weakened along the transition.

Contact Analysis For a broader view including hydrogen bonds as well as vdW interactions, we monitored inter-chain atom pairs showing a

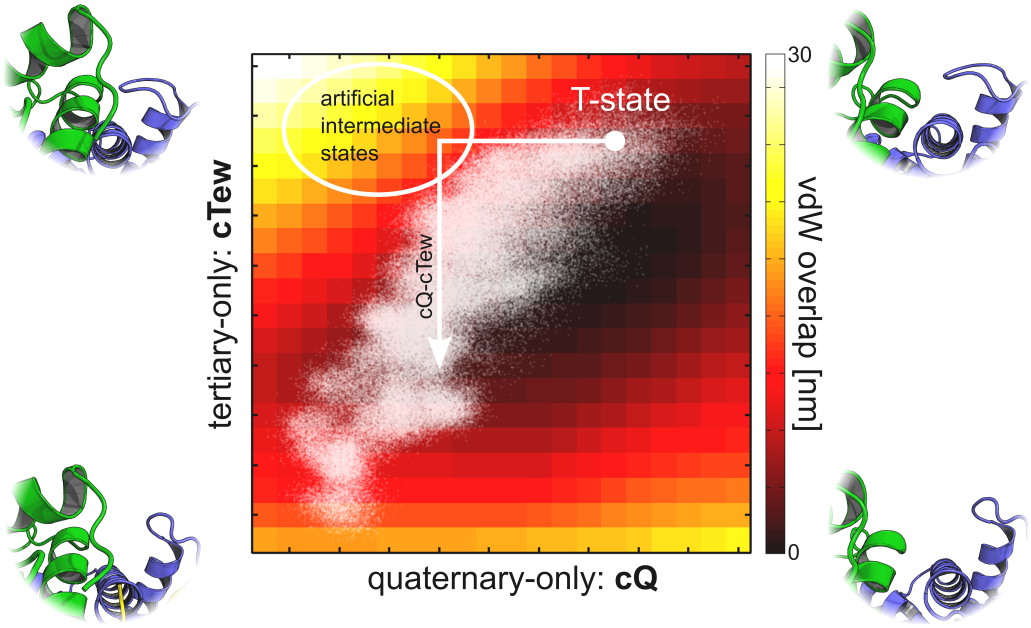

Figure 3.8: **Graphical Representation of the VdW-Overlap Analysis:** Van der Waals overlaps were calculated for structures in the plane spanned by *cQ* (x-axis) and *cTew* (y-axis). For the extreme structures in the four corners a zoomed-in part of Hb is shown to illustrate the motions. Projections of the original simulation data onto this plane are shown as white dots.

distance smaller than 3Å. This analysis was carried out along *cQ-cTew* which allows us to classify the contacts according to when the contact is or is not formed along this specific path. For specific interaction types following contact patterns are expected. The two residues[3] in contact are:

pulling at each other, breaking the contact when in the off-diagonal intermediate artificial states (see Figure 3.8) and maintaining it close to the T- and R-state,

pushing each other getting close only in the off-diagonal intermediate artificial states when not moving along *cQ* and *cTew* together,

switching one of the interacting residues for another while moving along *cQ-cTew*.

Pulling Contacts In Table 3.1 observed contacts are listed that fall in the first category. These contacts only stay intact if the system moves along *cQ* and *cTew* together, but break if moving in one or the other direction

[3] Atom contacts were translated into residue contacts if at least one atom of both residues was closer than 3 Å.

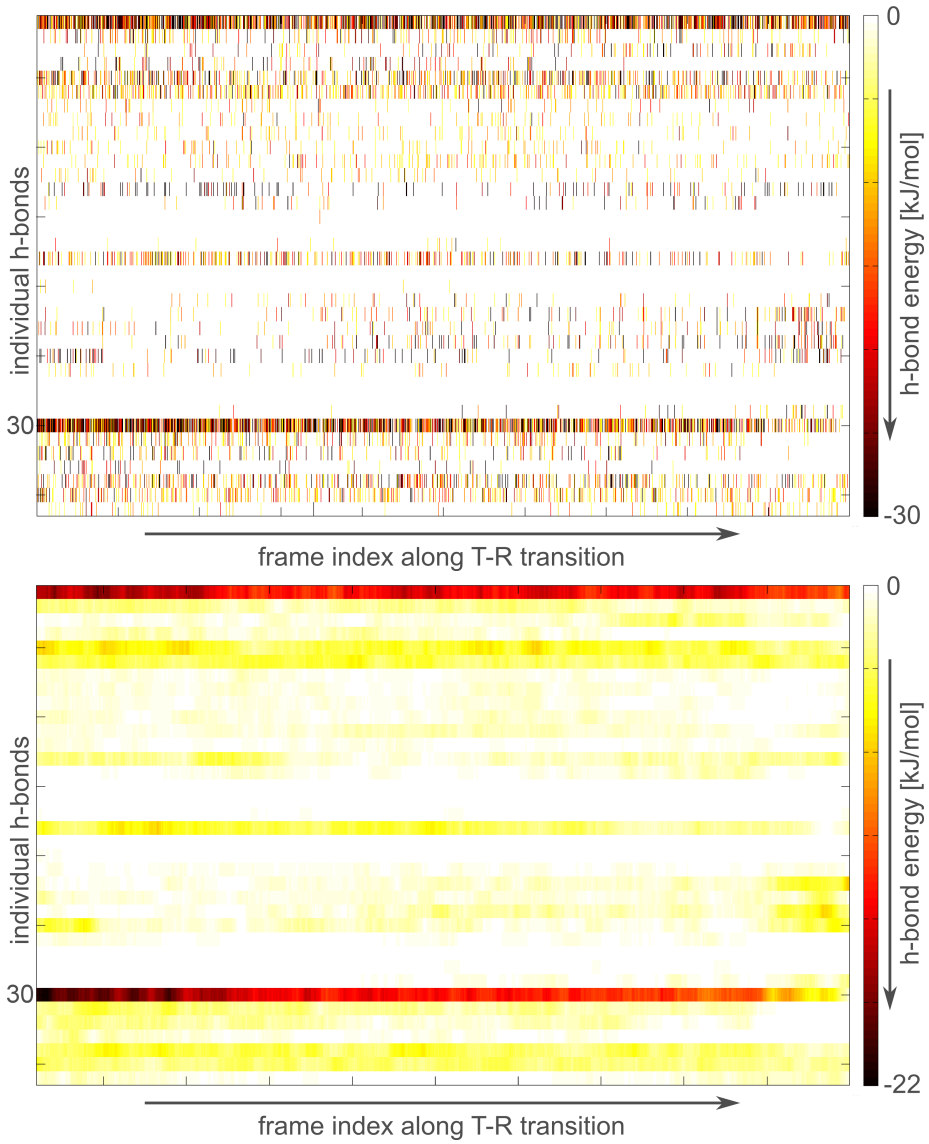

Figure 3.9: **Hydrogen Bond Analysis:** For 5000 MD structures ordered along the T-to-R transition, hydrogen bond energies were estimated using the Espinosa formula (top). The figure below shows the same data while averaging the energies of 50 neighbouring structures.

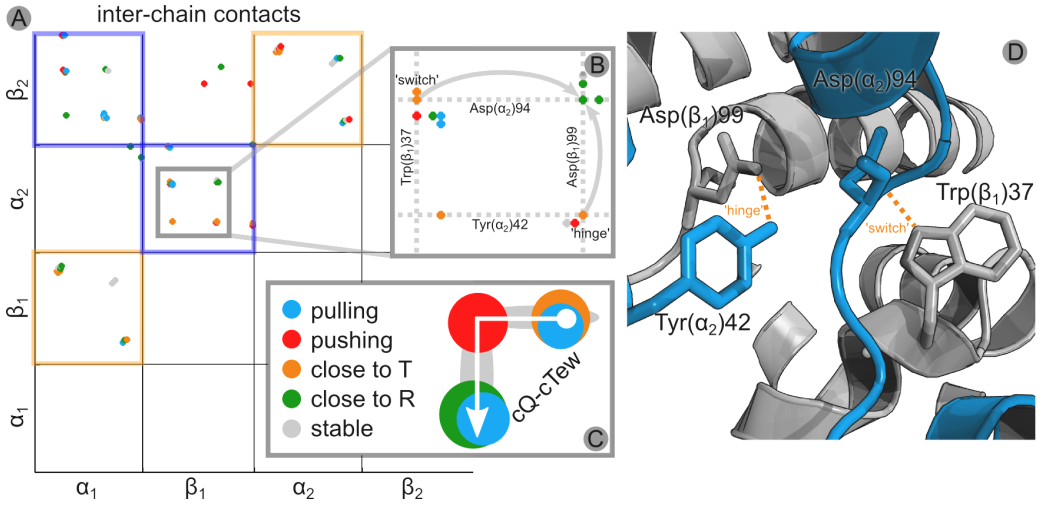

Figure 3.10: **Inter-Chain Contact Analysis:** (A) The matrix of observed contacts is depicted with the colour indicating the contact class. (B) An exemplary close-up on the contact region including the 'switch' and 'hinge' contacts (structure shown in (D)) as defined by Balakrishnan et al. [74]. The arrows illustrate the switching behaviour of the Asp(α_2)94 and the Asp(β_1)99 residues. (C) Schematic representation of the contact classifications along *cQ-cTew*.

independently. This is the expected behaviour for contacts which must remain intact for the allosteric mechanism to function. Exemplarily, this was observed for Phe($\alpha_{1/2}$)117 and Arg($\beta_{1/2}$)30. The hydrogen bond between the carboxylic oxygen of Phe and the side chain of Arg breaks while moving from the T-state towards the off-diagonal intermediate artificial states (see Figure 3.8), and forms again when approaching the R-state.

Table 3.1: **List of observed pulling contacts (type 1)**

residue 1	residue 2
Lys(α_1)40	His(β_2)146
Tyr(α_1)42	Asp(β_2)99
Arg(α_1)92	Gln(β_2)39/Glu(β_2)43
Pro(α_1)95	Trp(β_2)37
Phe($\alpha_{1/2}$)117	Arg($\beta_{1/2}$)30
Tyr(α_1)140	Pro(β_2)36
Tyr($\alpha_{1/2}$)140	Trp($\beta_{2/1}$)37
Arg($\beta_{1/2}$)40	Arg($\alpha_{2/1}$)92/Leu($\alpha_{2/1}$)91
Ala(α_2)110	His(β_2)116

Pushing Contacts Contacts of the second category, which appear only while moving along *cQ* and *cTew* individually, are listed in Tab. 3.2. One scenario how these contacts could be leading to the allosteric mechanism is the residues getting too close when not moving along *cQ* and *cTew* together. This could be the case for repulsive vdW or coulomb interactions. A clear example for this was the interaction of Lys(β_1)82 and Lys(β_2)82 we observed: Close to the T-state both side chains are pointing into the solvent. While moving along *cQ*, the two β chains approach each other and bring both positively charged side chains unfavourably close. The motion along *cTew* relaxes this repulsive interaction by bending the N-terminal ends of the F helices (the helix notation goes back to Watson, Kendrew and Perutz [90]).

Experimental studies introduced cross-links between the two lysines [91, 92]. The derived structure was described to be an intermediate between T- and R-state with characteristics of both states but no cooperativity. This is in accord with our analysis, from which we saw that a linker between the lysines would make the F helix bending impossible.

Table 3.2: **List of observed pushing contacts (type 2)**

residue 1	residue 2
Pro($\alpha_{1/2}$)37	His($\beta_{2/1}$)146
Thr(α_1)38	Pro(β_2)100/Tyr(β_2)145
Thr($\alpha_{1/2}$)41	Val($\beta_{2/1}$)98
Trp(β_1)37	Arg(α_2)92
Lys(β_1)82	Lys(β_2)82
His(β_1)143	Lys(β_2)82
Leu(α_2)34	Ala(β_2)128
Phe(α_2)36	Gln(β_2)131
Asp(α_2)126	Tyr(β_2)35

Switching Contacts If during the transition one residue switches an interaction partner, we expect to see the first contact disappearing and a contact with the new residue appearing. This was observed e.g. for the C-terminal Arg(α_1)141. Its side chain interacts with the carboxyl group of Val(β_2)34, and switches along *cQ*-*cTew* so that a salt-bridge is formed between the Arg terminus and the side chain of Lys(α_2)127. This event also has been seen in the symmetry-related counterpart independently. Further contacts of this type are listed in Table 3.3.

3.2.4 Rotational Correlation of Amino Acids

To further investigate structural fluctuations within and between the protein chains, we looked at how the rotations of each amino acid backbones

Table 3.3: **List of observed contacts switching during transition (type 3).** (*) In Hb Kempsey the mutation Asp(β)99→Asp(β)99 increases the O_2 affinity [93, 94]). The hydrogen bond Asp(β_1)99-Tyr(α_2)42 was analysed by Balakrishnan et al. and named "switch contact" [74]. (**) Asp(α_2)94-Trp(β_1)37 is the "hinge contact" analysed by Balakrishnan et al. Both hydrogen bonds are reported to form during transition from R to T.

residue	contact in T	contact in R
Arg($\alpha_{1/2}$)141	Val($\beta_{2/1}$)34	Lys($\alpha_{2/1}$)127
Asp(β_1)99(*)	Tyr(α_2)42	Asp(α_2)94/Val(α_2)96
Asp(α_2)94(**)	Trp(β_1)37	Asp(β_1)99/Glu(β_1)101
His(β_1)146	Lys(α_1)40	His(β_2)2

are correlated. During my diploma thesis, I developed the method applied here [95]. In short, it calculates for each amino acid how the relatively rigid fragment of C_α-C_β-N-C rotates between two structures. That way we assign a rotation vector to each amino acid. The norm of the difference of two rotation vectors then defines a distance metric between two amino acids. For a trajectory, the distances are summed up to yield an averaged rotational distance between two amino acids. This allows to separate rigid domains from flexible protein regions.

Applied to the Hb simulations the correlations of the amino acid pairs can be collected in a matrix (see Fig. 3.11). As it contains four rigid bodies, the Q trajectory shows the expected block behaviour, indicating that the amino acids within a protein chain are correlated more strongly than between the chains.

From the matrix for the full MD trajectory in contrast, it is not possible to detect the four protein chains. This suggests that the tertiary motions, which constitute the difference between both matrices, are dominating the dynamics from the perspective of this analysis.

The maximum rotational-similarity distance is higher in the case of the full MD trajectory than the Q motions only, which is consistent with the lower number of degrees of freedom.

3.2.5 Influence of Histidine Protonation

We aimed to analyse the effect of the histidine protonation states on the transition probabilities in addition to the protonation states used by J. Hub et al. [88]. For this purpose, a second set of protonation state was simulated. We chose to use the protonation states as reported by Kovalevsky et al., who used neutron protein crystallography to measure the protonation of histidine residues in the T-state [96].

Both histidine protonation states are listed in Table 3.4. In the case

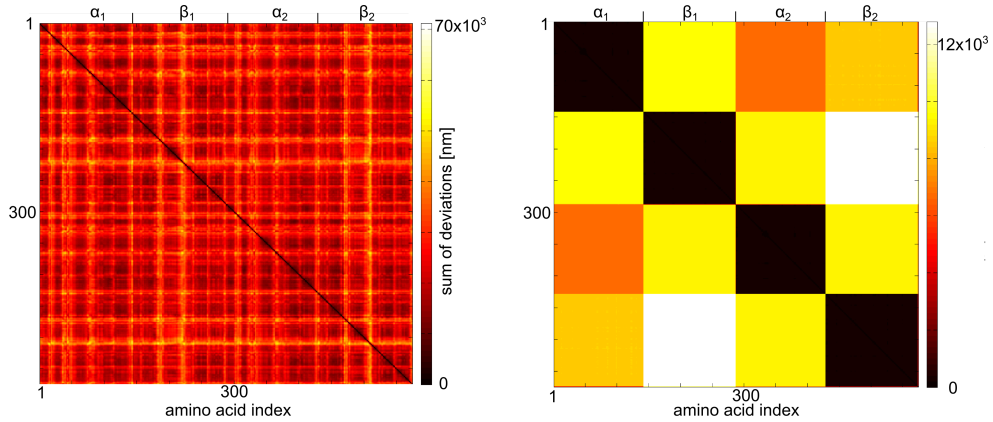

Figure 3.11: **Rotational Correlation of Amino Acids:** Pairwise rotational correlation for each amino acid is shown for the full MD trajectory (right) and the Q trajectory (left).

of the protonations used by Hub et al., 13 out of 20 simulations (only comparing simulations with cut-off of 1.4 nm for the vdW interactions) showed a transition, while in the case of the protonation state described by Kovalevsky et al., only 4 out of 20 simulations did (see Table 3.5). This suggests a clear dependence of the transitions on the histidine protonation state (further details on the calculation of the statistical significance in Sec. 3.5.4).

3.3 Discussion

3.3.1 Coupling of Quaternary and Tertiary Motions

The correlation of $R = 0.98$ for the model fitting (and $R_C = 0.83$ for the cross-validation) between the quaternary mode and the detected tertiary mode is high, and allows us to predict Hb's quaternary conformation for a given tertiary conformation. We were able to detect this coupling despite the fact that we did not take the full Q motions into account, but rather reduced the motions to the first PCA eigenvector cQ. In a future study, a more complete interaction picture may be derived by coupling the tertiary motions to the full 18 dimensional quaternary subspace.

The model that we used to describe the coupling of quaternary and tertiary motions is of linear nature. On the one hand, since there is no necessity for a linear coupling, non-linear models could be more suitable. On the other hand, the fact that we found a linear model with an acceptable correlation in the cross-validation makes us confident that already a major part of the coupling can be described linearly. We investigated only instantaneous cQ to T coupling, although in reality there might be a lag-time due to the time required for the signal to pass. Our results, however, show that the coupling can already be identified from

Table 3.4: **Comparison of the histidine protonation states used in this study:** "0" indicates the neutral side chain, "+1" the doubly protonated and thereby positively charged side chain. Histidines bound to heme groups are shown by "H". Residues for which the protonation was not derived are shown in brackets. In that case we used the protonation by Hub et al. The measured protonations by Kovalevsky et al. are different for the both α resp. β subunits whereas Hub et al. used symmetric protonation states.

	Hub	Kovalevsky			Hub	Kovalevsky	
His(α)	$\alpha_{1/2}$	α_1	α_2	His(β)	$\beta_{1/2}$	β_1	β_2
20	0	+1	0	2	0	(0)	(0)
45	0	0	0	63	0	+1	0
50	0	0	+1	77	0	0	0
58	0	+1	0	92..Fe	H	H	H
72	0	+1	+1	97	0	+1	+1
87..Fe	H	H	H	116	0	+1	+1
89	0	0	+1	117	0	(0)	0
103	0	+1	+1	143	0	0	+1
112	0	+1	+1	146	+1	+1	0
122	0	0	0				

an analysis of instantaneous correlations. Future work may include non-linear models as well as delayed responses for the coupling, especially for systems in which the order of events is known.

The rotational correlation of amino acids was not able to distinguish the four protein chains when applied to the MD trajectories. In contrast, when applied to the Q motions, the protein chains separated clearly as expected. This suggests that the tertiary motions are dominating the rotational dynamics of Hb. Hemoglobin seems not to behave like four rigid bodies with local structure changes, but more like one clump of four soft chains. This hypothesis still needs to be tested, and it may be true for other proteins as well. Also, this direct analysis of motions does not yield any coupling between local and global motions and thus hints at the necessity of our specialized procedure to identify the coupling coordinate from the T coordinates.

3.3.2 Molecular Coupling Mechanism

VdW Repulsions & Hydrogen Bonds The interaction picture we derived from our analysis suggests that van der Waals interactions are a global driving force for the allosteric coupling. The fact that MD trajectories projected onto the plane *cQ -cTew* coincide with the region around the diagonal in Fig. 3.8 that corresponds to low van der Waals overlap

strongly points at the underlying coupling mechanism: With *cTew* being derived to optimize the coupling between local and global motions, it also shows a strong coupling of sterical repulsions between local and global motions.

Nevertheless, further efforts to break down the global interactions to individual repulsive contact pairs did not yield a conclusive picture. This suggests that the vdW repulsions at Hb's inter-chain interfaces do not act on a residue level, but on a broader, collective scale.

In contrast to the vdW analysis, our hydrogen bond analysis did not produce strong global trends. In additional correlation measurements of individual hydrogen bonds strengths with quaternary transitions, we were not able to identify key hydrogen bonds, indicating that these play only a secondary role in the allosteric coupling mechanism.

Contact Analysis The contact analysis along *cQ*-*cTew* allowed us to classify contacts according to their behaviour along this path. By picking this path in the plane spanned by *cQ* and *cTew*, we ensure that the observed contacts are important for the coupling. If any contact pair did not play a role in the allosteric coupling, it would not have been part of the coupling of global and local motions. For a number of contacts we observed also the symmetry-related residue-pairs (if not a contact of the same type, at least the contact itself), which assured us of the significance of these contact pairs. The similar number of pulling and pushing interface interactions suggests that both types contribute to the allosteric coupling in equal measure. The fact that sterical repulsions and hydrogen bonds could not be unambiguously traced down to a residue level individually, but could in the generalized contact analysis, points at an interplay of repulsive and attractive interactions.

Notable Addition In our analysis the two residues $\mathrm{Asp}(\alpha_{1/2})126$ and $\mathrm{Arg}(\alpha_{2/1})141$ stay in contact along the full *cQ*-*cTew* path, that is close to R and T as well as in the off-diagonal intermediate artificial states. Hence, since the contact is not changing in the coupling space spanned by *cQ* and *cTew*, this salt-bridge does not seem to play a role in the coupling of local and global motions. Nevertheless, it was shown in Hb Montefiore that the mutation of $\mathrm{Asp}(\alpha_{1/2})126$ to a Tyr breaks down the cooperativity [97]. Further studies are needed to investigate why we did not detect this contact pair. The applied dimensionality reduction from Q to cQ may have been causing this.

Mutations In this study, we assigned different roles in the coupling mechanism (like forming hydrogen bonds or repulsion due to van der Waals interaction) to individual amino acids. Mutagenic studies of these

amino acids provide a direct means to validate the predicted role of individual amino acids in the allosteric mechanism.

The observed contacts which are caused by van der Waals repulsions may be reduced by mutating to residues with smaller side chain sizes. In the case of charged side chains introducing an additional charge of the same sign may increase the repulsion. Contacts including hydrogen bonds can be suppressed by using unfavourable mutations.

We suggest mutations affecting the "hinge" and "switch" contacts [74] in Table 3.3. Our observations extended this region by interaction partners in the R-state allowing to choose mutations affecting either the T- or R-state. By introducing a hydrogen bond donor the mutation Val(α_2)96→Thr may stabilize the R-state. The central role of the Asp(β)99 in this area makes it an interesting mutation site since it might separate the two timescales associated with the "hinge" and "switch" contacts as described by Balakrishnan et al.

The repulsive interaction of the two Lys(β)82 can be explored by either a mutation to Arg or even by switching both charges by a mutation to Glu, keeping the repulsive Coulomb interaction.

Further, we suggest a mutation of the Arg(α)141→Lys to analyse the details of this salt-bridge by this conservative mutation, leaving the charges untouched and only changing the side chain length.

3.3.3 Influence of Protonation

The measured histidine protonations in the T-state by Kovalevsky et al. [96] showed a high number of doubly protonated and thus positively charged side chains. This is in agreement with the Bohr effect, stating the T-state to be more stable at low pH. By applying these protonations to our MD simulations we were able to observe significantly lowered transition probabilities from the T- to R-state.

Even though our simulations are not long enough to be considered at equilibrium conditions, our observations can be taken as a proof of concept of the electrostatic interactions stabilizing the T-state and thereby underlying the pH-driven Bohr effect. During the T-to-R transition residues at the α_1/α_2 and the β_1/β_2 interfaces get closer. Positively charged histidine residues at these interfaces add a repulsive Coulomb force, rendering the transition energetically more unfavourable. Interestingly, most of the histidines that are changed from neutral to positive when applying the T-state protonations, are located on the outside of Hb and not at chain interfaces. A careful introduction of additional histidines by mutation – as pH sensitive switches – may increase the Bohr effect. Also, calculation of free energy differences between the T- and R-state upon protonating histidine residues may yield direct, quantative insight into the contribution of individual histidines to the total Bohr effect.

3.4 Conclusions

We developed a novel method to extract the underlying allosteric coupling mechanism from spontaneous Hb transition trajectories.

One fundamental component of this method is the separation of local/tertiary and global/quaternary degrees of freedom (DOF). In this study, we presented a method that strictly decomposes both. This method guarantees that any observed coupling between both subspaces is not due to linear dependence of the respective basis vectors, but a real feature of the allosteric mechanism.

The suggested separation algorithm is not limited to hemoglobin and can be applied to other systems with multiple chains. Also, the algorithm can be used for any definition of domains in the broader sense to separate the motions within the domains and between the domains.

The second component of our method, the PLS-based functional mode analysis, allowed us to find a linear coupling coordinate between both subspaces and thereby identify the allosteric coupling. Applied to hemoglobin, the FMA revealed a remarkable correlation between collective coordinates from quaternary-only and tertiary-only motions. The tertiary-only coupling mode is markedly different from the tertiary structure differences between the known crystallographic R- and T-states. Thus, this mode could not have been derived solely based on the X-ray structures, but yields novel information directly based on transition trajectories between the T- and R-state.

The third part of this work is the interpretation of the identified coupling coordinates. We focused on the protein chain interfaces and detected key interaction residues. We suggest that the allosteric coupling between local and global motions in Hb consists of an interplay of repulsive and attractive interactions at the subunit interfaces in equal measure.

In addition, we were able to lower the T-to-R transition probability by increasing the amount of positively charged histidine residues as a possible molecular explanation of the Bohr effect. Future mutational studies may verify our predicted interactions and consolidate the molecular interaction picture of hemoglobin's allosteric coupling.

3.5 Materials and Methods

3.5.1 MD Simulations & Transition Trajectories

MD Simulations The starting structure for all simulations was the Hb T-state X-ray structure (Fermi et al. [9], PDB id: 2HHB). Each simulation was run with independent starting velocities and had a length of 200 ns. All simulations were carried out using the Gromacs software [30, 31] with the GROMOS 43a2 force field [33]. All simulation parameters used were the same as described before [88] with the exception of the Lennard-Jones

cut-off (rvdw). The key MD parameters are listed in Tab. 3.6. Figure 3.12 pictures a typical Hb simulation system with water molecules and physiological ion concentration. For the simulations a dodecahedral box was used for each periodic image, reducing the number of solvent molecules needed when compared to a rectangular box. Ten simulations used the same rvdw = 1.0 nm as Hub et al. Twenty additional simulations were run using a larger vdW-cutoff of 1.4 nm as consistent with the GROMOS force field parameterization. We could not observe differences in transition probabilities due to rvdw. Finally, 20 simulations using a second set of histidine protonations were performed (see Table 3.5). In total 10 μs of Hb simulations were carried out.

Table 3.5: **Simulation Setup:** Differences in parameters for the individual MD simulations carried out in this study: Lennard-Jones cut-off (rvdw), histidine protonation states (His prot.). Number of total simulations and transition simulations are shown for the different cases. (*) In this case 13 transitions were observed but only 12 trajectories were included in the analysis (see 3.5.1).

rvdw [nm]	His prot.	nr. of sims	nr. of transitions
1.0	Hub et al.	10	5
1.4	Hub et al.	20	13/12(*)
1.4	Kovalevsky et al.	20	4

Table 3.6: **Additional Simulation Parameters:** Parameters used for simulations in this study.

parameter	value	parameter	value
timestep	2 fs	steps	$1{\cdot}10^8$
solvent	SPC	salt	150 mM NaCl
force field	GROMOS43a2	electrostatics	particle-mesh Ewald
temperature	300 K	thermostat	v-rescale (τ=2.5 ps)
pressure	1 bar	barostat	Parrinello-Rahman (τ=5 ps)
# atoms	45946	# waters	13374
# Hb atoms	5730	# Na^+/Cl^-	49/45

T-R transition trajectories To judge whether a T-R transition occurred in the individual simulations, the following criterion was applied: Each simulation was projected onto the difference vector between the T-state and the R-state X-ray structure (Park et al. [10], PDB id: 1IRD; with applied symmetry for the full tetrameric state). If a projection at any time

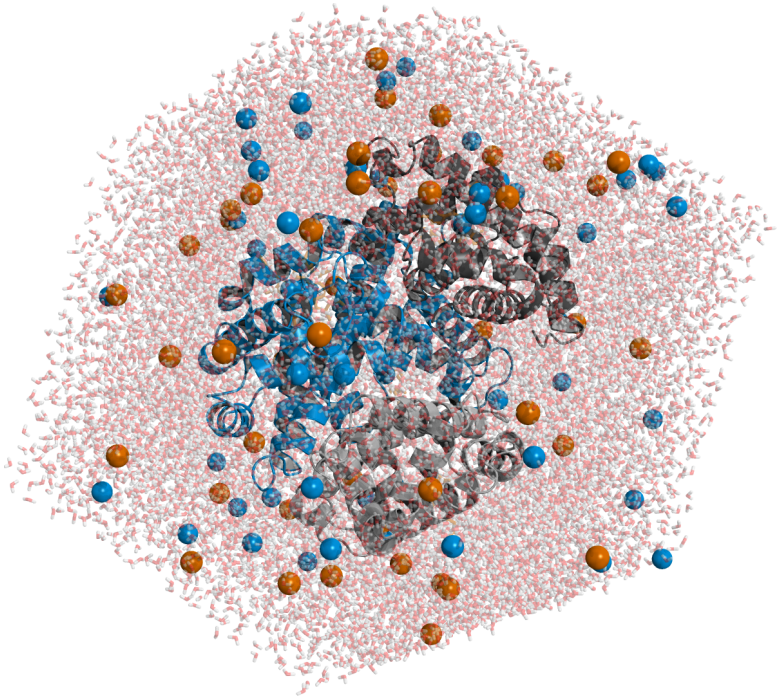

Figure 3.12: **Hemoglobin With Water and Ions in a Dodecahedral Box:** Hb (see Fig. 3.1) in a dodecahedral box filled with water molecules (red-white sticks) with ions in physiological concentration: Na^+ (blue) and Cl^- (orange).

covers 80% of the T-to-R distance, the whole simulation was considered a transition simulation. This gives us the information which simulations are making a transition to the R-state[4], but this does not automatically imply that these simulations reach a structure similar (e.g. in terms of root of mean square deviations) to the R-state. All transition trajectories covered a similar range on the T-R vector whereas one simulation exceeded that strongly. We excluded this outlier, because it would have dominated the global motions in terms of covariance despite its low statistical weight. This left us with 21 transition trajectories and 4.2 μs simulation time in total.

3.5.2 Separation and Coupling of Quaternary and Tertiary Motions

The following method was applied to hemoglobin, but it can be used to analyse other systems with multiple domains as well. Since we were interested in the coupling of local and global motions, we had to define a border that enclosed what we considered local and separated it from the global. Here, we chose the protein chains to each be a local domain. Note that also other choices would have been possible, e.g. grouping Hb into

[4]their motions have one component parallel to the T-R vector

two dimers instead of four monomers. When analysing the motions of a two domain protein consisting of only one chain, it may be of interest to consider each domain as one local entity. The border that we chose also defined the interface that we were analysing, in our case the chain-chain interfaces. This choice defined the local coordinates as tertiary coordinates (as they describe motions within single protein chains) and the global coordinates as the additional coordinates as they form the hemoglobin heterotetramer.

Tertiary-Only Motions: *T* To get the local/tertiary-only (T) coordinates, we superimposed the coordinates of each chain individually onto the respective chain in the T-state X-ray structure. This yielded an artificial structure, with all chains having the same center-of-mass and orientation as in the T-state and displaying only subunit-internal fluctuations. A trajectory processed in this way therefore consists of protein chains which do not move with respect to each other, but only show internal motions. The full MD trajectory has $3N$ degrees of freedom (DOF) with N=4556 being the number of atoms. Since all four subunits are superimposed individually, the DOF for T are reduced by 24 leaving 13644 DOF for T.

Quaternary-Only Motions: *Q* & *cQ* The complementary global/quaternary-only (Q) coordinates we obtained by doing the opposite and superimposing each chain with the respective chain from the T-state structure. Thus each chain was represented by a rigid body and lost all information on internal coordinate changes, only keeping fluctuating positions and orientations of the individual chains. Since the four chains have six translational and rotational DOF each, and the six overall positional and orientational DOF of the system are removed, the Q motions sample 18 DOF. For the coupling analysis, we simplified the Q motions further with a Principal Component Analysis (PCA; see 2.2.1). The resulting eigenvalue spectrum shows a steep decrease among the 18 non-zero eigenvalues with the first eigenvector (referred to as cQ) describing 22% of the total covariance of the Q motions. This principal mode was used for FMA.

The Q and T coordinates together carry the full information of the original trajectory (3N-6 DOF).

Application of FMA: *cT* & *cTew* Functional Mode Analysis (FMA) identifies collective motions maximally correlated to a specific functional property. The original implementation of FMA [55] accomplished a reduction of dimensionality of the underlying space based on PCA. Recently, a new version of FMA based on Partial Least Squares (PLS) has been published [56]. Therein a linear model is constructed based on a given number of latent vectors that are subsequently optimized.

We used FMA based on PLS to maximize the correlation of a given number of latent vectors within T to cQ. We constructed the FMA model on one half of the data first. To estimate the number of suitable latent vectors, we then used the second half for cross-validation to test the predictive power of the model. This resulted in 20 components as the optimal number, with decreasing predictive power for a higher number of components (indicative of overfitting). Applying this knowledge, we constructed the final model with 20 latent vectors on the full data set. The constructed model and the cross-validation are shown in Figure 3.6. The calculated Pearson correlation coefficients were 0.98 for fitting part and 0.83 for the cross-validation part indicating a strong coupling between motions along cQ and cT.

Comparison of Collective Coordinates To estimate how similar the cQ, cT and $cTew$ motions were, we calculated scalar products between the corresponding normalized vectors. In addition, we also compared them to the difference vector of the T- and R-state X-ray structures. To this end, the difference vector was decomposed into a quaternary-only and a tertiary-only part as were the simulations. All mutual scalar products are shown in Table 3.7.

The derived tertiary model cT shares no information (within statistical significance[5]) with the T-R difference vector (orange), which means that our model could not have been derived solely from the T and R X-ray structures. The same holds after application of our separation method to the crystallographic difference vector (T-R tertiary) yielding an overlap of 0.03 (grey). In contrast, it can be seen that on the level of quaternary motions the cQ vector is quite similar to the T-R X-ray difference vector (blue).

3.5.3 Molecular Coupling Mechanism

Van der Waals Overlaps To compute the inter-chain vdW overlap for a single structure we used a modified version of the `dist` program from the CONCOORD software [4]. For each atom we calculated how much its vdW sphere penetrates vdW spheres of atoms from other chains. The sum of vdW overlap values for all atoms gives a length in nanometer, representing a measure for the vdW overlap for each structure.

[5]Scalar products between high-dimensional random vectors are usually small. To determine at what value scalar products are significantly different from random vector scalar products, we computed scalar products between randomly oriented normal vectors in 13668 dimensions. 98% of the scalar products were below or equal to 0.02. In other terms, the probability for two random vectors having a scalar product of more than 0.02 is 2%.

Table 3.7: **Mutual Scalar Products of Specific Collective Coordinates:** Scalar products between the normalized vectors along collective coordinates including *cQ*, *cT*, *cTew*. For a comparison with the X-ray structures we also decomposed the difference vector between the T- and R-state (T-R full) in the same way it was done with the MD trajectories yielding T-R tertiary and T-R quaternary (values referred to in the text are coloured).

	cQ	*cT*	*cTew*	T-R full	T-R tertiary	T-R quaternary
cQ	1.00					
cT	0.01	1.00				
cTew	0.02	0.42	1.00			
T-R full	0.63	0.01	0.03	1.00		
T-R tertiary	0.03	0.03	0.02	0.52	1.00	
T-R quaternary	0.75	0.00	0.01	0.84	0.02	1.00

Hydrogen Bond Analysis Hydrogen bonding energies were estimated applying the Espinosa formula [89]. Here, only inter-chain hydrogen bonds were considered for each structure individually, and for computational efficiency reasons only 5000 equidistant structures were chosen from the ordered trajectory. The Espinosa formula has no repulsive term for small distances of the hydrogen atom to the acceptor, so we scaled down energies stronger than -30 kJ/mol to -30 kJ/mol to avoid counting spuriously low hydrogen bond energies due to artificially small distances. In case of distances too high to form a hydrogen bond, we cut off energies weaker than -5 kJ/mol, and reduced them to 0 kJ/mol. For the figure below in Fig. 3.9 a uniform running average with a window size of 50 was performed.

Contact Analysis In the contact analysis, atoms closer to each other than 3 Å were monitored and defined as contacts. For the 17 frames along *cQ*-*cTew* the contacts were defined individually, resulting in a binary on/off trajectory for every contact. We did not consider residues that were always[6], never, or in an irregular manner in contact along this

[6]These residues are in contact close to T and R, which qualifies for a pulling contact. But since they are also in contact in the artificial states in between, they allow for a motion along *cQ* and *cTew* independently, and hence do not explain the coupling of the motions.

pathway. A visualization of the contact analysis is shown in Fig. 3.10. In the description of the results, we named specific structures as "close to T" and "close to R"; "close to T"/"close to R" here means before/after the transition as defined in paragraph 3.5.1.

3.5.4 Influence of the Histidine Protonation on the T-to-R Transition Rates: Statistical Relevance

We observed T-to-R transitions in 13 out of 20 trajectories with the original histidine protonations [88] and in 4 out of 20 with the protonations measured by Kovalevsky et al. [96]. Since each of the simulations either did or did not show a transition, and all simulations are of equal length, they can each be considered as one event.

Assuming a probability p for a transition, the probability of observing k out of n simulations with a transition is given by the binomial probability distribution

$$F(n;k,p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

While p remains unknown, we can still estimate the probability for values of p, given that 13 positive events out of 20 were measured.

$$W(p) = 21 \cdot F(20;13,p) = 21 \cdot \binom{20}{13} p^{13} (1-p)^7.$$

W has to be normalized for p taking values between 0 and 1. The normalization factor 21 can be obtained using the beta function and its identity for integers: x and y

$$B(x,y) = \int_0^1 t^{x-1} (1-t)^{y-1} \, dt = \frac{(x-1)!\,(y-1)!}{(x+y-1)!}.$$

Each (unknown) underlying probability p from 0 to 1, is weighted with W accounting for 13 positive events observed out of 20. This results in the probability of measuring 4 or fewer events out of 20 with the same probabilities being

$$P(k \leq 4) = \sum_{k=0}^{4} \int_{p=0}^{1} W(p) \, F(20;k,p) \, dp \approx 0.0031.$$

Thus, the difference in the protonation states of the histidines has a significant effect on the transition probabilities with a p-value of 0.0031.

4 Allostery of the ATP-binding in ABCE1

This project is done in collaboration with Dr. Hadas Leonov from the Computational and Biomolecular Dynamics Group at the Max Planck Institute for Biophysical Chemistry, Göttingen. For this thesis, she contributed the Sequence and structure alignment in the ABC family including Fig. 4.3.

For the experimental part, we are collaborating with Prof. Robert Tampé's Cellular Biochemistry group at the Goethe University, Frankfurt.

4.1 Abstract

ABCE1 is a non-transporting member of the ATP binding cassettes (ABC) proteins family. It consists only of two nucleotide binding domains (NBDs), lacking a transmembrane domain found in ABC transporters, and was found to play a role in ribosome recycling. As in other ABC proteins, the NBDs can each hydrolyse ATP. This hydrolysis evokes a conformational change of the NBD from a more compact conformation ("closed") to a more open conformation ("open"). This motion is assumed to be responsible for releasing ABCE1 from aRF1 at the translation termination phase.

The two ATP binding sites appear structurally symmetric, but bear a peculiar asymmetry: Upon mutating one catalytic Glu to Gln in the first NBD, the overall activity decreased to less than half, while mutating the symmetric Glu at the second NBD caused a tenfold increase in activity.

In the following study, first steps towards understanding this unusual observation have been taken. While the structure of the open conformation was recently resolved by X-ray crystallography, a model of the closed conformation is still missing. By applying computational methods such as Molecular Dynamics and Essential Dynamics simulations, we derived a model of the closed structure and suggested mutagenic experiments to our experimental collaborators. These mutations are designed to stabilize the closed conformation, putatively allowing to derive a second high quality structure model with X-ray crystallography, which would be an important step towards characterizing the allosteric mechanism of the functional asymmetry.

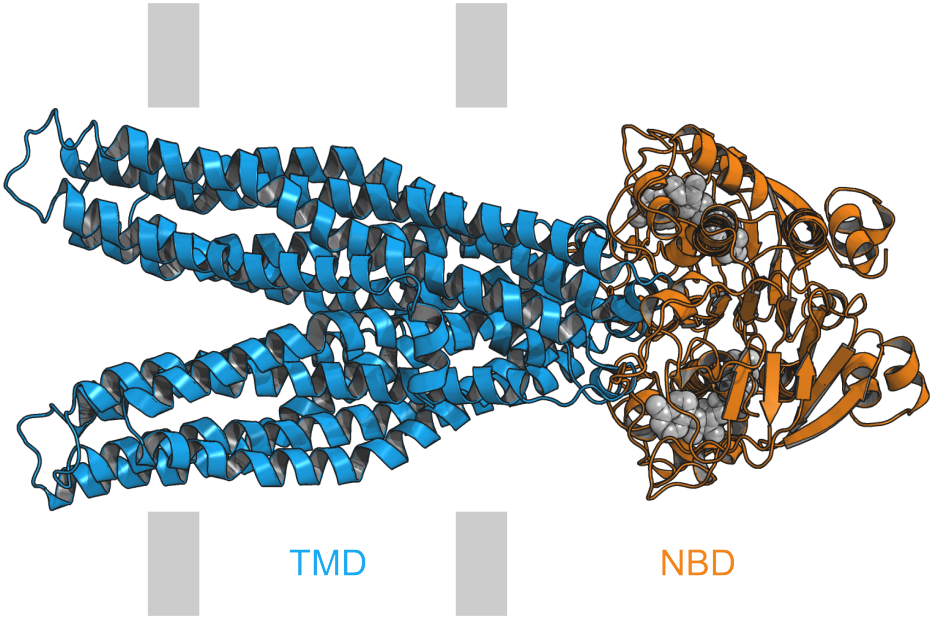

Figure 4.1: **Cartoon Representation of Sav1866:** This member of the ABC family has the typical ABC transporter characteristics: the trans-membrane domain (TMD, blue) with the membrane position indicated by grey bars and the nucleotide binding (NBD) or ATP twin-cassette (ABC) domain (orange). The bound nucleotides are shown in spheres. (PDB id 2ONJ [100])

4.2 Introduction

4.2.1 The ABC Family

ATP binding cassettes (ABC) proteins form a large superfamily of proteins mostly containing transporter proteins. ABC transporters are found in all kingdoms of life where they transport molecules through lipid membranes [98]. In difference to membrane channels, which allow a passive (but specific and mostly regulated) molecule exchange along a concentration gradient, ABC transporters can actively transport cargo against the concentration gradient at the cost of ATP hydrolysis [99]. These transporters consist of an α-helical trans-membrane domain (TMD) and nucleotide binding domains (NBD) for the ATP hydrolysis (see Fig. 4.1 for an example). Following the hydrolysis, the conformation of the NBDs change, and this change is leading to a rearrangement of the TMD, which allow molecules to translocate across the membrane. Hence, ATP hydrolysis in the NBD allosterically regulates the transport process.

ABC transporters are known to cause multi drug resistance in cancer cells, hindering chemotherapy. This is due to overexpression of ABC transporters, which extrude active ingredients used in chemotherapy [13, 14]. A profound understanding of ABC transporters may allow to

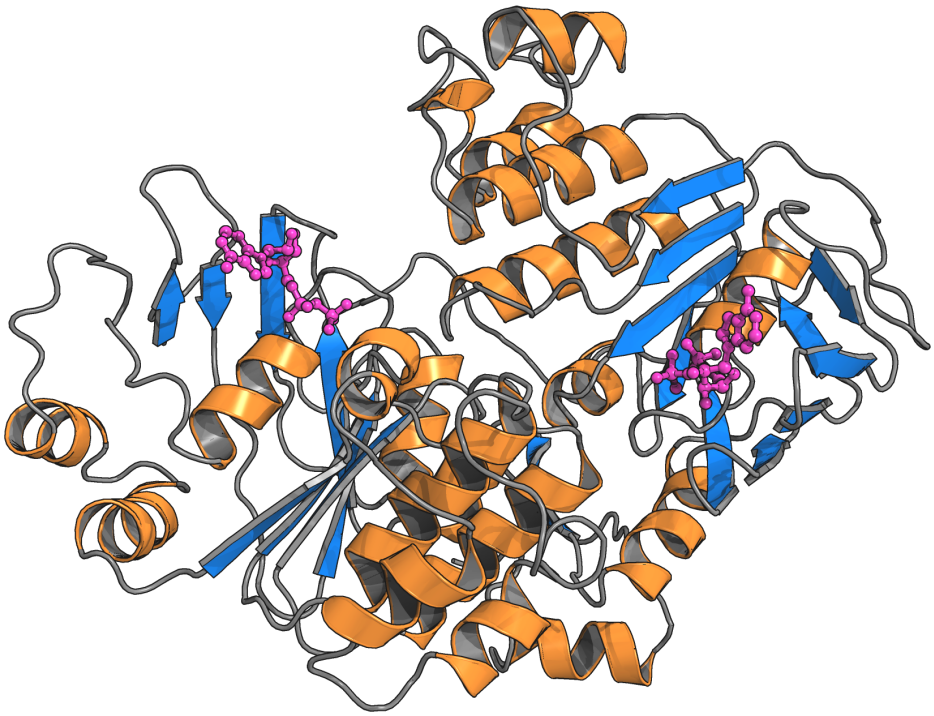

Figure 4.2: **Cartoon Representation of ABCE1:** This shows the ABCE1 X-ray structure from Barthelme et al. (PDB id 3OZX). The bound nucleotides are shown in magenta.

inhibit the transport to overcome this limitation in cancer treatment.

4.2.2 ABCE1

The protein ABCE1 is one of the few members of the ABC family that does not contain a TMD but only two NBDs. In addition, it contains an N-terminal iron-sulfur-cluster (FeS) domain. Originally, ABCE1 was discovered to act as a ribonuclease L inhibitor (RLI) [101], but later a much broader functional spectrum emerged: It was found to be critical for HIV-1 capsid assembly[1] [102], and during protein synthesis to interact with elongation, initiation [103] and release factors [104]. In 2011, Barthelme et al. published an X-ray crystal structure of ABCE1 of sulfolubus solfataricus (PDB id 3OZX; see Fig 4.2) [15]. The structure was crystalized without the FeS domain. It showed two similar ATP binding sites with a remarkable difference in function: stopping ATP hydrolysis by mutation of the catalytic Glu238 to Gln in the first NBD decreased the overall activity to 30-50%, while mutation of the symmetric Glu485 to Gln increased the activity by a factor of ten. As expected, simultaneous mutation in both sites renders the protein inactive [15].

[1] under the name HP68

4.2.3 Scope of This Study

In the present work, we want to investigate the reasons for the functional difference between the two ATP binding site despite their structural symmetry. The hydrolysis is predicted to be associated with a clamp-like conformational change. Our goal is to understand how both domains interact during the transition. At the beginning of the study, only a structure of the ADP-bound ("open") conformation was available. It is assumed, that an ATP-bound structure would adopt a "closed" conformation [15], but a structure model of it is still missing. To thoroughly analyse the conformational dynamics between both, high quality structure models of both states are necessary.

To this end, we applied computational methods to investigate mutations that could stabilize a closed conformation. In addition, we extended the 3OZX structure by adding the FeS domain according to the Pyrococcus abyssi X-ray structure by Karcher et al. ([105], PDB id: 3BK7), to investigate the role of the FeS cluster domain.

4.3 Results

It is hypothesised that ABCE1 undergoes a conformational change from "open" to "closed" upon ATP binding. The ADP-bound structure by Barthelme et al. represents the open conformation. The structure of the NBD domain consists of two homologous subunits expressed on the same protein chain. Following the residue numeration of 3OZX, the first subunit is from 76 to 329 (NBD one) and the second subunit from 340 to 599 (NBD two)[2]. This definition was used later to define contacts between the two domains.

In this work, we were applying computational methods to aid the identification of a closed ABCE1 conformation. First, we identified required structural features common to "closed" conformations of NBDs of other ABC proteins. Secondly, applying Essential Dynamics (ED) and Molecular Dynamics simulations, we drove the open ABCE1 structure to fulfil the structural features. As the last step, a contact analysis that focuses on differences in both conformations was applied to identify key residues. Suitable mutations of these residues may be used to stabilize the protein in a closed state.

4.3.1 Sequence And Structure Alignment in the ABC Family

The first step towards a closed structure model of ABCE1 was to get as much knowledge as we can have from existing structures of open and closed NBDs of ABC proteins. We combined ABC structures similar in

[2] The ten residues in between are in loop regions, not contained in the PDB. We included the residues according to experimental insight from our collaborators.

structure (using the DALI protein server [106]) in a core coordinate set of the NBD. From ABC proteins structures were selected to obtain maximal similarity between them, while keeping the number of common residues for each high. The resulting consensus set (called CS in the following) contained structures of 26 ABC proteins with a subset of 324 common amino acids each.

4.3.2 PCA of Similar ABC Structures

To identify different conformations within the CS, we applied a principal component analysis (PCA, see section 2.2.1). The first two eigenvectors represent a lateral and a cone shaped opening/closing motion. The projections on these motions show clustering of all ATP-bound structures and with one exception clustering of ADP-bound structures (see Fig. 4.3).

Since all ATP-bound structures occupy a single region in the first two PCA dimensions, we assume that the ATP-bound ABCE1 adopts a similar conformation and can also be found in that region (marked red in Fig. 4.3). This region is characterized by being maximally closed in the lateral and the cone shaped motion, explaining the denotation "closed". To obtain a structural model of the ATP-bound ABCE1, the next step was to drive the Barthelme et al. structure 3OZX in direction of the ATP-bound structures. Thereby we aimed at an ABCE1 structure with the characteristics of the closed conformation. This was done in the space given by the first PCA coordinates, while not restraining the vast majority of degrees of freedom.

4.3.3 Essential Dynamics Simulations

In order to drive the consensus set coordinates of the protein towards the closed target structure, we used essential dynamics simulations (ED; see Sec. 2.3). In the space spanned by the first five eigenvectors[3] of the PCA on the consensus set structures, we drove the simulations towards the target using the radial contraction (`radcon`) option of the ED implementation in GROMACS. To get a first idea how this works, see Fig. 4.4. Further details on the simulations are given in section 4.5.

In the PCA subspace, the ED simulations quickly reached the ATP-bound conformations (see Fig. 4.4) in $\sim$200 ps. From unrestrained MD simulations starting from the open conformation, we would estimate the actual timescale for this transition to be larger than 200 ns. For that reason, additional simulations are required after reaching the ATP-bound structures, to further equilibrate the degrees of freedom orthogonal to the PCA space.

[3] For the sake of clarity in illustrations only the first two eigenvectors are used.

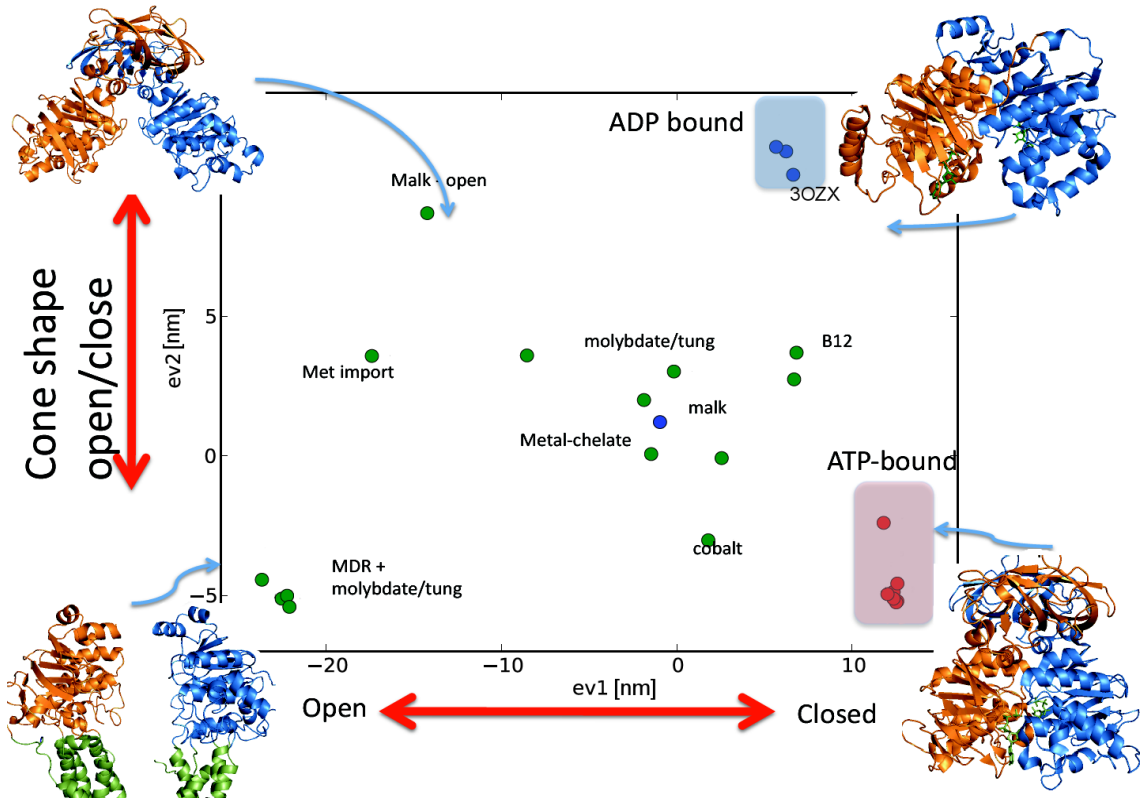

Figure 4.3: **PCA on the Consensus Set of ABC Proteins:** The projection of the consensus set structures on the first two PCA eigenvectors is shown. The structures are coloured corresponding to the bound ligand: ATP-bound (red), ADP-bound (blue) and empty (green). In the corners, structures in Cartesian coordinates represent the motions. The Barthelme et al. structure 3OZX is marked.

4.3.4 Unrestrained MD Starting from The Closed Model

From the ED simulations, we excluded the first 20 ns for equilibration and from the remainder, we chose the structure closest to the target as a model for the closed conformation. This was done separately for the ADP-, ATP-bound and empty protein in order to estimate the effect of the ligand in MD. Starting from the calculated models MD simulations without the ED restraints were run to further validate the model by allowing relaxation of all degrees of freedom. Figure 4.5 A shows the simulation projected onto the first two PCA eigenvectors of the CS. The simulations starting from the open conformation explore similar regions in the PCA space, with a tendency to evolve in direction of the closed conformation. Also, the simulations starting from the closed conformation evolve towards the open conformation, but do not behave as similarly as in the open case. The largest difference between the individual simulations was observed in the case of empty binding sites and starting from the closed conformation

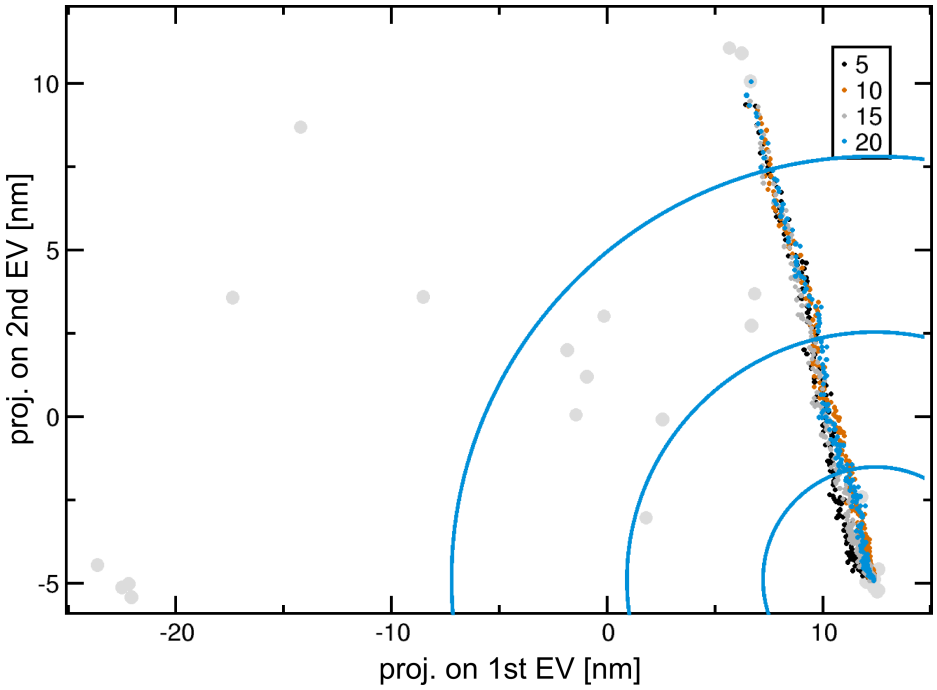

Figure 4.4: **ED Radial Contraction Simulations:** This illustrates how radial contraction works. The simulations are evolving towards the center of the circles, where the target is defined. The four simulations differ in the number of dimensions included into the ED subspace, but in this two-dimensional projection they behave similarly.

(see Fig. 4.5 B). Detailed information about the MD simulations can be found in Sec. 4.5.

4.3.5 Contact Analysis

In order to stabilize the protein in the closed conformation by mutations, we need to know the characteristical differences in residue interactions between the open and the closed state. Therefore, a contact analysis was performed to identify residues in contact across the NBD-NBD interface.

As structure ensembles for the open and closed state we used the ADP-bound simulations starting from the open conformation and the ATP-bound simulations starting from the closed conformation respectively. Two residues were registered as a contact, if at least 10% of the time two atoms of these residues are closer than 3 Å. In figure 4.6 an illustration of the results is shown.

Since it was our goal to stabilize the closed state with suitable mutations, we only focused on contacts that are present in the closed state but not in the open state. Being specific for the closed state, these contacts may give hints for promising mutation sites.

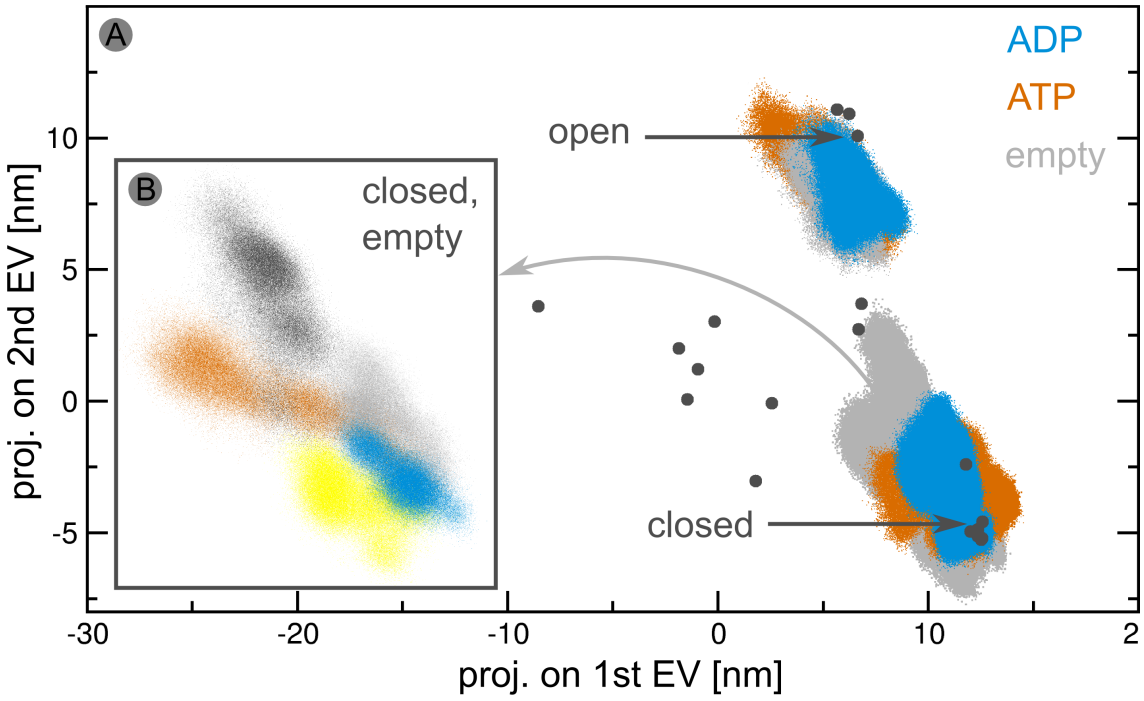

Figure 4.5: **Projection of Unrestrained MD Simulations on the Consensus Set:** (A) Projections of MD simulations starting from open structure (3OZX) and closed model are shown. The colours indicate the type of ligand bound in both binding sites. (B) Close up view of the five individual simulations (different colour) starting from the closed conformation with no ligand bound are shown.

4.3.6 Mutation Suggestions

To identify mutations that may stabilize the closed conformation, we applied following criteria on the full list of contacts: First, the two residues were much more frequently in contact in the closed conformation compared to the open state[4]. This was important, since we aimed at stabilizing the closed state over the open state. Second, (as mentioned before) the residues need to interact frequently. Third, none of the residues should be in direct proximity to the ATP binding site. This is necessary to reduce the probability that the mutation directly affects the ATP hydrolysis.

A possible mechanism to create a strong bonded interaction between two residues is to mutate both to cysteines so that a disulfide bond can form. In addition, other crosslinkers with different length can be introduced, or the sulfur atoms can be linked by mercury with $HgCl_2$.

We looked closer into the contact pairs and suggested the following that emerged as promising mutation candidates:

1. Asn204 – Glu538: Cys-Cys crosslink suggested; alternatively, the

[4] or not existing in the open state

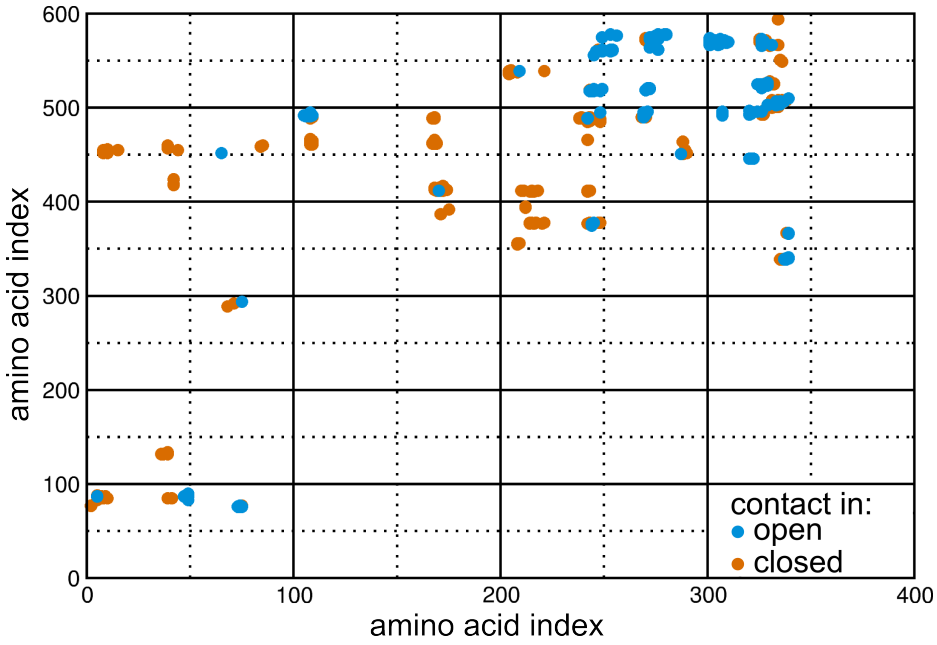

Figure 4.6: **Contacting Residues For Open And Closed State:** Only contacts present at least 10% of the time are included in the analysis.

mutation Asn204→Asp may form a salt-bridge to Glu538

2. Ser461 – Asn108: Cys-Cys crosslink suggested
3. Gln167, Tyr168, Gly462, Tyr489: hydrophobic cluster suggested, e.g. Gln167 → Phe or Tyr
4. Lys84 – Asp459: Cys-Cys crosslink with a longer linker suggested
5. Glu454 – Lys43: Lys43 is part of the FeS-cluster domain, a longer crosslink is required

At the moment, first mutation experiments had been carried out for the Ser461 – Asn108 contact. Between the residues different crosslinks were established, but further optimization on the experimental conditions is needed. Complete results on first mutations are expected within few months.

4.4 Discussion & Conclusion

In this project, collective coordinates were defined to drive the transition between the open and a putative closed conformation of ABCE1. The close proximity of all ATP-bound structures in the PCA projection, rendered us confident to also find the ATP-bound closed ABCE1 there. With the `radcon` algorithm, the transition towards the closed conformation

happened at least three orders of magnitude faster than in normal MD. Such structures have to be evaluated carefully, since they may be far away from equilibrium. For that reason, even after quickly reaching the closed conformation within the PCA space, we continued the simulations first with the ED restraints and after that unrestrained. In Fig. 4.7, the deviation from the target structure shows how the system "relaxes" after the fast conformational change.

The validation of the closed state model with unrestrained simulations showed, that the ABCE1 adopted to the new conformation. As seen in Fig. 4.5, the PCA projection of the simulations starting from the closed model occupy a similar area compared to the open state simulations, yielding a similar flexibility in the first two eigenvectors and thus pointing at the stability of the model.

The ensembles starting from both states move more towards each other than orthogonally. This might point towards an energetically favourable path connecting the two. In future computational studies, Umbrella Sampling [107,108] may be applied to connect both ensembles dynamically and to calculate free energy differences along this path.

Even though we were not yet able to work on the actual allosteric interactions between the two ABCE1 domains, obtaining a high resolution X-ray structure would be a big step in that direction. The mutations we have suggested should aid to advance towards this goal.

It was shown, that the FeS-cluster domain is essential for the interaction with the ribosome [15]. In our simulations, the FeS-cluster domain showed relatively high mobility, which may aid to establish a first contact of ABCE1 with the ribosomal 30S subunit. We observed a contact between the FeS-cluster domain and the second NBD (fifth mutation suggestion), which may affect the internal dynamics of ABCE1 and thereby the allosteric coupling between the two NBDs. Additional simulations are needed to further investigate the effect of the FeS-cluster on the dynamics of ABCE1. The large size of the ribosome, would still render it difficult to obtain MD simulations of ABCE1 in contact with the ribosome. But with growing computer power it may be possible to explore this important interaction soon and to gain deeper insight in the function of ABCE1.

4.5 Materials & Methods

4.5.1 PCA

For the PCA (see Sec. 2.2.1) a subset of residues from 3OZX was considered. The residues are given in section 7.3. The first two eigenvectors of the PCA on the CS structures already covered 81% of the total variance.

Table 4.1: **ABCE1 Simulations**

	ED	non-ED	
ligand		open	closed
ADP	5 ×100 ns	5 ×200 ns	5 ×200 ns
ATP	5 ×100 ns	5 ×200 ns	5 ×200 ns
empty	5 ×100 ns	5 ×200 ns	5 ×200 ns

Table 4.2: **MD Simulation Parameters:** Parameters used for simulations in this study. The atom numbers are taken from the ATP-bound simulations of the closed model; the exact numbers may vary slightly for the other setups.

parameter	value	parameter	value
timestep	4 fs	steps	$5{\cdot}10^7$
solvent	SPC	salt	150 mM NaCl
force field	AMBER99sb-ildn	electrostatics	particle-mesh Ewald
temperature	310 K	thermostat	v-rescale (τ=0.1 ps)
pressure	1 bar	barostat	Berendsen (τ=1 ps)
# atoms	88674	# waters	25982
# ABCE1 atoms	10474	# Na^+/Cl^-	85/81

4.5.2 MD Simulations

For the essential dynamics simulations as well as the subsequent MD simulations without the ED restraints, the main parameters are summarized in table 4.2. Virtual sites (see Sec. 2.1) have been applied to double the timestep to 4 fs. For the three binding site states (ADP, ATP, empty) ED simulations were carried out. Starting from the open structure and the closed ED model, unrestrained MD simulations were executed for ∼200 ns.

To obtain independent sampling and thereby increasing statistical significance, every simulation (including ED and non-ED) was carried out five times. Table 4.1 lists the simulations. In total, 7.5 μs of ABCE1 simulations were obtained.

Essential Dynamics Simulations The ED simulations (see Sec. 2.3) applied in this analysis used the radial contraction (`radcon`) option of the ED implementation in GROMACS. For a given collective subspace and the coordinates of a target in that subspace, this algorithm forces the simulation steps to reduce the distance to the target in the ED subspace. In two dimensions this would look like shown in Fig. 4.4: The simulations may move to a smaller circle around the target, but if the simulation

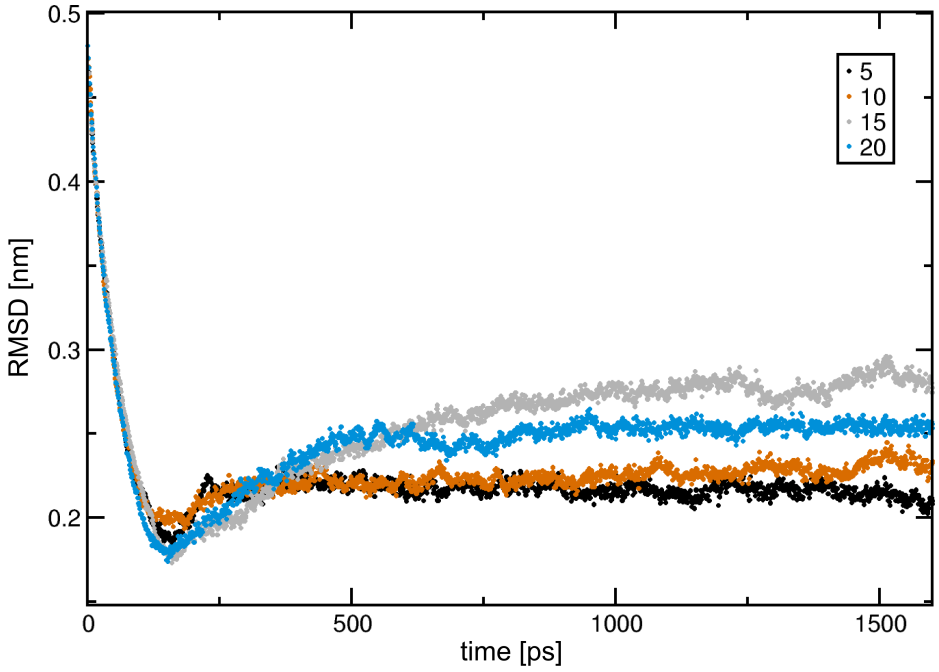

Figure 4.7: **RMSD From Target Structure:** For ED simulations with a different number of included PCA components, the distance from the target structure was measured.

proceeds to a larger radius, it will be altered such that it stays on the same hypersphere surface.

To estimate the effect of the number of included PCA components, we compared short simulations that use the 5, 10, 15 or 20 first eigenvectors. The convergence is on the same timescale as can be seen in Fig. 4.8. That is why we decided to run the extended simulations with the lowest number of additional restraints, namely five components. While the latter is measured within the ED subspace only, the root of mean square deviation (RMSD) includes all degrees of freedom. The RMSD from the target structure (see Fig. 4.7), shows a rapid decrease in RMSD due to a fast approaching to the target. The DOF not included in the ED subspace are relaxing to adopt the new conformation, as can be seen in the slight increase in RMSD after reaching a first minimum. This is the reason why we excluded the first 20 ns from the analysis for not being at equilibrium. The final ED simulations were run for ∼100 ns each.

4.5.3 Parameterization of the FeS-Cluster

The ABCE1 protein contains a so-called FeS-cluster domain itself incorporating two $(FeS)_4$-clusters. Figure 4.9 shows the two FeS-clusters within the domain, and a closer look onto one, with the cysteine residues coordinating it. The FeS-cluster domain was shown to be critical for the function of ABCE1 [15]. To include in in MD simulations, force field parameters

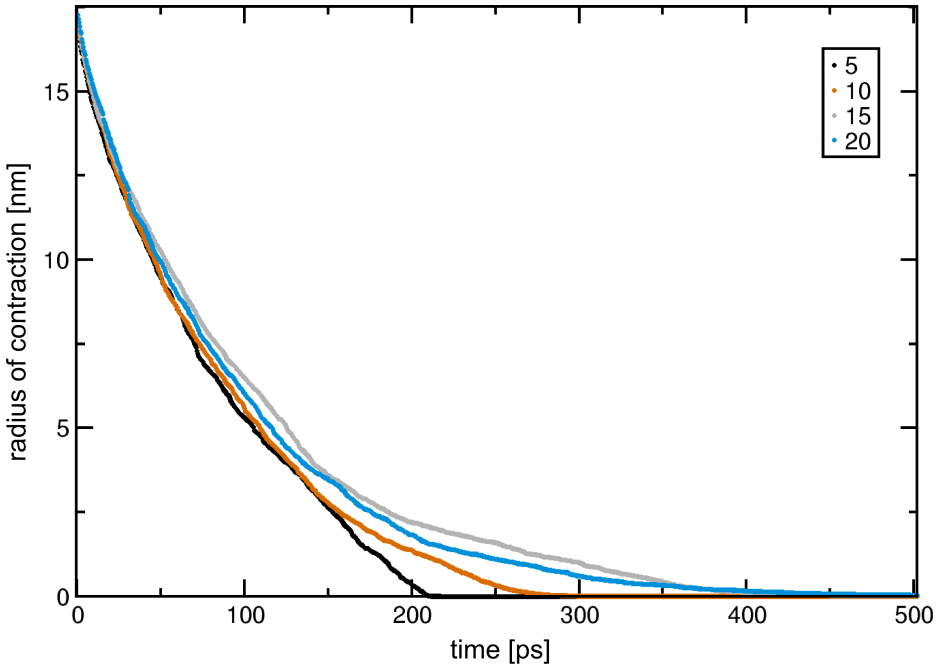

Figure 4.8: **Radius of Contraction:** For ED simulations with a different number of included PCA components, the distance from the target structure was measured. In contrast to an RMSD (see Fig. 4.7), this distance is measured in the ED subspace.

needed to be derived.

Banci et al. derived parameters for the FeS cluster and the coordinating cysteine residues for the AMBER force field from 1984 [109] with quantum mechanical calculations [110].

We adjusted the parameters such that they match the new AMBER99sb force field that was used in this project. Therefore, we changed the charges to yield the correct total charge of the backbone, and moved charges from the electron lone pairs (present in the older AMBER) to the cysteine sulfur atoms. The parameters for the bond lengths and angles for the FeS cluster and the coordinating cysteines were taken from the X-ray structure from Karcher [105]. Throughout all simulations, the FeS-cluster domain showed little internal mobility and mostly tumbled on top of the first NBD. The FeS-clusters themselves remained stable at all times. The final parameters used in this study are listed in Sec. 7.4.

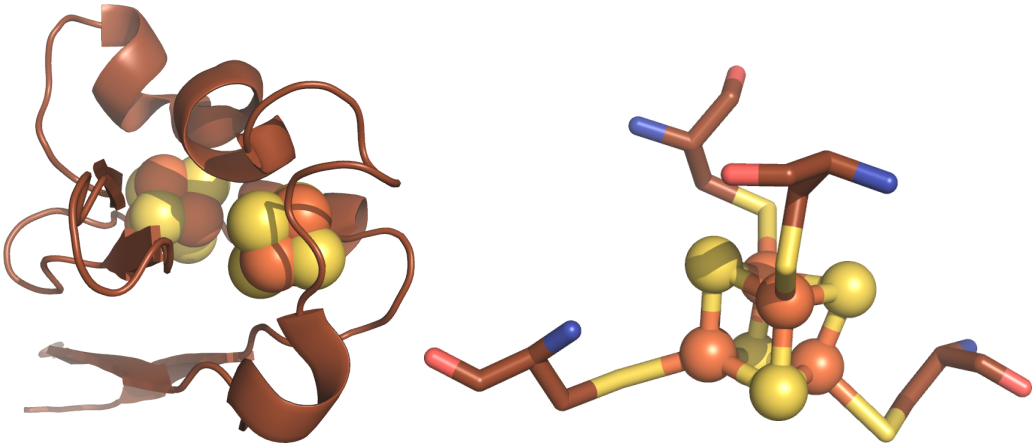

Figure 4.9: **Structural Details for the FeS-Cluster domain.** On the left, a cartoon representation and the localization of the two FeS clusters is shown. On the right, a close-up of a single FeS cluster (4Fe-4S) is depicted. Shown as spheres are the iron atoms (orange) and the sulfur atoms (yellow). The cluster is coordinated by four cysteine residues (shown as sticks).

5 Summary and Conclusions

Allostery is a universal principle in proteins that allows for powerful regulation that cannot be realized with orthosteric regulation. The number of identified allosteric interactions is increasing, but until today, there is no general answer to how allostery is manifested on an atomic level. In order to allosterically affect the binding affinity of one binding site through binding in a distant site, information has to pass between the sites. A possible means to connect local changes in two allosterically coupled binding sites is a conformational change.

In this thesis, we studied allosteric interactions in two proteins: hemoglobin and ABCE1. In both cases, we applied Molecular Dynamics simulations as the core method to obtain dynamical information. Rather than focusing on the fluctuations of the individual atoms, we looked at collective motions of the proteins. By relating collective dynamics to specific functions or other motions, we were able to identify the molecular constituents of the underlying mechanism.

Hemoglobin In the case of hemoglobin, it is known that a conformational change is required for the distant binding sites to interact. During such a conformational change, all atoms need to move collectively. We investigated the collective dynamics of hemoglobin during the conformational transitions that were observed with Molecular Dynamics simulations. We showed how different levels of collective dynamics – local and global motions – together form hemoglobin's allosteric mechanism. From that we were able to provide molecular insight into this mechanism that is crucial to the effectiveness of our respiratory system: The quaternary conformational change induces unfavourable[1] contacts that are accommodated by tertiary rearrangements.

ABCE1 The protein ABCE1 is part of one of the largest protein families, and understanding its function and malfunction is key in understanding many diseases on a molecular level. While most members of the ABC family are transporters with a transmembrane domain, ABCE1 only contains two nucleotide binding domains (NBDs) characteristic for ABC proteins. ABCE1 shows an allosteric interaction between the two NBDs: Stopping ATP hydrolysis in one NBD reduces the overall hydrolysis rate, while stopping hydrolysis in the other NBD increases the overall hydrolysis rate. How regulation is realized on a molecular level is still not understood.

[1] unfavourable through steric repulsions or missing hydrogen bonds

The two NBDs undergo a conformational change upon ATP hydrolysis, but only the "open" structure of this motion is resolved and the "closed" structure is still missing. Applying knowledge of collective protein motions in ABC proteins, we were able to characterize key interactions of a putative closed state. Experimental studies based on suggested mutations derived from the modeled closed state may soon result in a closed crystal structure, thereby providing an important next step to resolve the mechanism of allosteric coupling between the NBDs in ABCE1.

Outlook Collective motions are important for protein functions and describing conformational changes in suitable collective coordinates is key for describing large scale protein dynamics. The method we developed to separate local and global motions can readily be applied to other systems as well. Any multi-domain protein can be a suitable target. But the method may also yield interesting insight in protein aggregation, where local and global motions get in contact and the interface is directly given.

Allosteric regulation is important in pharmaceutical research. Understanding allostery on a molecular level may hence not only yield a deeper understanding of this phenomenon itself, but directly introduce new means to effect proteins with drugs. Profound knowledge of a number of allosteric interactions will show if there are common interaction motifs in proteins, or if each protein has its own characteristic collective dynamics underlying the allosteric coupling.

6 Acknowledgements

First and foremost, I thank my supervisor Prof. Bert de Groot for introducing me to computational biophysics and giving me the opportunity to do my doctorate studies in his group. He always – I cannot stress this enough – took the time for fruitful discussions and questions.

I would like to thank my other two thesis committee members Prof. Ralf Ficner and Prof. Marcus Müller, for helpful discussions. For their tireless organization of the International Max Planck Research School "Physics of Biological and Complex Systems" (IMPRS-pbcs), I am grateful to Antje Erdmann, Michaela Böttcher, and Frauke Bergmann. This thesis was financially supported by the IMPRS-pbcs.

I thank my parents for supporting me where they could, allowing me to focus on my work.

Also, I would like to thank the whole Department of Computational and Theoretical Biophysics. Especially Prof. Helmut Grubmüller for leading a department, in whose atmosphere of communication and discussion science feels alive. Especially, I thank Christian Blau, Rodolfo Briones, Carl Burmeister, Stephanus Fengler, Lars Bock, Vytautas Gapsys, Martin Höfling, and Martin Stumpe, Dirk Matthes and Camilo Aponte-Santamaría. Our computer experts Carsten Kutzner, Martin Fechner, and Ansgar Esztermann kept my workstation, the cluster and software working at all times: thank you. I would like to thank my colleagues Nicole Doelker and Hadas Leonov for the pleasant working atmosphere.

Jana, during my doctorate studies, I was able to gratefully enjoy your transition from my girlfriend to my wife. Thank you so much for your support!

7 Appendix

7.1 Overfitting and Cross-Validation

In the following paragraphs the notion of overfitting will be explained on a simple example and how cross-validation can be used to avoid it.

The relations between stochastic observables[1] can be explored by constructing models for these relations. The stochasticity makes the observations fuzzy and their relation noisy. Insufficient information about the probability distribution of the observables can make the search for a model challenging. A good model distinguishes which part of the data is due to the actual relation between the observables and which is due to the noise of the data.

For an examined model relation, free model parameters are aligned so that the model fits the data, the so called *fitting* process. *Overfitting* is the phenomenon that occurs if a model with too many parameters is fitted to the observations; the model is not only aligning to the coarse features of the data, but it starts fitting the noise.

Imagine a set of two-dimensional data points scattered with some noise around a straight line. Now, for the sake of the argument, forget about the underlying line and ask the question what is a suitable model for the x-y dependence of the data[2].

Let us construct two models with the aim of describing the data dependence: (1) a linear model, and (2) a high-order polynomial model. In Fig. 7.1 to the grey data points a linear (blue) and a polynomial model (orange) are fitted. The straight line does not capture every single detail, but it follows the global trend of the data set, whereas the polynomial visits the individual points[3]. From this perspective, the polynomial fit appears to represent the better description of the data.

Since a good model is not only able to describe already measured data but also predicts future measurements, the quality of a model should also contain its predictive power. For this a principle called *cross-validation* has been established. It can be achieved by removing a small random part (e.g. 5-10%) of the original data – the cross-validation set – and using the remaining 90-95% – the model fitting set – for the fitting of the model. In Fig. 7.1 the black dot represents the cross-validation set

[1] observables whose values are subject to randomness

[2] As the a straight line was used to construct the data, so it is an obvious candidate for a good model.

[3] For N data points a polynomial of the order $N-1$ can be constructed such that it crosses every single point perfectly. This would create a model that is as complex as the data itself.

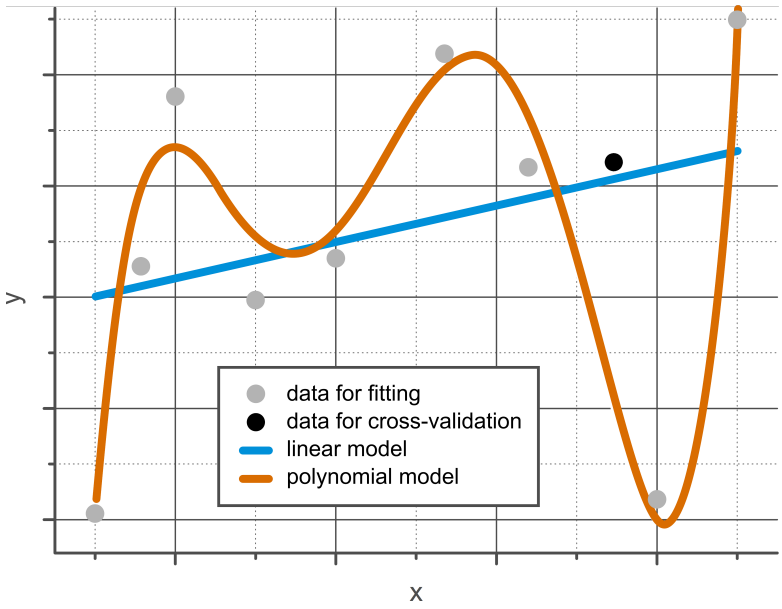

Figure 7.1: **Example of Overfitting:** To the data points (grey dots) two models are fitted: a linear model (blue) and a polynomial model (orange). One point (black dot) was excluded from the fitting and used for cross-validation (see text).

and the remaining grey dots the model fitting set. For both sets the deviation of the model from the original data is calculated (e.g. with the Pearson correlation coefficient): The deviation from the model fitting set measures the goodness of the individual fit; the deviation from the cross-validation set measures how well a model describes actual data not included into the fit, i.e. predicting them. The polynomial model performs better on the model fitting set, whereas the linear model performs better on the cross-validation set. This indicates that in contrast to the linear model the polynomial model is overfitted since it fails at predicting the cross-validation set.

This example was constructed to show the drastic effect of overfitting on model construction and validation. In real-life applications overfitting may not so easily be revealed. For example, in X-ray crystallography a structural (atomistic) model is fitted to the measured X-ray reflections from the crystal. The goodness of this fit is (amongst other measures) described by the R-factor that compares the actual reflections with the reflections calculated from the model. For protein crystals of intermediate resolution the number of measured data and the number of atom coordinates is often in the same order of magnitude, which makes overfitting an imminent risk.

To this end, Brünger suggested to expand the usage of R-factor to the so-called free R-factor R_{free}. It is calculated in the same way as the R-factor but on a randomly chosen set of reflections that was not used in

the construction of the structure model. Thereby R_{free} cross-validates the structure and identifies overfitting.

7.2 Used Software

For all MD simulations on the in-house computer cluster the GROMACS software package version 4.53 was used [30,31]. This includes the ED simulations and most of the analysis tools. For further calculations MATLAB R2011a (Mathworks, Inc.) and `bash`, `awk` and `sed` were used.

Pictures of protein structures were rendered in `PyMol` [111]. Schemes were drawn with `inkscape` and `gimp`. This thesis was typesetted with LaTeX 2_ε.

7.3 List Of Consensus Residues

The residues from 3OZX forming the consensus set are:
subunit A:
Ile80-Arg82, Phe88-Asn98, Thr100-Phe127, Lys153-Glu154, Lys161-Lys165, Gln167-Tyr168, Gly178-Ile187, Lys192-Ala210, Ile212-Gln238, Ser240-Tyr276, Thr278-Gly286, Gly292-Ser295, Ser297-Ala299, Arg301-Gly303
subunit B:
Met344-Trp346, Lys348-Lys352, Leu358-Val359, Asp361, Gly363-Glu367, Glu369-Ala394, Glu396-Thr400, Ile405-Pro410, Ile413-Pro415, Tyr417-Ala428, Lys430, Asn448-Val457, Asp459-Gly462, Glu464-Gln485, Ser487-Arg507, Ala511-Asp524, Ala527-Gly535, Leu542-Thr544, Val547-Thr551, Met553-Asn554, Phe556, Arg558

7.4 FeS-Cluster Parameters

Following, the parameters are given that can be used to include $(FeS)_4$-clusters in MD simulations with GROMACS.

NB: To adopt the angle potential from Banci et. al to the GROMACS software the harmonic force constants had to be multiplied by two, since in AMBER's potential function a factor 0.5 is missing when compared with GROMACS. The simulations were run with the LINCS algorithm to maintain the bond-lengths, the denoted bond energies are arbitrary and will not be used.

aminoacids.rtp

```
[ SF4 ]
[ atoms ]
FE1     FS      1.465       1
FE2     FS      1.465       2
FE3     FS      1.465       3
FE4     FS      1.465       4
S1      S       -0.915      5
S2      S       -0.915      6
S3      S       -0.915      7
S4      S       -0.915      8
[ bonds ]
FE1     S2      0.2298      100000
FE1     S3      0.2316      100000
FE1     S4      0.2285      100000
FE2     S1      0.2323      100000
FE2     S3      0.2247      100000
FE2     S4      0.2183      100000
FE3     S1      0.2306      100000
FE3     S2      0.2228      100000
FE3     S4      0.2230      100000
FE4     S1      0.2138      100000
FE4     S2      0.2143      100000
FE4     S3      0.2103      100000
[ angles ]
S2      FE1     S3          92.970      460.2
S2      FE1     S4          103.009     460.2
S2      FE3     S1          100.043     460.2
S2      FE3     S4          107.137     460.2
S2      FE4     S3          104.018     460.2
S2      FE4     S1          108.531     460.2
S3      FE1     S4          102.404     460.2
S3      FE2     S1          95.786      460.2
S3      FE2     S4          108.102     460.2
S3      FE4     S1          106.155     460.2
S4      FE2     S1          101.824     460.2
S4      FE3     S1          100.939     460.2
FE2     S1      FE3         73.974      460.2
FE2     S1      FE4         75.172      460.2
FE2     S3      FE1         73.534      460.2
FE2     S3      FE4         77.488      460.2
FE2     S4      FE3         78.249      460.2
FE2     S4      FE1         75.356      460.2
FE3     S1      FE4         73.083      460.2
FE3     S2      FE1         74.155      460.2
FE3     S2      FE4         74.565      460.2
FE3     S4      FE1         74.399      460.2
FE4     S2      FE1         78.152      460.2
FE4     S3      FE1         78.532      460.2
[ CYF]
[ atoms ]
N       N       -0.41570    1
H       H       0.27190     2
CA      CT      0.02130     3
HA      H1      0.11240     4
CB      CT      -0.33740    5
HB1     H1      0.00405     6
HB2     H1      0.00405     7
SG      SH      -0.49000    8
C       C       0.59730     9
O       O       -0.56790    10
[ bonds ]
N       H
N       CA
CA      HA
CA      CB
CA      C
CB      HB1
CB      HB2
CB      SG
C       O
-C      N
[ impropers ]
-C      CA      N           H
CA      +N      C           O
```

aminoacids.hdb

```
CYF     3
1       1       H       N       -C      CA
1       5       HA      CA      N       CB      C
2       6       HB      CB      CA      SG
```

ffbonded.itp

```
[ bondtypes ]
FS      SH      1       0.226       120000.0
[ angletypes ]
SH      FS      S       1       180.000     0.000
CT      SH      FS      1       180.000     0.000
[ dihedraltypes ]
CT      SH      FS      S       9       180.0     0.00000     1
FS      S       FS      X       9       180.0     0.00000     1
```

atomtypes.atp

```
FS      55.00000
```

Bibliography

[1] F.L. Hünefeld. *Der Chemismus in der thierischen Organisation*. Brockhaus, 1840.

[2] Thomas E. Creighton. *Proteins: Structures and Molecular Properties*. WH Freeman, 1993.

[3] M. D. Naresh, V. Subramanian, S.M. Jaimohan, A. Rajaram, V. Arumugam, R. Usha, and A. B. Mandal. Crystal structure analysis of hen egg white lysozyme grown by capillary method. unpublished, Deposition: 29.03.2007, Release 17.04.2007.

[4] B. L. de Groot, D. M. F. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. C. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins-Structure Function And Genetics*, 29(2):240–251, 1997.

[5] R.H. Erickson and Y.S. Kim. Digestion and absorption of dietary protein. *Annual review of medicine*, 41(1):133–139, 1990.

[6] J.B. Wittenberg and B.A. Wittenberg. Myoglobin function reassessed. *Journal of experimental biology*, 206(12):2011–2020, 2003.

[7] Vijayalakshmi Santhakumar, Martin Wallner, and Thomas S. Otis. Ethanol acts directly on extrasynaptic subtypes of gabaa receptors to increase tonic inhibition. *Alcohol*, 41(3):211 – 221, 2007.

[8] G.A.R. Johnston. Gabaa receptor pharmacology. *Pharmacology & Therapeutics*, 69(3):173 – 198, 1996.

[9] G. Fermi, MF Perutz, B. Shaanan, and R. Fourme. The crystal structure of human deoxyhaemoglobin at 1.74 å resolution. *Journal of molecular biology*, 175(2):159–174, 1984.

[10] Sam-Yong Park, Takeshi Yokoyama, Naoya Shibayama, Yoshitsugu Shiro, and Jeremy R.H. Tame. 1.25 Å resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. *Journal of Molecular Biology*, 360(3):690 – 701, 2006.

[11] C. Viappiani, S. Bettati, S. Bruno, L. Ronda, S. Abbruzzetti, A. Mozzarelli, and W.A. Eaton. New insights into allosteric mechanisms from trapping unstable protein conformations in silica gels. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14414, 2004.

[12] S. Bettati, C. Viappiani, and A. Mozzarelli. Hemoglobin, an "evergreen" red protein. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1794(9):1317 – 1324, 2009.

[13] A. Hinz and R. Tampé. ABC Transporters and Immunity: Mechanism of Self-Defense. *Biochemistry*, 51(25):4981–4989, 2012.

[14] M.M. Gottesman and S.V. Ambudkar. Overview: ABC transporters and human disease. *Journal of bioenergetics and biomembranes*, 33(6):453–458, 2001.

[15] D. Barthelme, S. Dinkelaker, S.V. Albers, P. Londei, U. Ermler, and R. Tampé. Ribosome recycling depends on a mechanistic link between the FeS cluster domain and a conformational switch of the twin-ATPase ABCE1. *Proceedings of the National Academy of Sciences*, 108(8):3228–3233, 2011.

[16] H.A. Scheraga, M. Khalili, and A. Liwo. Protein-folding dynamics: overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.*, 58:57–83, 2007.

[17] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glattli, P.H. Hunenberger, et al. Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie International Edition in English*, 45(25):4064, 2006.

[18] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[19] D. Seeliger and B.L. de Groot. Protein thermostability calculations using alchemical free energy simulations. *Biophysical journal*, 98(10):2309–2316, 2010.

[20] G. Portella, J.S. Hub, M.D. Vesper, and B.L. De Groot. Not only enthalpy: large entropy contribution to ion permeation barriers in single-file channels. *Biophysical journal*, 95(5):2275–2282, 2008.

[21] D.A. Köpfer, U. Hahn, I. Ohmert, G. Vriend, O. Pongs, B.L. de Groot, and U. Zachariae. A molecular switch driving inactivation in the cardiac K+ channel hERG. *PloS one*, 7(7):e41023, 2012.

[22] D. Matthes, V. Gapsys, and B.L. de Groot. Driving Forces and Structural Determinants of Steric Zipper Peptide Oligomer Formation Elucidated by Atomistic Simulations. *Journal of Molecular Biology*, 421(2-3):390–416, 2012.

[23] F. Gräter, J. Shen, H. Jiang, M. Gautel, and H. Grubmüller. Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations. *Biophysical journal*, 88(2):790–804, 2005.

[24] Kresten Lindorff-Larsen, Nikola Trbovic, Paul Maragakis, Stefano Piana, and David E. Shaw. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society*, 134(8):3787–3791, 2012.

[25] L.V. Bock, C. Blau, G.F. Schröder, N. Fischer, H. Stark, M.V. Rodnina, A.C. Vaiana, and H. Grubmüller. Energy barriers and driving forces of tRNA translocation through the ribosome. 2012.

[26] O.F. Lange, N.A. Lakomek, C. Farès, G.F. Schröder, K.F.A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B.L. de Groot. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *science*, 320(5882):1471–1475, 2008.

[27] J.H. Peters and B.L. de Groot. Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. *PLoS computational biology*, 8(10):e1002704, 2012.

[28] R.O. Dror, R.M. Dirks, JP Grossman, H. Xu, and D.E. Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual Review of Biophysics*, 41:429–452, 2012.

[29] M. Born and W. Heisenberg. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 379(9):1–31, 1924.

[30] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, and H.J.C. Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

[31] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[32] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[33] W.F. van Gunsteren, SR Billeter, A. A. Eising, P.H. Hünenberger, P. Krüger, A.E. Mark, WRP Scott, and I.G. Tironi. *Biomolecular Simulation: The GROMOS96 manual and user guide*, 1996.

[34] C. Oostenbrink, A. Villa, A.E. Mark, and W.F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry*, 25(13):1656–1676, 2004.

[35] W.L. Jorgensen, D.S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[36] G.A. Kaminski, R.A. Friesner, J. Tirado-Rives, and W.L. Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487, 2001.

[37] B.R. Brooks, C.L. Brooks, A.D. MacKerell, L. Nilsson, RJ Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[38] A.D. MacKerell Jr, D. Bashford, M. Bellott, R.L. Dunbrack Jr, JD Evanseck, MJ Field, S. Fischer, J. Gao, H. Guo, S. Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[39] S. Miyamoto and P.A. Kollman. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry*, 13(8):952–962, 1992.

[40] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, et al. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[41] K.A. Feenstra, B. Hess, and H.J.C. Berendsen. Improving efficiency of large timescale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem*, 20:786–798, 1999.

[42] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81:511, 1984.

[43] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity-rescaling. *arXiv preprint arXiv:0803.4060*, 2008.

[44] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[45] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684, 1984.

[46] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089, 1993.

[47] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103:8577, 1995.

[48] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[49] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

[50] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[51] Y.S. Liu, Y. Fang, and K. Ramani. Using least median of squares for structural superposition of flexible proteins. *BMC bioinformatics*, 10(1):29, 2009.

[52] V. Gapsys and B. L. de Groot. 2012.

[53] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics*, 7:412–425, 1993.

[54] S. Mika, B. Schölkopf, A. Smola, K.R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and de-noising in feature spaces. *Advances in neural information processing systems*, 11(1):536–542, 1999.

[55] J.S. Hub and B.L. de Groot. Detection of functional modes in protein dynamics. *PLoS computational biology*, 5(8):e1000480, 2009.

[56] T. Krivobokova, R. Briones, J.S. Hub, A. Munk, and B.L. de Groot. Partial Least-Squares Functional Mode Analysis: Application to the Membrane Proteins AQP1, Aqy1, and CLC-ec1. *Biophysical Journal*, 103(4):786–796, 2012.

[57] M. C. Denham. Implementing partial least squares. *Statistics and Computing*, 5:191–202, 1995.

[58] Inge S. Helland. On the structure of partial least squares regression. *Communications in Statistics - Simulation and Computation*, 17(2):581–607, 1988.

[59] H. Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, 52:2893–2906, 1995.

[60] O.F. Lange, L.V. Schäfer, and H. Grubmüller. Flooding in GROMACS: Accelerated barrier crossings in molecular dynamics. *Journal of computational chemistry*, 27(14):1693–1702, 2006.

[61] BL De Groot, A. Amadei, RM Scheek, NAJ Van Nuland, HJC Berendsen, et al. An extended sampling of the configurational space of hpr from e. coli. *Proteins: Structure, Function & Genetics*, 26(3):314–322, 1996.

[62] BL De Groot, A. Amadei, DMF Van Aalten, and HJC Berendsen. Towards an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. *Journal of Biomolecular Structure and Dynamics*, 13(5):741–751, 1996.

[63] F. Lipmann. Metabolic generation and utilization of phosphate bond energy. *Adv. Enzymol. Relat. Areas Mol. Biol*, 1:99–162, 1941.

[64] R.I. Weed, C.F. Reed, and G. Berg. Is hemoglobin an essential structural component of human erythrocyte membranes? *Journal of Clinical Investigation*, 42(4):581, 1963.

[65] F. Hoppe-Seyler. *Medicinisch-chemische Untersuchungen: aus dem Laboratorium für angewandte Chemie zu Tübingen*. August Hirschwald, 1866.

[66] R. Benesch and RE Benesch. The effect of organic phosphates from the human erythrocyte on the allosteric properties of hemoglobin. *Biochemical and biophysical research communications*, 26(2):162, 1967.

[67] R.E. Benesch, R. Benesch, and C.I. Yu. Oxygenation of hemoglobin in the presence of 2,3-diphosphoglycerate. Effect of temperature, pH, ionic strength, and hemoglobin concentration. *Biochemistry*, 8(6):2567–2571, 1969.

[68] Chr. Bohr, K. Hasselbalch, and August Krogh. Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt. *Skandinavisches Archiv Für Physiologie*, 16(2):402–412, 1904.

[69] R. Benesch and R.E. Benesch. The chemistry of the bohr effect. *Journal of Biological Chemistry*, 236(2):405–410, 1961.

[70] C. Ho and I.M. Russu. How much do we know about the bohr effect of hemoglobin? *Biochemistry*, 26(20):6299–6305, 1987.

[71] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, and H. Wyckoff. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

[72] M. F. Perutz, M. G. Rossman, A. F. Cullis, H. Muirhaed, G. Will, and A. C. T. North. Structure of haemoglobin. a three dimensional fourier synthesis at 5.5 Å resolution obtained by x-ray analysis. *Nature (Lond.)*, 185:416–422, 1960.

[73] H. Lehmann and RW Carrell. Variations in the structure of human haemoglobin with particular reference to the unstable haemoglobins. *British medical bulletin*, 25(1):14–23, 1969.

[74] G. Balakrishnan, M.A. Case, A. Pevsner, X. Zhao, C. Tengroth, G.L. McLendon, and T.G. Spiro. Time-resolved absorption and UV resonance Raman spectra reveal stepwise formation of T quaternary contacts in the allosteric pathway of hemoglobin. *Journal of molecular biology*, 340(4):843–856, 2004.

[75] E. Antonini. Interrelationship between structure and function in hemoglobin and myoglobin. *Physiological reviews*, 45(1):123–170, 1965.

[76] J. Monod, J. Wyman, and J.P. Changeux. On the nature of allosteric transitions: a plausible model. *Journal of molecular biology*, 12(7):88–118, 1965.

[77] DE Koshland Jr, G. N methy, and D. Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1):365–385, 1966.

[78] W.A. Eaton, E.R. Henry, J. Hofrichter, A. Mozzarelli, et al. Is cooperative oxygen binding by hemoglobin really understood? *Nature structural biology*, 6:351–358, 1999.

[79] W.A. Eaton, E.R. Henry, J. Hofrichter, S. Bettati, C. Viappiani, and A. Mozzarelli. Evolution of allosteric models for hemoglobin. *IUBMB life*, 59(8-9):586–599, 2007.

[80] E. R. Henry, S. Bettati, J. Hofrichter, and W. A. Eaton. A tertiary two-state allosteric model for hemoglobin. *Biophysical Chemistry*, 98(1–2):149 – 164, 2002.

[81] A. Cooper and DTF Dryden. Allostery without conformational change. *European Biophysics Journal*, 11(2):103–109, 1984.

[82] Vincent J. Hilser, James O. Wrabl, and Hesam N. Motlagh. Structural and energetic basis of allostery. *Annual Review of Biophysics*, 41(1):585–609, 2012.

[83] M.S. Shadrina, A.M. English, and G.H. Peslherbe. Effective Simulations of Gas Diffusion Through Kinetically Accessible Tunnels in Multisubunit Proteins: O_2 Pathways and Escape Routes in T-state Deoxyhemoglobin. *Journal of the American Chemical Society*, 2012.

[84] S.V. Lepeshkevich, S.A. Biziuk, A.M. Lemeza, and B.M. Dzhagarov. The kinetics of molecular oxygen migration in the isolated [alpha] chains of human hemoglobin as revealed by molecular dynamics simulations and laser kinetic spectroscopy. *Biochimica et Biophysica Acta (BBA)-Proteins & Proteomics*, 2011.

[85] N. Ramadas and J.M. Rifkind. Molecular Dynamics of Human Methemoglobin: The Transmission of Conformational Information between Subunits in an $\alpha\beta$ Dimer. *Biophysical journal*, 76(4):1796–1811, 1999.

[86] Liliane Mouawad, David Perahia, Charles H. Robert, and Christophe Guilbert. New insights into the allosteric mechanism of human hemoglobin from molecular dynamics simulations. *Biophysical Journal*, 82(6):3224 – 3245, 2002.

[87] O.K. Yusuff, J.O. Babalola, G. Bussi, and S. Raugei. Role of the subunits interactions in the conformational transitions in adult human hemoglobin: an explicit solvent molecular dynamics study. *The Journal of Physical Chemistry B*, 2012.

[88] J.S. Hub, M.B. Kubitzki, and B.L. de Groot. Spontaneous Quaternary and Tertiary TR Transitions of Human Hemoglobin in Molecular Dynamics Simulation. *PLoS computational biology*, 6(5):e1000774, 2010.

[89] E. Espinosa, E. Molins, and C. Lecomte. Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities. *Chemical physics letters*, 285(3-4):170–173, 1998.

[90] M.F. Perutz. Structure and function of haemoglobin: I. a tentative atomic model of horse oxyhaemoglobin. *Journal of Molecular Biology*, 13(3):646 – IN2, 1965.

[91] E.J. Fernandez, C. Abad-Zapatero, and K.W. Olsen. Crystal structure of lys$\beta_1$82-lys$\beta_1$82 crosslinked hemoglobin: A possible allosteric intermediate. *Journal of Molecular Biology*, 296(5):1245–1256, 2000.

[92] S.-Y. Park, N. Shibayama, T. Hiraki, and J.R.H. Tame. Crystal structures of unliganded and half-liganded human hemoglobin derivatives cross-linked between lys 82β_1 and lys 82β_2. *Biochemistry*, 43(27):8711–8717, 2004.

[93] C.O.N.S. Reed, R. Hampson, S. Gordon, R.T. Jones, M.J. Novy, B. Brimhall, M.J. Edwards, and R.D. Koler. Erythrocytosis secondary to increased oxygen affinity of a mutant hemoglobin, hemoglobin kempsey. *Blood*, 31(5):623, 1968.

[94] H.F. Bunn, R.C. Wohl, T.B. Bradley, M. Cooley, and Q.H. Gibson. Functional properties of hemoglobin kempsey. *Journal of Biological Chemistry*, 249(23):7402–7409, 1974.

[95] M.D. Vesper. Verfeinerung der Proteinkristallographie durch Flexibilitätsvorhersagen (Diplomarbeit). Master's thesis, Max-Planck-Institut für Biophysikalische Chemie & Georg-August-Universität Göttingen, 2008.

[96] A.Y. Kovalevsky, T. Chatake, N. Shibayama, S.Y. Park, T. Ishikawa, M. Mustyakimov, Z. Fisher, P. Langan, and Y. Morimoto. Direct determination of protonation states of histidine residues in a 2 Å neutron structure of deoxy-human normal adult hemoglobin and implications for the bohr effect. *Journal of molecular biology*, 398(2):276–291, 2010.

[97] H. Wajcman, J. Kister, F. Galactéros, A. Spielvogel, M.J. Lin, G.J.A. Vidugiris, R.E. Hirsch, J.M. Friedman, and R.L. Nagel. Hb Montefiore (α126 (H9) Asp$\rightarrow$ Tyr) High Oxygen Affinity and Loss of Cooperativity Secondary to C-Terminal Disruption. *Journal of Biological Chemistry*, 271(38):22990–22998, 1996.

[98] I Klein, B Sarkadi, and A Váradi. An inventory of the human abc proteins. *Biochimica et Biophysica Acta - Biomembranes*, 1461(2):237 – 262, 1999.

[99] ML Mimmack, MP Gallagher, SR Pearce, SC Hyde, IR Booth, and CF Higgins. Energy coupling to periplasmic binding protein-dependent transport systems: stoichiometry of atp hydrolysis during transport in vivo. *Proceedings of the National Academy of Sciences*, 86(21):8257, 1989.

[100] Roger J.P. Dawson and Kaspar P. Locher. Structure of the multidrug ABC transporter Sav1866 from Staphylococcus aureus in complex with AMP-PNP. *FEBS Letters*, 581(5):935 – 938, 2007.

[101] C. Bisbal, C. Martinand, M. Silhol, B. Lebleu, and T. Salehzada. Cloning and Characterization of a RNase L Inhibitor. *Journal of Biological Chemistry*, 270(22):13308–13317, 1995.

[102] J.R. Lingappa, J.E. Dooher, M.A. Newman, P.K. Kiser, and K.C. Klein. Basic residues in the nucleocapsid domain of Gag are required for interaction of HIV-1 gag with ABCE1 (HP68), a cellular protein

important for HIV-1 capsid assembly. *Journal of Biological Chemistry*, 281(7):3773–3784, 2006.

[103] Z. Chen, J. Dong, A. Ishimura, I. Daar, A.G. Hinnebusch, and M. Dean. The essential vertebrate ABCE1 protein interacts with eukaryotic initiation factors. *Journal of Biological Chemistry*, 281(11):7452–7457, 2006.

[104] D. Barthelme, U. Scheele, S. Dinkelaker, A. Janoschka, F. MacMillan, S.V. Albers, A.J.M. Driessen, M.S. Stagni, E. Bill, W. Meyer-Klaucke, et al. Structural organization of essential iron-sulfur clusters in the evolutionarily highly conserved ATP-binding cassette protein ABCE1. *Journal of Biological Chemistry*, 282(19):14598–14607, 2007.

[105] A. Karcher, A. Schele, and K.P. Hopfner. X-ray structure of the complete ABC enzyme ABCE1 from Pyrococcus abyssi. *Journal of Biological Chemistry*, 283(12):7962–7971, 2008.

[106] L. Holm and P. Rosenström. Dali server: conservation mapping in 3d. *Nucleic acids research*, 38(suppl 2):W545–W549, 2010.

[107] B. Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1):275–282, 1995.

[108] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[109] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984.

[110] L. Banci, I. Bertini, P. Carloni, C. Luchinat, and P.L. Orioli. Molecular dynamics simulations on HiPIP from chromatium vinosum and comparison with NMR data. *Journal of the American Chemical Society*, 114(27):10683–10689, 1992.

[111] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.