Karlsruhe Institute of Technology

Multiscale Molecular Dynamics Simulations of Histidine Kinase Activity

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

von der KIT-Fakultät für Chemie und Biowissenschaften

des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Fathia Idiris, M.Sci.

Tag der mündlichen Prüfung: 27. April 2022

1. Referent: Prof. Dr. Marcus Elstner

2. Referent: Prof. Dr. Alexander Schug

Erklärung zur Dissertation

Ich erkläre hiermit, dass ich die vorliegende Dissertation selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Weiterhin versichere ich, dass ich die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Fathia Idiris

14.03.2022

Acknowledgments

Firstly, I would like to express my gratitude to my supervisor, Alex Schug, for his continuous support and encouragement throughout my time at KIT. I would also like to thank Tomáš Kubař and Mayukh Kansari for the great collaboration and productive discussions regarding kinases and enhanced sampling techniques. I want to extend my thanks to my first supervisor Marcus Elstner for all his support and guidance.

A huge thanks to all the Karlsruhe and Jülich MBS group members. I've thoroughly enjoyed working alongside you all. Thank you to all the GRK members for all of the insightful discussions and for sharing this experience with me.

I am immensely grateful to my family and friends for all the support they have given me. A special thanks to my parents for always believing in me. Without their encouragement, this would not have been possible. And finally, a big thank you to Eathaar for being the most supportive friend.

Abstract

Two-component systems (TCS) comprising sensor histidine kinases (HK) and response regulator (RR) proteins are key players in bacterial signal transduction mechanisms. The ability of bacteria to detect and respond appropriately to a variety of chemical and physical stimuli is crucial for their survival. Thus, it comes as no surprise that TCS are among the most scrutinized bacterial proteins. Sensor histidine kinases are typically homodimeric integral membrane proteins consisting of multiple domains. Signal detection at the sensor domain of HK triggers a series of transient large-scale conformational changes along the domains. While the structural properties of HKs differ, they all have a conserved kinase core consisting of the catalytic ATP-binding (CA) and dimerization histidine phospho-transfer (DHp) domains. During the signal transduction cascade, the kinase core adopts an asymmetric conformation such that one of the protomers is kinase active, while the other is inactive. This means that the ATP contained in one of the CA domains is in an ideal position to transfer its γ-phosphate group to the conserved histidine in the DHp. This phosphoryl group is then transferred to the response regulator protein, which in turn ensures an appropriate cellular response.

In this work, I examined the conformational dynamics of the kinase core using molecular dynamics (MD) simulations. Two types of HKs were studied: WalK and CpxA. Due to the large system sizes and the required biological time-scales, exploration of these conformational changes is infeasible using classical MD simulations. To circumvent this issue, I introduce a dual-basin structure-based model for WalK histidine kinase. This coarse-grained model provides insights into the transition pathway between the inactive and active states at a considerably low computation cost. After building an intuition with the simplified model, I employed enhanced sampling methods with the atomistic models to gain more detailed insights into the dynamics. The results presented here indicate that the conformational motions of the individual subdomains of the kinase core are tightly coupled.

Zusammenfassung

Zweikomponentensysteme (TCS), bestehend aus einer Sensorhistidinkinase (HK) und einem Antwortregulationsprotein, sind Schlüsselbausteine in bakteriellen Signaluebertragungsmechanismen. Die Fähigkeit von Bakterien auf eine breite Vielfalt von chemischen und physikalischen Stimuli angemessen zu reagieren ist ausschlaggebend für ihr Überleben. Es ist daher nicht überraschend, dass TCS zu den meistuntersuchten bakteriellen Proteinsystemen gehört. Sensorhistidinkinasen sind typischerweise in die Zellmembran integrierte, homodimere Proteine bestehend aus mehreren Domänen. Reizwahrnehmung an der Sensordomäne von HK löst eine Reihe von großskaligen Konformationsübergangen entlang der Domänen aus. Während sich die strukturellen Eigenschaften von verschiedenen HKs unterscheiden können, erhalten sie alle einen katalytischen ATP-bindenden Kern (CA) und dimerisierende Histidinphosphotransferdomänen (DHp). Während der Signalkaskade nimmt der Kern eine asymmetrische Konformation an, sodass die Kinase an einem der Protomere aktiv ist und die der anderen inaktiv. Das ermöglicht es dem ATP in einer der CA-Domänen seine γ-Phosphatgruppe an das Histidin der DHp abzugeben. Diese Phosphorylgrouppe wird anschließend an das Antwortregulationsprotein weitergegeben, die eine angemessene Reaktion der Zelle veranlasst.

In der vorliegenden Arbeit untersuche ich die Konformationsdynamik des Kinasekerns mithilfe von Molekulardynamiksimulationen (MD). Der Fokus der Arbeit liegt auf zwei verschiedenen HKs: WalK und CpxA. Wegen der Größe der Systeme und den erforderlichen biologischen Zeitskalen, ist es nicht möglich die relevanten Konformationsübergänge in klassischen MD-Simulationen zu berechnen. Um dieses Problem zu umgehen, verwende ich ein strukturbasiertes Modell mit paarweisen harmonischen Potentialen. Diese Näherung erlaubt es, den Übergang zwischen dem inaktiven und den aktiven Zustand mit wesentlich geringerem rechnerischen Aufwand zu untersuchen. Nachdem ich das System mithilfe dieses vereinfachten Modells erkundet habe, benutze ich angereicherte Stichprobenverfahren mit atomistischen Modellen um detailliertere Einsichten in die Dynamik zu gewinnen. Die Ergebnisse in dieser Arbeit legen nahe, dass das Verhalten der einzelnen Unterdomänen des Kinasekerns eng miteinander gekoppelt ist.

Contents

Abbreviations

2D Two-dimensional

3D Three-dimensional

ADP Adenosine diphosphate

AMBER Assisted Model Building with Energy Refinement

ATP Adenosine triphosphate

CA Catalytic ATP-binding

CG Coarse-grained

CK Chapman-Kolmogorov

COM Center of Mass

CV Collective Variable

DHp Dimerization and Histidine Phosphotransfer

FEP Free Energy Profile

GAF GMP-specific phosphodiesterases, Adenylyl cyclases and FhlA

GROMACS GROningen MAchine for Chemical Simulations

HAMP Histidine kinase, Adenylyl Cyclase, Methyl-accepting Chemotaxis Protein, and Phosphatase

HK Histidine Kinase

MD Molecular Dynamics

MSM Markov State Model

PAS Per-Arnt-Sim

PC Principal Component

PCA Principal Component Analysis

PCCA+ Robust Perron Cluster Analysis

PDB Protein Data Bank

PME Particle Mesh Ewald

PMF Potential of Mean Force

REC Receiver domain

Rg Radius of Gyration

RMSD Root-mean-square deviation

RR Response Regulator

SBM Structure-based Model

SMD Steered Molecular Dynamics

TCS Two-component System

TM Transmembrane

US Umbrella Sampling

WHAM Weighted Histogram Analysis Method

1. Introduction

1.1. Two-component Systems (TCS)

The growth and survival of a bacterium is dependent on the organism's ability to sense and appropriately respond to environmental changes. Environmental stressors include changes in temperature, osmotic activity, changes in pH, and ligand binding. Two-component systems (TCS) comprising sensor histidine kinase (HK) and response regulator (RR) proteins are key players in archael and bacterial signal transduction mechanisms [1, 2]. While TCS are ubiquitous among bacteria, they are also present in some eukaryotes. However, TCS are notably absent from the human genome, and thus these enzymes are promising targets for developing novel compounds that selectively inhibit the growth of bacteria. For example Waldiomycin, an angucycline antibiotic, inhibits WalK-histidine kinase activity [3]. Anti-TCS drugs work in a different manner to conventional antibiotics and could possibly be effective against various multi-drug-resistant bacteria such as methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus* (VRE) [4–7].

A prototypical two-component system is shown in Figure 1.1. When a signal is detected at the sensor domain of the HK, this signal is propagated along the domains via a series of transient conformational changes [8]. Three types of phosphoryl-transfer reactions can occur along the signalling cascade as a means of conveying the information. Firstly, there is an autokinase (or autophosphorylation) reaction which occurs within the hisitidine kinase. The phosphoryl group is transferred from HK to the response regulator protein. This is often referred to as the phosphotransferase state. Bifunctional histidine kinases (e.g., EnvZ) also function as phosphatase for the RR, and catalyse the cleavage of the phosphoryl group [9, 10]. The autokinase and phosphotransferase states correspond to the active state of HK, whereas the phosphatase state corresponds to the inactive state.

The following section provides an overview of the structural properties of histidine kinases and response regulator proteins, detailing their roles in the signalling pathway.

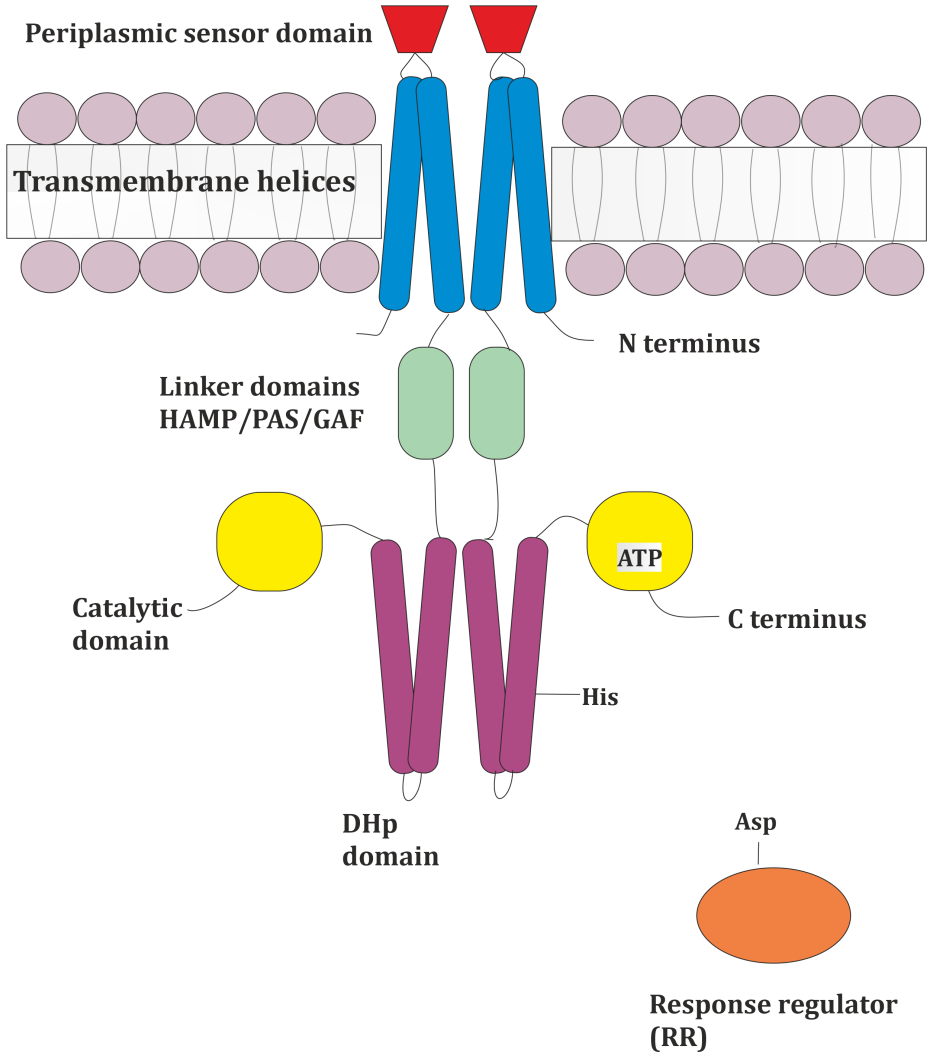
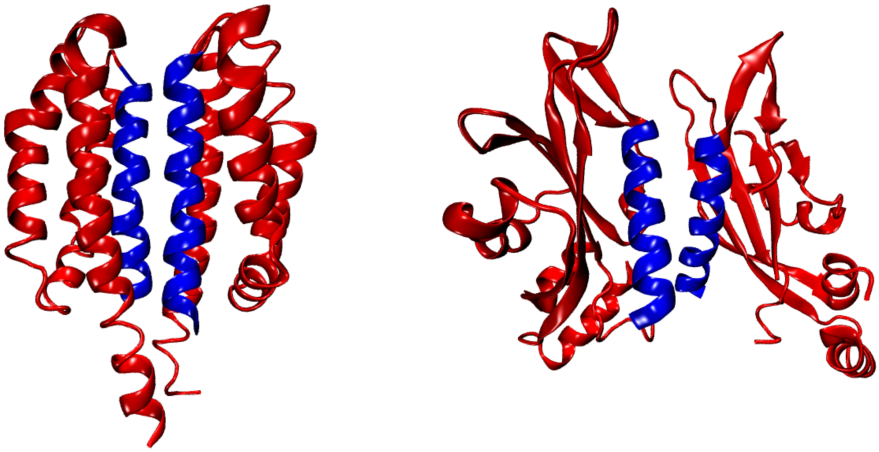


Figure 1.1.: Schematic representation of a protypical two-component system (TCS) featuring domains for signal recognition, transmission, and catalysis. A stimulus is first detected at the periplasmic sensor domain (red) and this signal is transmitted along the transmembrane helices (blue) and the linker domains (green) before reaching the catalytic core at the C-terminus. The catalytic core comprises the dimerization and histidine phosphotransfer (DHp, purple) and catalytic ATP-binding (CA, yellow) domains. Signal detection results in a phosphoryl transfer reaction from ATP in the CA domain to a conserved histidine in the DHp domain. This phosphoryl group is then transferred to an aspartic acid in the response regulator (orange) protein, resulting in an appropriate cellular response.

1.1.1. Histidine Kinases

Hisitidine kinase is an integral membrane protein consisting of multiple domains as shown in Figure 1.1. While the majority of HKs are homodimeric, monomeric conformations of HKs such as HisKA_2-type kinase EL346 from *Erythrobacter litoralis*, have been shown to exist [11, 12].

A signal is first detected at the extracytosolic sensor domain which is flanked by two transmembrane (TM) helices. Different TCS detect different stimuli at their sensor domain, for example DesK detects cold temperature [13], EnvZ senses and responds to osmotic stress [14], and CheA mediates bacterial chemotaxis [15, 16]. As the sensor domains of different TCS have diverse functions, they share little sequence identity [17]. However, available structures for various sensor domains display common structural folds which suggests that the signalling mechanism is conserved [17]. Common structural folds that have been identified include all α-helical structures such as in NarX (Fig. 1.2, left) [18] and TorS [19], and mixed α/β folds such as in PhoQ (Fig. 1.2, right) [20] and CitA [21].



NarX (3EZH) PhoQ (3BQ8)

Figure 1.2.: Examples of two sensor domains of histidine kinases illustrating two common structural folds. On the left is the sensor domain of NarX (PDB ID: 3EZH [18]) which has an all α-helical structure. On the right is the sensor domain of PhoQ (PDB ID: 3BQ8 [20]) which has an α/β fold. The sensor domains (red) are flanked by two transmembrane helices (blue).

Many HKs feature one or more linker domains which serve as a means to transmit the N-terminal signal to the catalytic domains at the C-terminal (see Fig. 1.3). The most widely studied of these linker domains is known as HAMP (histidine kinase, adenylyl cyclase, methyl-accepting chemotaxis protein, and phosphatase) and is found in 30% of HK [22–24]. Other known linker domains include PAS (Per-Arnt-Sim) [25], and GAF (GMP-specific phosphodiesterases, adenylyl cyclases and FhlA) [26].

Several HAMP domain structures have been solved and reveal that they form parallel, homodimeric four-helix coiled coils built from two α-helices connected via a loop region (see Fig. 1.3) [24]. Many models have been proposed to describe the dynamics of the HAMP domain during the signal transduction mechanism. These models are not always mutually exclusive.

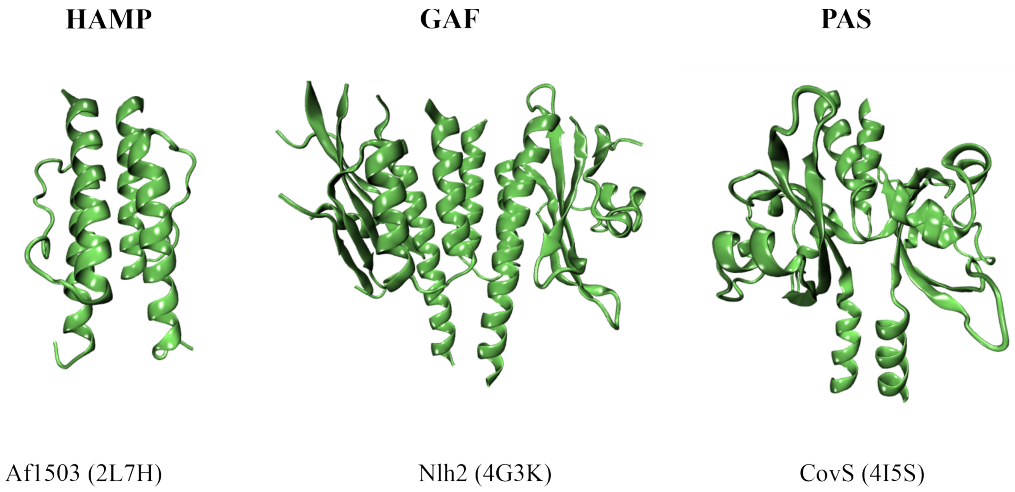


Figure 1.3.: Cytoplasmic linker domains that connect the sensor domain to the catalytic core. Example structures of HAMP (PDB ID: 2L7H [23]), GAF (PDB ID: 4G3K [27]) and PAS (PDB ID: 4I5S [28]) domains are shown.

The gearbox rotation model was suggested based on the first HAMP structure, Af1503 [29]. This structure of HAMP displayed a knobs-to-knobs coiled-coil packing which differed from what had been typically observed in other four-fold symmetric homotetrameric coiled coils. Accordingly, Hulko et al. suggested that a second knobs-into-holes state must also exist [29]. In the gearbox rotation model, the two distinct conformations of HAMP inter-convert via helical rotations around the central axis of the α-helices bundle. The availability of more structures and with the aid of disulfide crosslinking studies, the diagonal scissoring mechanism was proposed [30]. In this model, the domain alternates between one state in which the α-helices are tightly packed and a second state where the helices are splayed out at the C terminus, forming a more loosely packed bundle.

The focus of this thesis is on the conformational dynamics of the kinase core. The kinase (or catalytic) core consists of the dimerization and histidine phosphotransfer (DHp) and catalytic ATP-binding (CA) domains (see Fig. 1.4A). These two domains are highly conserved across HKs. Similarily to the HAMP domain, the DHp domain is typically homodimeric with each protomer comprising two α-helices that form an antiparallel coiled-coil. The CA domain, on the other hand, has an α/β sandwich fold made up of a five-stranded β-sheet flanked by three α-helices. The nucleotide binds between two α-helices and is held by a highly mobile loop known as the ATP lid. Well conserved nucleotide-binding sequence motifs known as the N, G1, F, and G2 boxes comprise the binding site [32]. The DHp domain also contains a conserved sequence motif known as the H box. This region contains a phosphorylatable histidine which is involved in an autophosphorylation reaction with the ATP contained in the CA domain.

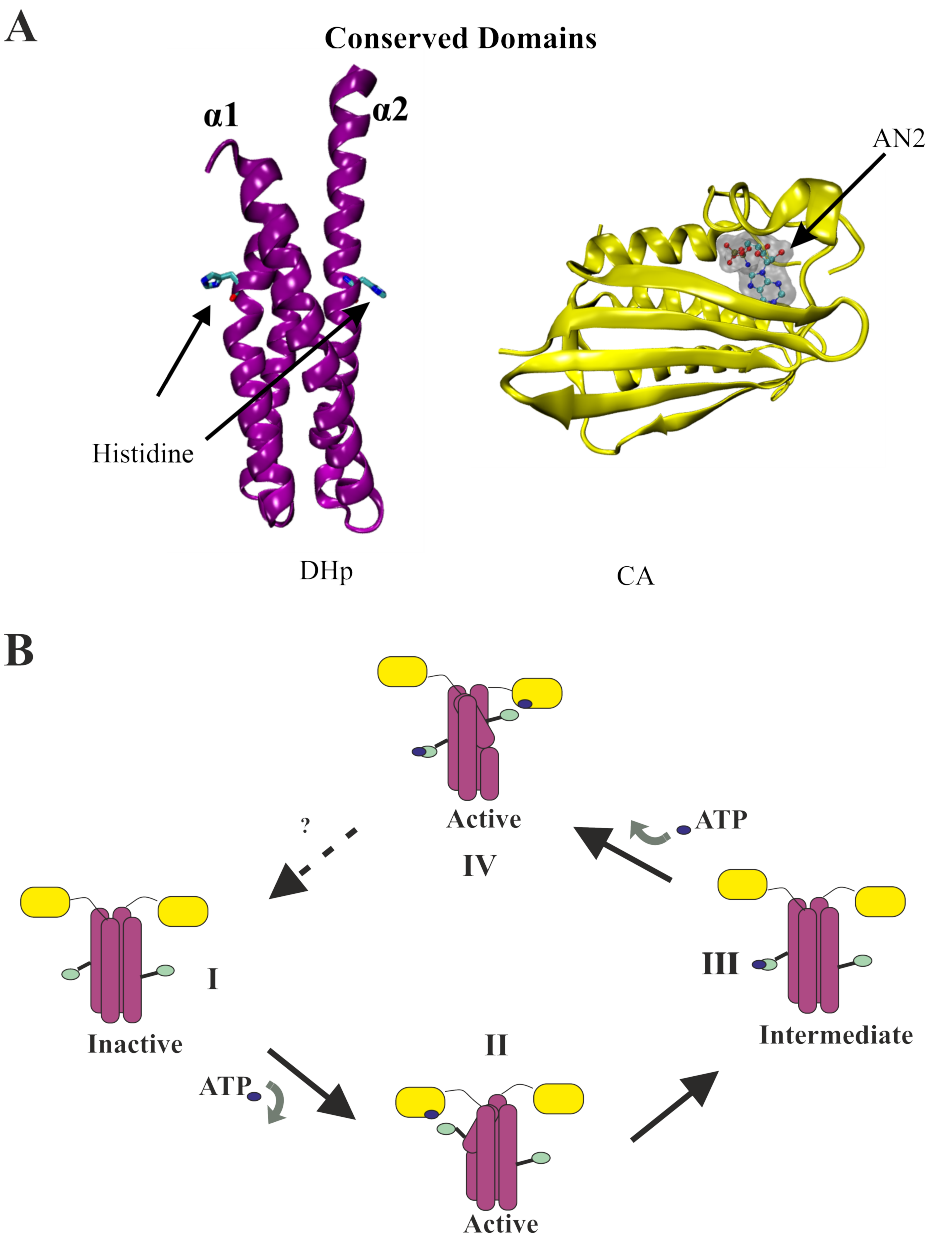


Figure 1.4.: Activity at the catalytic core of histidine kinases. A) The conserved catalytic core comprising the dimerization and histidine phosphotransfer (DHp) and catalytic ATP-binding (CA). The domains of the kinase core of WalK is illustrated here (PDB ID: 4U7O [31]). The DHp (purple, left) domain forms a four-helix bundle and contains a conserved histidine. The CA (yellow, right) domain has an α/β sandwich fold and a highly model loop region which facilitates ATP binding. Here, the ATP analogue, AN2 (Amp phosphoramidate), is found in the ATP-lid. B) The autokinase cycle of the conformational transitions of a *cis*-phosphorylating histidine kinase. The γ-phosphate of ATP and the phosphorylatable histidine are depicted in blue and green, respectively. Redrawn from ref. [28]

.

Following signal detection, the conserved core adopts an asymmetric conformation such that one of the two protomers of the homodimer is kinase active, while the other is inactive (see Fig. 1.4B) [34]. This means that the ATP molecule contained in one of the CA domains

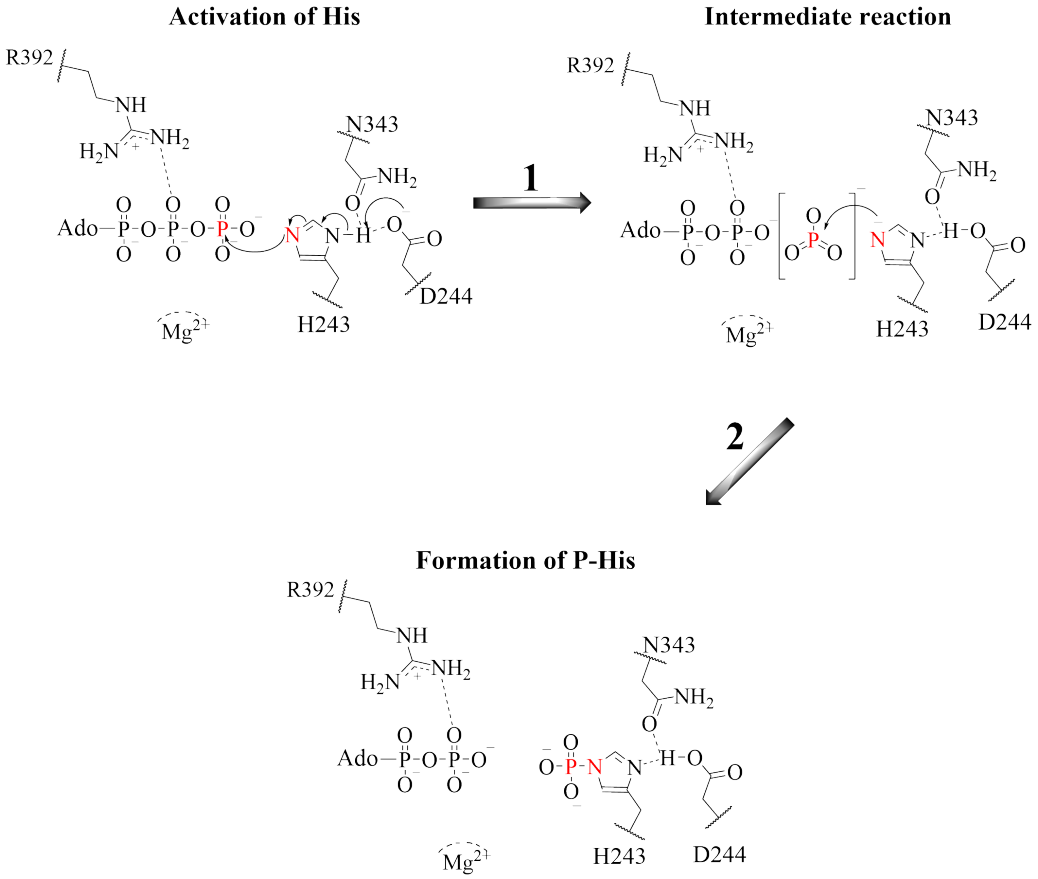


Figure 1.5.: Proposed mechanism for histidine kinase autophosphorylation. In **step 1** His nucleophilicity is enhanced by an acidic residue that acts as a general base and a polar residue that interacts with this base. In **step 2**, the phosphoryl group is transferred to the activated His to form the final products P~His and ADP. The resultant negatively charged β-P in the ADP is neutralized by a positively charged residue in the ATP lid and by a magnesium cation. The interacting residues and residue numbers reflect those in EnvZ. [33].

is in close proximity to the phosphorylatable histidine found in the DHp domain. In the proposed reaction mechanism by Casino et al. [33], the nucleophilicity of the histidine is first enhanced by a neighboring acidic residue (e.g. Asp244 in EnvZ) that acts as a general base (see Fig. 1.5). Once the appropriate tautomeric form of histidine has been induced, a nucleophilic attack of the ϵ-N to the γ-phosphate of ATP can occur. The transfer of this phosphoryl group to a conserved aspartatic acid in the response regulator protein, results in an appropriate cellular response.

Whether the autophosphorylation reaction occurs in *cis* or *trans* is believed to be dependent on the intrinsic handedness of the hairpin loop between the two DHp α-helices [35]. If this hairpin loop turns right, the CA domain of one protomer is brought closer to the phosphorylatable histidine of the DHp domain of the other protomer. Therefore, autophosphorylation proceeds in *trans* (see Fig. 1.6, right). Alternatively, if the loop

turns left, the CA domain is positioned closer to the histidine within the DHp domain of the same protomer. As a result, autophosphorylation occurs in *cis* (see Fig. 1.6, left). In this work, I studied the kinase core dynamics of WalK and CpxA which are *cis*- and *trans*-phosphorylating HKs, respectively.

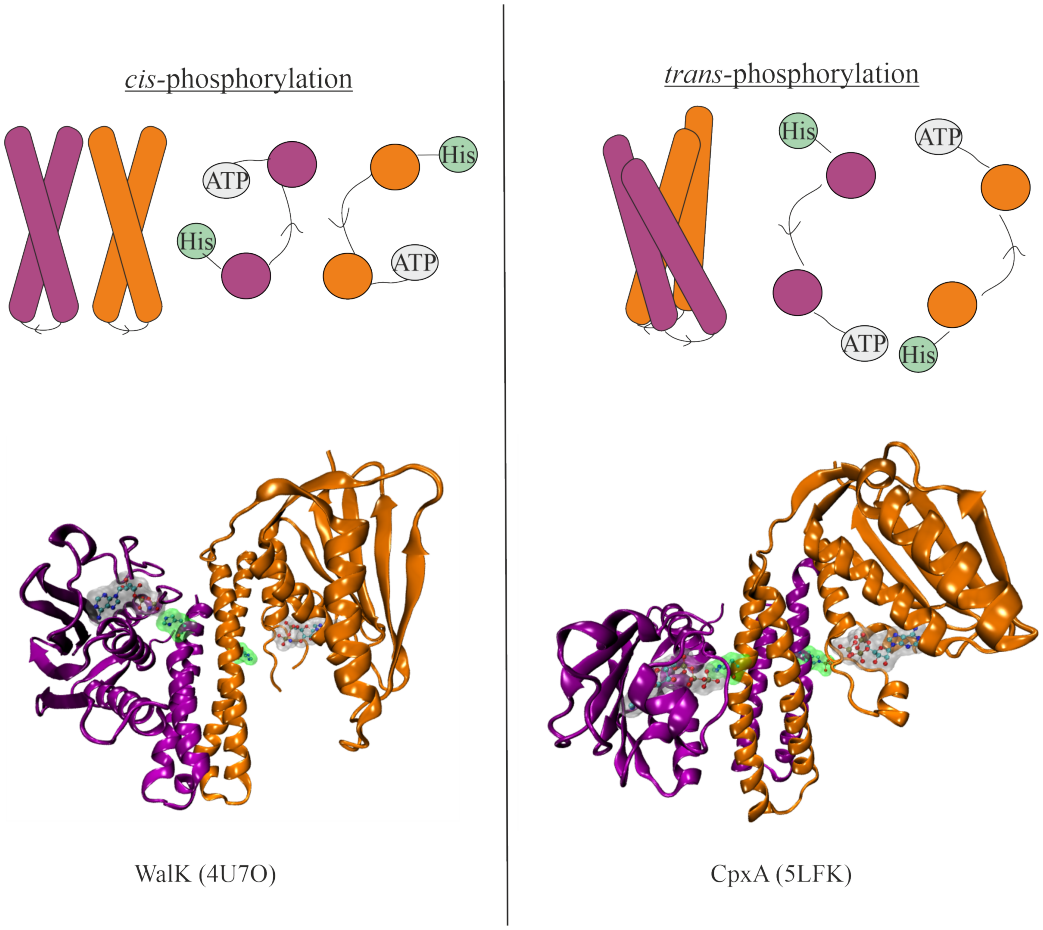


Figure 1.6.: Schematic view of the orientation of the DHp helices which determine whether the autophosphorylation reaction of histidine kinases occurs in *cis* or *trans*. Kinases with a right-handed four-helix bundle phosphorylate in *trans*, whereas those with a left-handed four-helix bundle phosphorylate in *cis*. The structures of the *cis*- (PDB ID: 4U7O [31]) and *trans*-phosphorylating (PDB ID: 5LFK [36]) histidine kinases that are studied in this work are depicted here.

1.1.2. Response Regulator (RR) Proteins

The prototypical response regulator protein consists of a receiver (REC) domain and an effector domain [37]. The REC domain contains the phosphorylatable aspartate residue within a conserved Rossmann-like α/β fold. Activation of an RR in a TCS is associated with the phosphoryl transfer reaction from the histidine residue in the DHp domain of HK

to the aspartate residue in the REC domain. Although some RRs such as CheY [38] and Spo0F [39] only contain an REC domain, the majority of bacterial RRs are transcription regulators with a DNA-binding effector domain [37]. This reflects the importance of transcriptional regulation as a response to environmental change. An example of such a response regulator is PhoB of the PhoB-PhoR two-component system (see Fig. 1.7) [40–42]. Other RRs have effector domains for binding ligands, RNA or proteins.

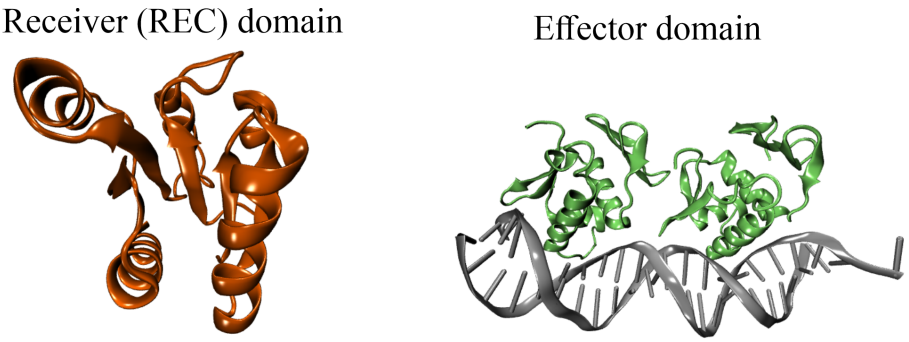


Figure 1.7.: Response regulation in two-component systems. The receiver domain (PDB ID: 1B00 [40]) and DNA-binding effector (PDB ID: 1GXP [41]) domains of the response regulator protein, PhoB, are shown in orange and green, respectively.

Early insights into the crucial HK-RR interactions were gained through analysis of the *B. subtilis* Spo0F-Spo0B complex structure [39, 43]. Spo0B itself is not an HK but instead a phosphotransferase with structural similarities to HKs [44, 45]. The single-domain RR Spo0F contains an aspartate residue that is phosphorylated by sensor HK KinA. This phosphoryl group is transferred to a histidine residue on Spo0B, which is then transferred to another response regulator known as Spo0A. The notion that the interactions formed in the Spo0F-Spo0B complex are related to the interactions in HK-RR two-component systems was later confirmed when the first HK-RR structure of *T. maritima* proteins HK853 and RR468 was published [46]. As seen in Figure 1.8, Spo0F binds to the DHp-type four-helix bundle of Spo0B in a similar manner as the HK853-RR468 complex. Both complexes formed significant contacts between residues in the $\alpha 1$ helix of the DHp and the $\alpha 1$ helix and $\beta 5$-$\alpha 5$ loop of the RR. In the HK853-RR468 complex, additional contacts also formed between the $\alpha 2$ helix in the DHp domain and the $\alpha 1$ helix of the RR. Due to the small differences in the orientation of these helices in the Spo0B/Spo0F structure, these interactions are not realized.

The focus of this work lies in the autokinase functionality of histidine kinases. Since this stage is prior to the formation of HK-RR interaction, I simulated HK in the absence of its corresponding RR. Moreover, including RR in the simulation box would increase the computational costs tremendously due to the additional degrees of freedom. Nonetheless,

more detailed discussions on RR structure and signalling mechanisms can be found in refs. [37] and [47].

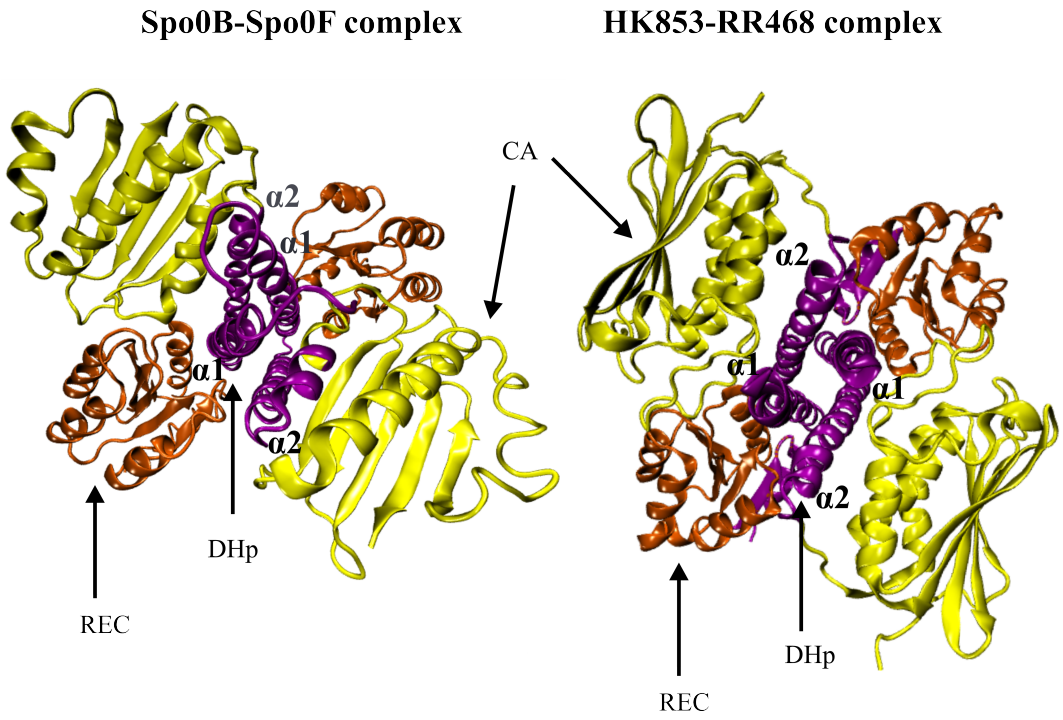


Figure 1.8.: Structural comparison between complexes of sporulation phosphorelay proteins Spo0B/Spo0F (PDB ID: 1F51 [39]) and the HK/RR pair HK853/RR468 (PDB ID: 3DGE [48]). DHp domains, CA domains, and RR REC domains are illustrated in purple, yellow, and orange, respectively.

2. Theoretical Background

This chapter outlines the computational methods used in this work. I performed molecular dynamics (MD) simulations to explore the conformational dynamics of two types of histidine kinases: WalK and CpxA. Firstly, an introduction into classical MD simulations and molecular mechanics is given in section 2.1. Despite the rapid increase in computer power and advances in MD algorithms in recent years, classical MD simulations are still limited to relatively small systems and biologically fast processes. To address this limitation, coarse-grained (CG) models have been developed and have since surged in popularity (see sec. 2.2). CG models are simplified representations of complex systems which accelerate the computation time. I used structure-based models (SBM), which are a class of CG models based on energy landscape theory and the principle of minimal frustration (see sec. 2.2.1).

In this work, I constructed a dual-basin SBM which is a modified SBM that comprises structural informational of two known protein conformations. Methods for constructing multiple basin models are discussed in section 2.2. Analytical techniques for interpreting molecular dynamics simulations are explained in section 2.3. Finally, the enhanced sampling methods for predicting the free energy profiles of the large-scale conformational changes are detailed in section 2.4.

2.1. Molecular Dynamics

Molecular dynamics (MD) simulations are an effective and widely used computational tool for investigating large biomolecular systems. In this technique, the time evolution of a system of interacting particles is computed according to the Born–Oppenheimer (BO) approximation. Due to the significantly lower mass and higher velocity of electrons, nuclei and electrons can be treated separately. With this assumption, electronic motion is not consider in MD simulations and the nuclei are treated as point particles which move

according to Newtonian dynamics. To compute the phase space trajectory, the equations of motion must be numerically solved for the interacting particles. Newton's second law describes the motion of a classical particle of mass m with acceleration $\overrightarrow{a}$ moving under the influence of a force $\overrightarrow{F}$.

$$\overrightarrow{F} = m \cdot \overrightarrow{a} \tag{2.1}$$

The classical equations of motion expressed as a function of time are:

$$\overrightarrow{F}(\overrightarrow{r}) = m \cdot \frac{d^2\overrightarrow{r}}{dt^2} \tag{2.2}$$

$$\frac{d\overrightarrow{r}(t)}{dt} = \frac{\overrightarrow{p}(t)}{m} \tag{2.3}$$

$$\frac{d\overrightarrow{p}(t)}{dt} = \overrightarrow{F}(\overrightarrow{r}) \tag{2.4}$$

where $\overrightarrow{p}$ and $\overrightarrow{r}$ are the vector momentum and positions of the particles respectively. The trajectory is defined by the integration of equations 2.3 and 2.4, to determine the positions and velocities of atoms at every time step. Choosing an appropriate time step length is crucial in yielding a realistic simulation of the molecular changes. A time step that is too large results in instabilities in the integration while conversely, not enough phase space is explored when the time step is too small. All bonds involving hydrogen are constrained by the LINCS algorithm [49], allowing the use of a time step of 2 fs for the all-atom MD simulations in this work.

Classical force fields describe the interactions between the particles using potential energy functions. The force $\overrightarrow{F}$ is a gradient of the potential energy V which can then be derived by using the following expression:

$$\overrightarrow{F}_i = - \bigtriangledown V(\overrightarrow{r}_1, \overrightarrow{r}_2, \ldots \overrightarrow{r}_N). \tag{2.5}$$

Several algorithms are available for the numerical integration of the Newton's equations of motion with GROMACS software package [50]. I applied the leap-frog integration method [51] in all the atomistic MD simulations presented in this work.

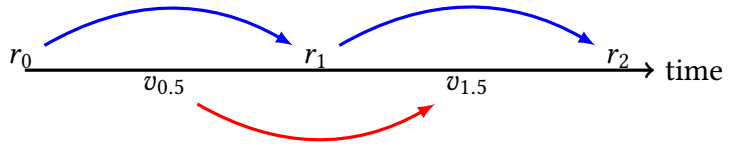


Figure 2.1.: Schematic of the Leap-frog integration method. The positions (r) and velocities (v) leap over each other.

2.1.1. Leap-Frog Integration Method

The leap-frog integration method [51] is one of the variants of the Verlet scheme [52] for integrating the equations of motion in molecular dynamics simulations. This algorithm uses the velocities v at half-integer time steps and the positions r at time t to update the positions and velocities based on the following relations:

$$v(t + \frac{1}{2}\Delta t) = v(t - \frac{1}{2}\Delta t) + a(t) \cdot \Delta t \tag{2.6}$$

$$r(t + \Delta t) = r(t) + v(t + \frac{1}{2}\Delta t) \cdot \Delta t. \tag{2.7}$$

The origin of the name "leap-frog" comes from the way that the values of r and v are leaping over each other as illustrated in Fig. 2.1.

2.1.2. Stochastic Dynamics

Stochastic or Langevin dynamics [53] is an approach that incorporates additional friction and noise terms to Newton's equations of motion to implicitly model solvent effects. For a system of N interacting particles with masses M and coordinates x, the Langevin equation is:

$$M\frac{d^2x}{dt^2} = - \bigtriangledown V(x) - M\gamma\frac{dx}{dt} + R(t) \tag{2.8}$$

where V is the potential energy, γ is the friction constant, and $R(t)$ is a stationary Gaussian process with mean, $\langle R(t) \rangle$, of zero. Hydrodynamic interactions are not explicitly modelled, instead the random force R and friction constant γ mimic the molecular collisions and viscosity of the solvent. The reduction in the degrees of freedom means that Langevin models are much less computationally expensive than the corresponding all-atom in

explicit solvent MD simulations. The random force and the frictional force are related by the fluctuation/dissipation theorem. This relation can be expressed by the following equation:

$$\left\langle R(t)R(t')^T \right\rangle = 2\gamma k_B T M \delta(t - t'). \tag{2.9}$$

Here, δ is the Dirac delta, T is the temperature and k_B is the Boltzmann's constant. In MD simulations of canonical ensembles (NVT), the number of particles (N), the volume of the system (V) and the temperature (T) must remain constant. A thermostat is required to ensure that the temperature of the system remains constant. As seen in equation 2.9, the temperature is implicitly controlled in Langevin dynamics. To obtain a stochastic trajectory, GROMACS numerically integrates equation 2.8 by using a leap-frog integrator which applies the friction and velocity term in an impulse fashion. In this thesis, stochastic dynamics were used in the coarse-grained simulations of histidine kinase.

2.1.3. Pressure and Temperature Coupling

MD simulations in the canonical (NVT) and isothermal–isobaric (NPT) ensembles require temperature or pressure coupling, respectively. I used the velocity-rescaling thermostat [54] and Parrinello-Rahman barostat [55] implemented in GROMACS, to simulate the conformational changes of histidine kinases.

Due to the changes in the velocities of atoms over time, the system's temperature also changes. The velocity-rescaling thermostat is a variant of the Berendsen thermostat [56] which suppresses the fluctuations in temperature by scaling the velocities of the atoms. The system's temperature is slowly corrected according to:

$$\frac{dT}{dt} = \frac{1}{\tau_T}(T_0 - T_t). \tag{2.10}$$

Here, T_0 is the desired reference temperature, T_t is the temperature at time t and τ_T is the temperature time constant. This means that temperature deviation decays exponentially with time constant, τ_T. The Berendsen thermostat consequently constrains the fluctuations of the kinetic energy and as a result, does not generate a proper canonical ensemble. To ensure that a correct kinetic energy distribution is derived, an additional stochastic term is

introduced by the velocity-rescaling thermostat. The kinetic energy is modified according to the following equation:

$$dK = (K_0 - K)\frac{dt}{\tau_T} + 2\sqrt{\frac{KK_0}{N_f}}\frac{dW}{\sqrt{\tau_T}}, \tag{2.11}$$

where N_f the number of degrees of freedom, K is the kinetic energy, and dW is a Wiener process. In this way, a canonical ensemble is maintained. In the NVT simulations of histidine kinase, a τ_T of 0.1 ps and a reference temperature of 300 K was used.

MD simulations in the NVT ensembles have a constant volume and thus a fixed periodic box size. In contrast, simulations in the isothermal-isobaric ensembles require a barostat which functions by adjusting the box size to maintain constant pressure. I used the Parrinello-Rahman algorithm [55], which is conceptually similar to the extended-ensemble approach of the Nosé-Hoover thermostat [57]. With this barostat, the simulation box vectors obey the matrix equation of motion,

$$\frac{db^2}{dt^2} = VW^{-1}b'^{-1}(P - P_{ref}), \tag{2.12}$$

where b is a 3×3 matrix, V is the volume of the box and W is the matrix parameter that calibrates the strength of the coupling. The current and reference pressures are denoted by P and P_{ref} respectively. This results in the following modified equations of motion:

$$\frac{d^2r_i}{dt^2} = \frac{F_i}{m_i} - M\frac{dr_i}{dt} \tag{2.13}$$

$$M = b^{-1}b'^{-1}\left[b\frac{db'}{dt} + b'\frac{db}{dt}\right] \tag{2.14}$$

The strength of the pressure coupling is determined by the inverse mass parameter matrix W_1. GROMACS automatically calculates an appropriate value for this parameter given that the user supplies an approximate isothermal compressibility, β and the pressure time constant, τ_p. These variables are related through the following equation:

$$(W^{-1})_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2 L}, \tag{2.15}$$

where L is the largest box matrix element. In all NPT simulations of histidine kinase, the pressure time constant, τ_p, was 2.0 ps and the isothermal compressibility, β, was $4.5x10^{-5}$ bar^{-1}.

2.1.4. Force Fields

An empirical force field refers to the functional form of the potential energy, which includes a set of parameters that describe the bonded and non-bonded interactions present in the simulated system. Many force fields have been developed that differ in the formulation of the energy terms and the strategy of parameterization. The parameters are often calibrated to experimental data and/or approximate quantum mechanical calculations [58]. Therefore, force fields are designed for a specific purpose. In the present work, I used AMBER-ff99SB (Assisted Model Building with Energy Refinement) [59]. This is the latest version of AMBER available in GROMACS, with improved side chain dihedral parameters for amino acids.

The summation of the bonded and non-bonded potentials yields the total potential energy of the system:

$$V_{total} = V_{bonded} + V_{nonbonded}. \tag{2.16}$$

The different types of bonded interactions are illustrated in Fig 2.2. These interactions between covalently bonded atoms can be represented by harmonic oscillators centered around a reference ground state. The ground state parameters are defined at the start of simulation and are often taken from x-ray crystallography. For proteins, these structures can be obtained from the protein data bank (PDB) [60]. The bonded interactions potentials are given by:

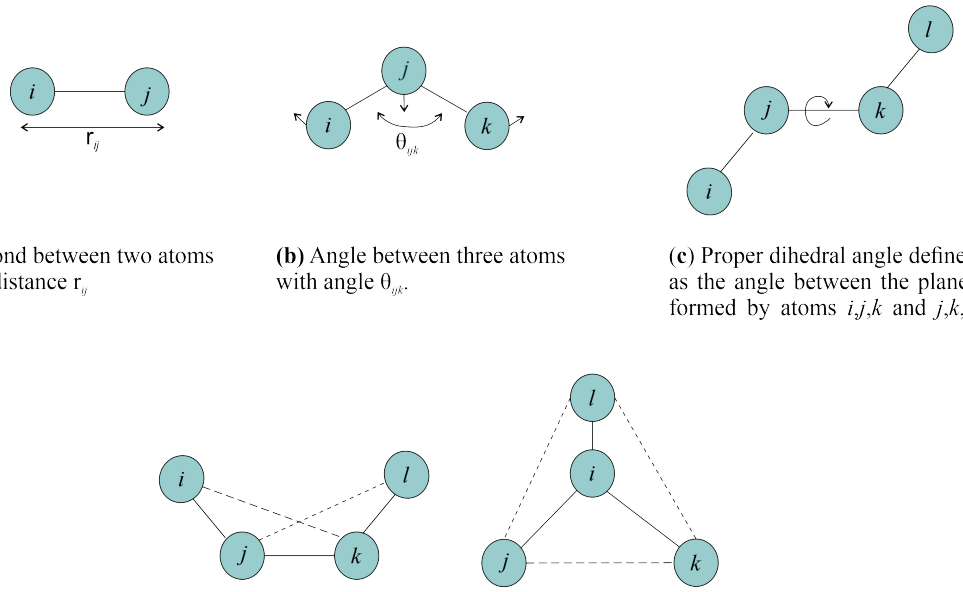
$$V_{bond} = \frac{1}{2}k_b(r - r_0)^2, \tag{2.17}$$

$$V_{angle} = \frac{1}{2}k_\theta(\theta - \theta_0)^2, \tag{2.18}$$

$$V_{dihedral} = k_d(1 + \cos(n\phi - \phi_0)) \text{ and} \tag{2.19}$$

$$V_{improper} = \frac{1}{2}k_i(\chi - \chi_0)^2. \tag{2.20}$$

The bond stretching between two atoms can be modelled by a harmonic spring potential with a force constant, k_b, centered at the predefined equilibrium length, r_0. Similarly, the



(a) Bond between two atoms with distance r_{ij}

(b) Angle between three atoms with angle θ_{ijk}.

(c) Proper dihedral angle defined as the angle between the planes formed by atoms *i,j,k* and *j,k,l*.

(d) Two types of improper dihedral angles defined as the angle between planes formed by atoms *i,j,k* and *j,k,l*. Out of plane bending for rings (left) and out of tetrahedral angle (right).

Figure 2.2.: Summary of the types of bonded interactions formed between atoms.

angle θ between three atoms fluctuates around the reference angle θ_0, by a harmonic angle potential with force constant, k_θ.

Dihedral angles describe the angle between two intersecting planes. Two types of dihedral angles are defined in the force field (see equations 2.19 and 2.20). The first is the proper dihedral angle, ϕ, which is a periodic angle between four atoms with a multiplicity, n, to take into account conformational isomerism. In this periodic function, the reference proper dihedral angle is ϕ_0 and the force constant is k_d. The other type are improper dihedral angles (χ), which are meant to restrain atoms in a planar group to a plane (e.g. aromatic rings). A harmonic potential is again utilized with a force constant, k_i, and a reference angle, χ_0.

The non-bonded potential ($V_{nonbonded}$) is composed of the van der Waals and Coulomb energy terms. The van der Waals potential for two non-bonded atoms at distance r_{ij} is often described by the 12-6 Lennard-Jones potential:

$$V_{LJ} = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right), \tag{2.21}$$

where ϵ_{ij} is the well depth and σ_{ij} is the interatomic distance at which the potential is zero (see Fig. 2.3a). This potential includes a long-ranged attractive term, r^{-6}, due to dispersion forces and a short-ranged repulsive term, r^{-12}, due to Pauli repulsion at short distances.

While the Lennard Jones potential describes the interaction between uncharged particles, the coulombic potential describes the electrostatic interaction between charged particles (see Fig. 2.3b). The Coulombic interaction between atoms with partial charges q_i and q_j is given by,

$$V_{coulomb} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}. \tag{2.22}$$

Here, the distance between the atoms is r_{ij}, and parameters ϵ_0 and ϵ_r are the vacuum and relative permittivity, respectively. Computing the long-range interactions, which decay rapidly in these non-bonded terms, is relatively computationally expensive. For van der Waals interactions, a cutoff can be set such that beyond a predefined interatomic separation distance, the potential is considered zero. This speeds up the computation time. In contrast, electrostatic interactions are by nature, long ranged, and so simply using a cutoff scheme may provide a poor approximation [61]. A more efficient approach is the Particle mesh Ewald (PME) method [62], which separates the slowly-converging potential into two terms that converge more quickly. One is a short-ranged potential in real space, and the other is a long-ranged potential in Fourier space.

2.2. Coarse-Grained Potentials

In recent years, we have witnessed remarkable advances in computing power and the accuracy of force-fields. However, exploring the large-scale conformational changes of complex biomolecular systems remains infeasible using classical MD simulations. This is due to the system size and the immense simulation time required to compute these relatively slow biological processes. The approximate range of time and length scales covered by various molecular models are depicted in Fig. 2.4. This has lead to the development of coarse-grained (CG) models. CG models are simplified representations of complex systems where pseudoatoms replace atoms or even whole molecules. The aim is to reduce the computation cost by reducing the degrees of freedom ,while also being able to reproduce experimental data or behavior observed in high resolution atomistic simulations. A plethora of CG models have been developed with varied levels of resolutions [63–66]. A

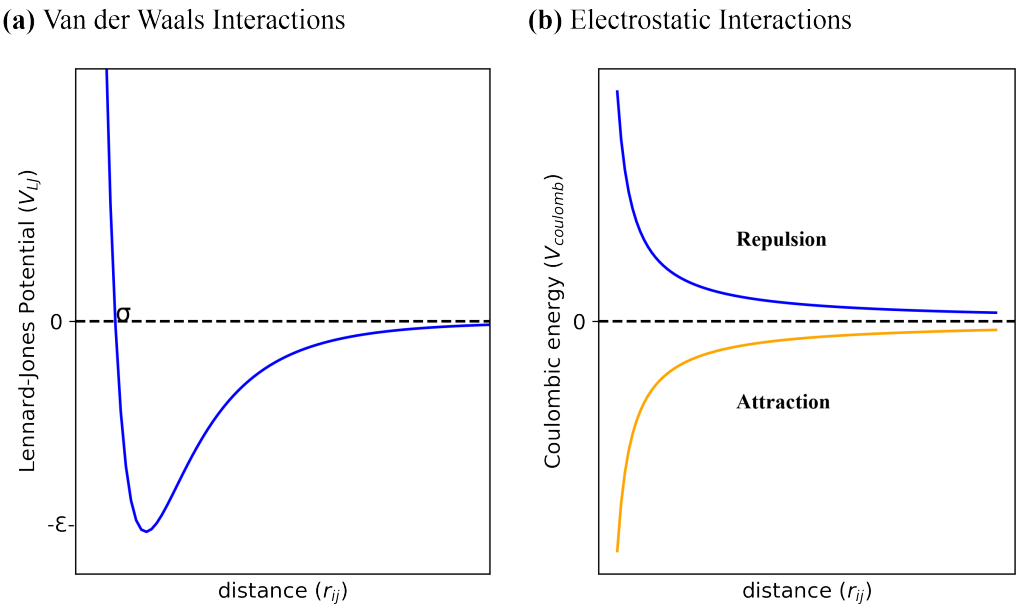


Figure 2.3.: Two types of non-bonded interactions between atoms i and j. (a) Lennard-Jones potential for uncharged atoms i and j, with a well depth ϵ and a zero-potential distance σ. (b) Electrostatic interactions between charged atoms i and j, described by the Coulombic potential.

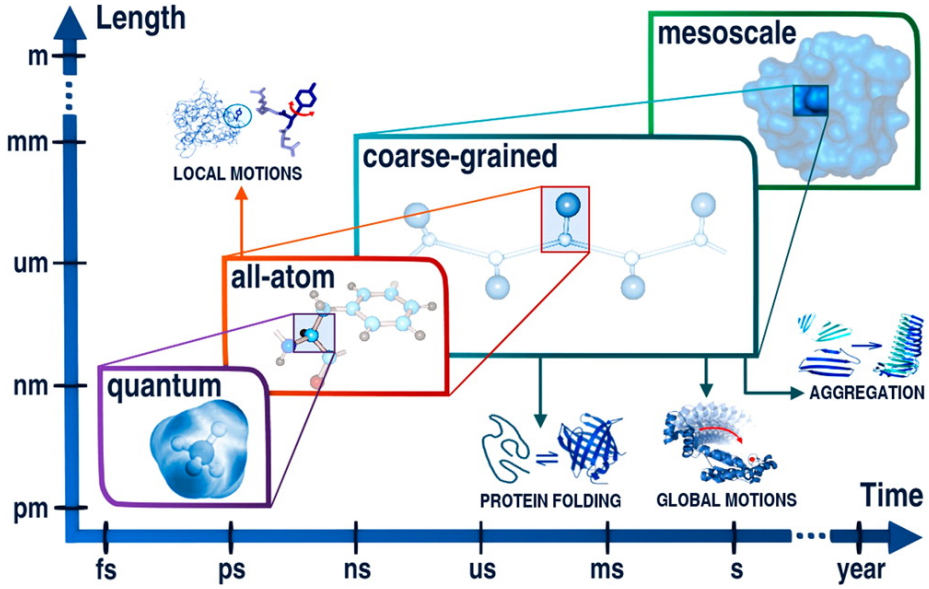


Figure 2.4.: Application ranges for molecular modeling at different resolutions: quantum, all-atom, coarse-grained, and mesoscale. Approximate ranges of time scales and system sizes (lengths) are shown. Examples of processes that can be studied using the models are also depicted. Reprinted with permission from Ref. [66]. Copyright © 2016 American Chemical Society.

few examples of coarse-grained models of a tripeptide are shown in Fig. 2.5. In this work however, I have chosen to use structure-based models (SBMs) [67, 68].

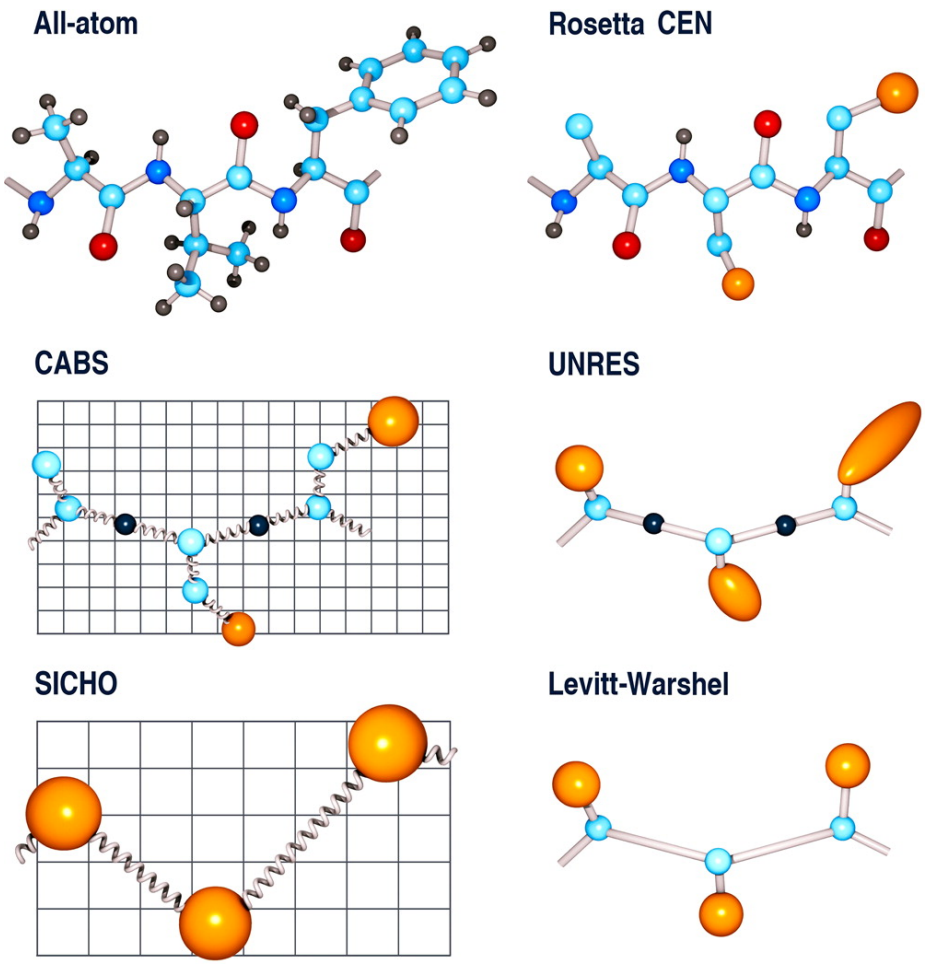


Figure 2.5.: All-atom representation of a tripeptide and the corresponding coarse-grained models. Various coarse-grained models are presented: Rosetta centroid mode (CEN) representation [69], CABS (C-alpha, beta, and side chain) [64], UNRES (united residue) [65], SICHO (side chain only) [63], and Levitt and Warshel model [70]. United side chain atoms are colored in orange. Pseudobonds of fluctuating length are represented as springs and lattice models are shown on the underlying lattice slide. Reprinted with permission from Ref. [66]. Copyright © 2016 American Chemical Society.

2.2.1. Structure-Based Models (SBMs)

Pioneered by Gō et al, native structure-based models (SBMs) or Gō models are a class of coarse-grained models based on energy landscape theory and the principle of minimal frustration [67, 68]. This theory states that the energy landscapes of naturally evolved proteins have a funnel-like shaped bias towards its native folded state (see Fig.2.6) [71]. Energetic traps in local minima in the energy landscape counteract efficient protein folding. These can be either kinetic or thermodynamic traps. Kinetic trapping occurs when the thermal motions of the protein is too small to overcome a large energetic barrier. Thermodynamic trapping refers to a scenario in which the probability of a transition

occurring is low, because the energy gained by overcoming a barrier is too small. The topological feature which accounts for the distribution of these traps is known as energetic roughness. Minimization of frustration during evolution of proteins has ensured efficient folding.

In SBMs, non-bonded interaction pairs are categorized as being either native or non-native contacts. The formulation of the SBM force-field is such that the native contact pair interactions are favourable, and the non-native pairs are less favourable, neutral, or repulsive. This results in a smooth unfrustrated energy landscape, enabling more efficient exploration of the conformational space.

SBMs were initially introduced for studying protein folding and have since been applied to study a wide range of phenomena. This includes protein structure prediction [72], protein allostery [73, 74], misfolding [75], and conformational dynamics within the ribosome [76]. Despite the simplicity of these models, SBM simulations have shown good agreement with experimental results [77–80].

2.2.1.1. Structure-Based Potential

In this work, an SBM of histidine kinase with each amino acid residue represented by a single bead at the α-carbon position was generated using SMOG webtool [81]. The functional form of the C-α Hamiltonian is,

$$\begin{aligned} V_{C\alpha} = & \sum_{bonds} \varepsilon_r \left(r - r_0\right)^2 + \sum_{angles} \varepsilon_\theta \left(\theta - \theta_0\right)^2 + \sum_{backbone} \varepsilon_D F_D \left(\phi\right) \\ & + \sum_{contacts} \varepsilon_C \left[5\left(\frac{\sigma_{ij}}{r}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r}\right)^{10}\right] + \sum_{non-contacts} \varepsilon_{NC} \left(\frac{\sigma_{NC}}{r}\right)^{12} \end{aligned} \tag{2.23}$$

where the dihedral potential F_D is,

$$F_D \left(\phi\right) = \left[1 - cos(\phi - \phi_0)\right] + \frac{1}{2}\left[1 - cos(3(\phi - \phi_0))\right]. \tag{2.24}$$

As seen in the empirical force field, bond stretching and angle bending contributions are expressed by harmonic potentials with reference values r_0 and θ_0 respectively. The backbone dihedrals are given a proper dihedral angle potential, which accounts for conformational isomerism. The reference backbone dihedral is ϕ_0. The energetic weights, ε (in

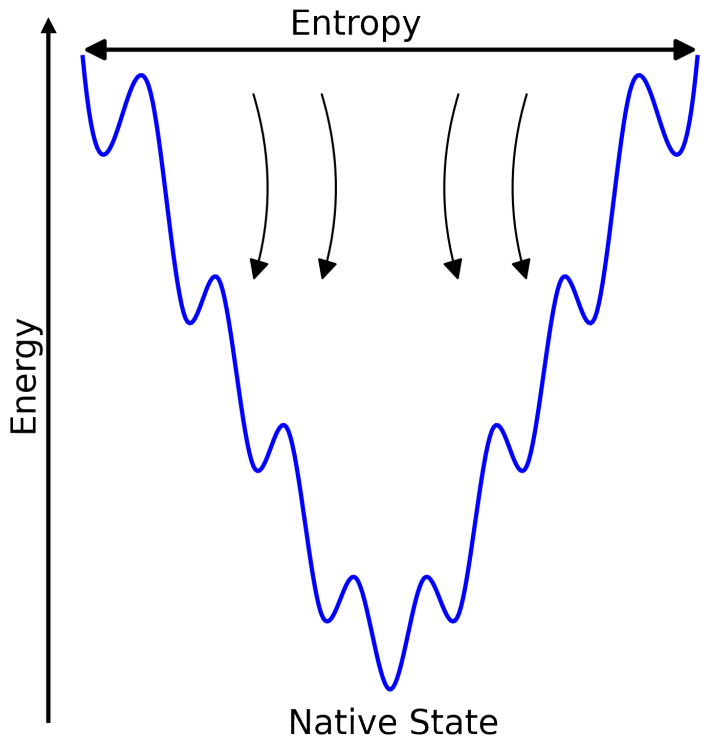


Figure 2.6.: A funneled protein folding energy landscape. A funnel-like shaped bias towards the minimum energy native state through an ensemble of converging pathways. Protein folding is hindered by the overall loss in conformational entropy, but is driven by stabilizing effects such as hydrogen bonding or water expulsion out of hydrophobic core.

reduced units), are ε_r =20000, ε_θ = 40, ε_D = 1, ε_{NC} = 1, and ε_C = 1. Non-bonded interactions are divided into native and non-native contact pairs. The native contact pair potential is a variant of the Lennard Jones potential. In contrast to the standard Lennard Jones potential, the attractive term is proportional to r^{-10}. Longer contact distances are possible in the C-α only models, and so the formation of unphysical unfolded states is more probable when using the 6-12 potential. Using a 10-12 potential minimizes this error. The non-native contacts are deemed unfavorable and so the potential only has a repulsive term, r^{-12}.

2.2.1.2. Determination of Native Contacts

A simple method for identifying native contact pairs is by setting a cutoff distance. At larger cutoff distances, e.g. for C-α-only models of proteins, several nonphysical contacts

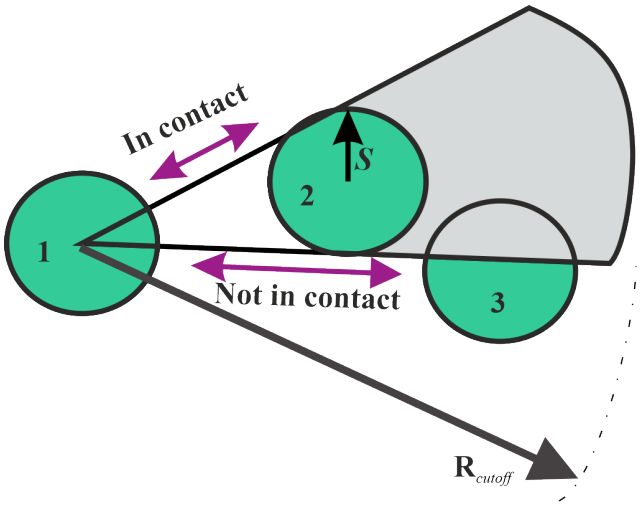


Figure 2.7.: Shadow contact algorithm. All atoms are given a shadowing radius S. To determine the contacts, atoms within cutoff radius R_{cutoff} are considered. All atoms within the cutoff radius that also have a shadow cast upon them due to an intervening atom are discarded.

are introduced. The shadow contact map algorithm circumvents this limitation through introducing a screening term S, which is a radius given to each atom, along with the cutoff radius [82]. First, atom pairs within the cutoff are considered possible native contacts. Then, those that have an intervening atom are screened out (see Fig. 2.7). Contact pair formation in proteins are often visualized using contact maps (see Fig. 2.8). A contact map is a 2D binary matrix that provides a reduced representation of a protein's 3D structure.

2.2.1.3. Units in Structure-based Models

MD simulations with SBMs are run with reduced units. Reduced units are dimensionless units that are often defined based on the Lennard-Jones potential parameters. The length, time, mass and energy scales are all set to 1. In GROMACS [84], however, the default set of units are nm length scale, ps time scale, amu mass scale, and kJmol^{-1} energy scale. Reduced units can be used directly in GROMACS with the exception of temperature, which is expressed in 0.008314 (i.e. the Boltzmann's constant, k_B) reduced units. Therefore, the reduced temperature of 1 is equivalent to ~120 GROMACS temperature. In SMOG webtool, the input PDB structure file contains length scales in Å which is easily converted into nm, but the time scale, mass scale and energy scale are 'free'.

While it is possible to convert the scales back from reduced units to physical units, this can be quite challenging and thus should be done with care. Time scales can be estimated from comparisons to experimental observations such as folding rates and rotational correlation

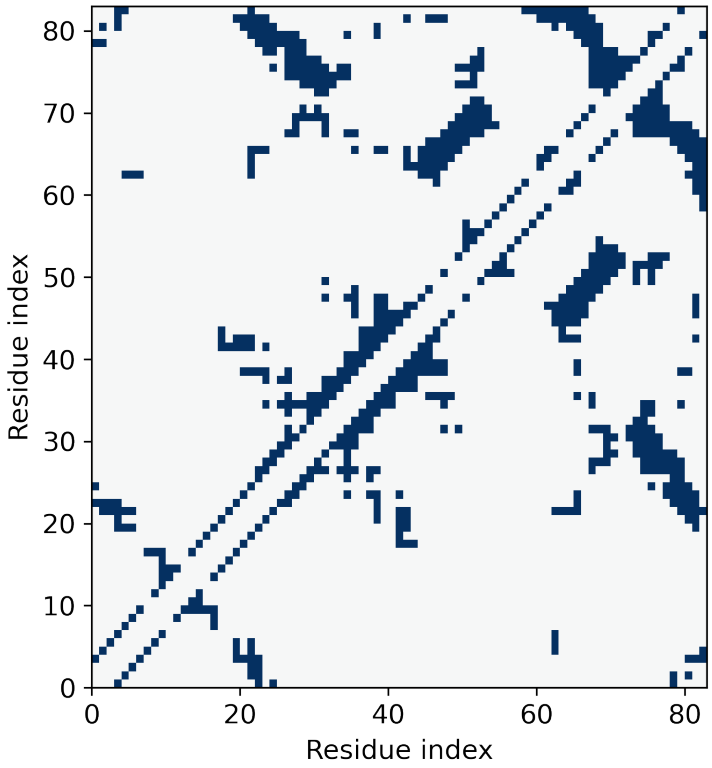


Figure 2.8.: Contact map for the serine proteinase inhibitor, chymotrypsin inhibitor 2 (CI-2) (PDB ID: 2CI2 [83]). Each point in the contact map represents the presence (black square) or absence (no marking) of a contact formed between two residues.

times [80]. It is worth noting that these time estimates do not necessarily take into account solvent friction effects resulting from the use of Langevin dynamics in the absence of water molecules. It is also essential to choose an appropriate temperature. The necessary simulation temperature can be determined by comparing atomic root mean squared fluctuation (RMSF) values between the SBM and the corresponding all-atom model in explicit solvent simulations, or experimental B-factors [85].

2.2.2. Multiple Basin Models

Force-fields containing structural information of a single conformation can be referred to as single-basin models. The conformational transitions along a pathway between known states can be explored through combining the individual single-basin potentials to form a multiple basin potential. Many studies have successfully demonstrated conformational transitions between two states by mixing two single-basin potentials to form a dual-basin

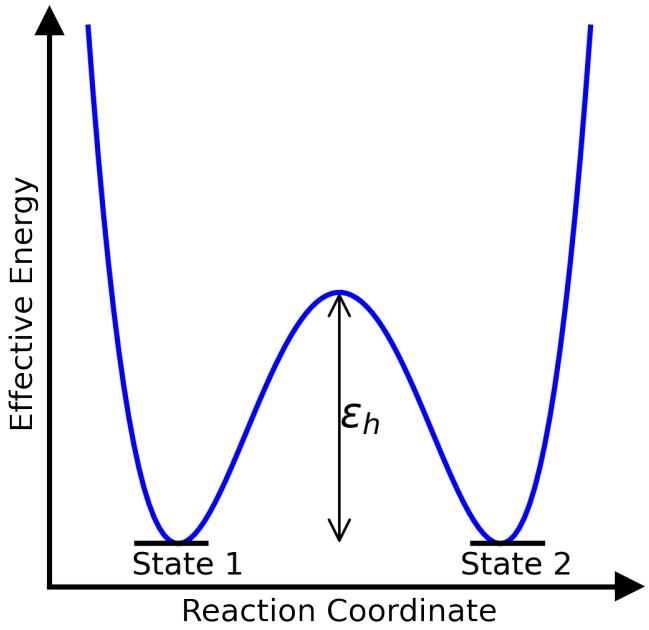


Figure 2.9.: Double-well potential. The energy landscape between two local minima with a barrier height, ε_h, can be obtained by mixing the single-basin potentials of each conformer.

[86–92]. Kinases, namely adenylate [88–90] and Src kinase [93], have been extensively studied using dual-basin models.

The mixed potentials can be obtained using two main approaches. In the first approach, often known as micro-mixing [89], contact energies of one state are added to the single-basin potential of the other state as perturbations. Conversely, in macro-mixing all energetic contributions of the two single-basin potentials are coupled to each other in the mixed potential [86]. The micro-mixing approach is suitable for systems in which the two states have a high degree of similarity such that they can be described primarily by the addition of contacts. In systems with little overlap between native contact sets, the macro-mixing approach would be more suitable. Two widely used macro-mixing approaches in combining potentials are the exponential Boltzmann weighing [94] and the superposition method [95]. Both approaches ensure a smooth transition between the basins, whilst retaining their respective minima (see Fig. 2.9).

2.2.2.1. Macro-Mixing Potentials: Exponential Boltzmann Weighing

With the exponential Boltzmann weighing method, a dual-basin potential is constructed by the summation of the partition functions of the two single-basin potentials. The corresponding partition functions are,

$$Z_1 = \int d\mathbf{R} \exp\left(-\beta E_1(\mathbf{R})\right), \tag{2.25}$$

$$Z_2 = \int d\mathbf{R} \exp\left(-\beta E_2(\mathbf{R})\right), \tag{2.26}$$

and the new potential function for N potential surfaces $E_i(R)$ would be:

$$\exp\left(-\beta E(\mathbf{R})\right) = \sum_{i=1}^{N} \exp\left(-\beta(E_i(\mathbf{R}) + \varepsilon_i)\right) \tag{2.27}$$

In this expression, $E_i(\mathbf{R})$ are the single-basin potentials, β is a mixing parameter and ε_i is an energetic offset applied to E_i. The ε_i parameter calibrates the relative stability of the states, and β calibrates the barrier height and hence the conformational transition rate. Energy surfaces can be taken from SBMs, elastic models, or all-atom potentials. This scheme has been applied to the adenylate [92], Src [93], and cyclin-dependent kinases [96].

2.2.2.2. Macro-Mixing Potentials: Superposition Method

In the superposition approach, Clementi et al.'s[78] version of the off-lattice Gō potentials are combined by solving an eigenvalue equation [95]. Firstly, two single-basin potentials, $V(R|R1)$ and $V(R|R_2) + \Delta V$, are constructed. Here, R represents the coordinates of the protein structure, while R_1 and R_2 correspond to the coordinates of the two fiducial structures. An additional term, ΔV, is introduced to calibrate the relative stability of the two basins. The smoothed dual-basin potential, V_{DB}, is defined as the eigenvalue of the following equation:

$$\begin{pmatrix} V(R|R_1) & \Delta \\ \Delta & V(R|R_2) + \Delta V \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = V_{DB} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{2.28}$$

where (c_1, c_2) is the eigenvector and Δ is a coupling constant which modifies the energy barrier. This leads to the secular equation:

$$\begin{vmatrix} V(R|R_1) - V_{DB} & \Delta \\ \Delta & V(R|R_2) + \Delta V - V_{DB} \end{vmatrix} = 0 \tag{2.29}$$

The lower-energy solution is used as the dual-basin potential:

$$V_{DB} = \frac{V(R|R_1) + V(R|R_2) + \Delta V}{2} - \sqrt{\left(\frac{V(R|R_1) - V(R|R_2) - \Delta V}{2}\right)^2 + \Delta^2} \tag{2.30}$$

In the initial studies [95], double-well energy landscapes for glutamine-binding protein, S100A6 (a structural analogue of calmodulin), dihydrofolate reductase, and HIV-1 protease were successfully computed such that a reversible transition was obtained. The possibility to modulate the barrier heights allowed the "basin-hopping" similarly as seen with exponential Boltzmann weighing.

2.2.2.3. Micro-Mixing Potentials: Contact Perturbations

The introduction of local perturbations characteristic of a specific conformation into the force-field of another conformer, is sometimes referred to as micro-mixing. The perturbations are native contact energies of the second known state, and in some cases, dihedral angle terms are also introduced [89, 97]. This framework is employed when the known conformations have only a few overlapping native contact pairs. The formulation of an all-atom dual-basin structure-based potential by Singh et al [97], is defined as:

$$V_{DB} = V_{backbone} + V^{LJ}_{mixed} + V^{\phi}_{mixed}, \tag{2.31}$$

where $V_{backbone}$ encompasses the standard bonded potentials for bond stretching, angle bending and improper dihedral angles introduced and defined in section 2.1.4. The mixed Lennard-Jones potential for conformations α and β is,

$$
\begin{aligned}
V_{LJ} =& \lambda_\alpha \sum_{contacts}^{\alpha} \epsilon_C^\alpha \left[\left(\frac{\sigma_{ij}^\alpha}{r^\alpha} \right)^{12} - 2 \left(\frac{\sigma_{ij}^\alpha}{r^\alpha} \right)^{6} \right] \\
&+ \lambda_\beta \sum_{contacts}^{\beta} \epsilon_C^\beta \left[\left(\frac{\sigma_{ij}^\beta}{r^\beta} \right)^{12} - 2 \left(\frac{\sigma_{ij}^\beta}{r^\beta} \right)^{6} \right] \\
&+ \sum_{noncontacts}^{\alpha\beta} \epsilon_{NC}^{\alpha\beta} \left[\left(\frac{\sigma_{ij}^{\alpha\beta}}{r^{\alpha\beta}} \right)^{12} \right] \\
&+ \sum_{mixed}^{\alpha\beta} (\lambda_\alpha \epsilon_C^\alpha - \lambda_\beta \epsilon_C^\beta) \left[\left(\frac{\sigma_{ij}^{\alpha\beta}}{r^{\alpha\beta}} \right)^{12} - 2 \left(\frac{\sigma_{ij}^{\alpha\beta}}{r^{\alpha\beta}} \right)^{6} \right].
\end{aligned}
\tag{2.32}
$$

Here, there are separate LJ terms for α-only, β-only and $\alpha\beta$ shared native contacts. Non-native contact pairs are only given a repulsive term with energetic weight, ε_{NC}. Additional weighting parameters, λ_α and λ_β, are introduced to modulate the depths of the basins. Therefore, these weights induce the desired conformational transition, while allowing the control of the residency time at a given basin. The same weighting parameters are also used to scale the backbone and side-chain proper dihedral potential energy contributions to the force-field. The mixed proper dihedral potential is,

$$
\begin{aligned}
V_{mixed}^{\phi} =& \sum_{backbone}^{\alpha} \lambda_\alpha \varepsilon_{BB} F_D \phi \\
&+ \sum_{backbone}^{\beta} \lambda_\beta \varepsilon_{BB} F_D \phi \\
&+ \sum_{sidechain}^{\alpha} \lambda_\alpha \varepsilon_{SC} F_D \phi \\
&+ \sum_{sidechain}^{\beta} \lambda_\beta \varepsilon_{SC} F_D \phi,
\end{aligned}
\tag{2.33}
$$

where ε_{BB} and ε_{SC} are the backbone and side-chain energetic weights, respectively. The term, $F_D\phi$, was defined previously in equation 2.24. A similar micro-mixing framework is adopted in this thesis to construct a dual-basin structure-based model of WalK histidine kinase.

2.2.3. Reconstruction of Atomistic Models from Coarse-grained Models

In the present work, a structure-based potential was modified to form a dual-basin SBM of histidine kinase. When developing a new coarse-grained model, it is essential to evaluate its accuracy through comparison with either experimental data and/or high-resolution atomistic simulations. The process of recovering the all-atom protein structure can be divided into two main stages: backbone reconstruction from C-α positions and side chain reconstruction from the protein backbone. Protein backbone prediction algorithms have been developed that are based on libraries of peptide fragments from known structures [98, 99], analytical methods [100], and/or statistical propensities [101]. Most side-chain prediction methods use a rotamer library of known discrete side-chain conformations and a search algorithm for efficient sampling of the conformational space to find the global energy minimum [102]. I have chosen to use PD2 ca2main [98] for the initial step, followed by side chain reconstruction with SCWRL4 [103].

PD2 ca2main [98] uses a library of short peptide backbone fragments obtained using Gaussian mixture models, to predict the positions of the backbone atoms. To further refine the backbone atom placement, an optional energy minimization step is provided by the server. SCWRL4 [103] is a widely used protein side-chain prediction algorithm which uses a backbone-dependent rotamer library derived from kernal density estimates.

2.3. Analysis of Molecular Dynamics Trajectories

2.3.1. Principal Component Analysis (PCA)

Biomolecular simulations using MD yield several configurations of the test system along the conformational space. Given the sheer amount of simulation data, and the many degrees of freedom to account for, obtaining a concise but accurate interpretation of the system's dynamics poses a challenge. To address this issue, many analytical techniques have been developed including a variety of dimension reduction methods [104, 105] and Markov state models (see section 2.3.2) [106, 107]. One powerful and widely used dimension reduction technique that can be applied to MD trajectories is principal component analysis (PCA) [108]. In the context of protein simulations, PCA is also often referred to as Essential Dynamics (ED), since the aim is to extract the "essential" motions from the set of sampled

conformations [109, 110]. Defining the essential subspace using PCA has been applied with great success in protein functional and folding studies (e.g. in kinases [111–114]), enhanced sampling techniques [115, 116], and in the construction of peptide folding free energy landscapes [117]. In recent work by Khan et. al [118], PCA of different mutants of the N-terminal RNA binding domain (N-NTD) of SARS-CoV-2 provided insights into the key residues for the dynamics and binding of RNA (see Fig. 2.10).

The correlated motions of a molecule with N atoms can be described by a covariance matrix [119],

$$C_{ij} = \left\langle M_{ii}^{\frac{1}{2}}(x_i - \langle x_i \rangle) M_{jj}^{\frac{1}{2}}(x_j - \left\langle x_j \right\rangle) \right\rangle, \tag{2.34}$$

where M is a diagonal matrix containing the masses of the atoms in mass-weighted analysis, or a unit matrix in the case for non-mass weighted analysis. The covariance matrix C is a $3N \times 3N$ matrix, which can then be diagnolized with an orthonormal transformation matrix R:

$$R^T C R = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_{3N}), \quad \text{where } \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{3N}. \tag{2.35}$$

Diagonalization of this covariance matrix yields the eigenvectors (also called principal or essential modes), which are the columns of R, and a set of eigenvalues, λ. This describes the principal modes of the collective motion and their respective amplitudes. The principal components $p_i(t)$ can be obtained by projecting the MD trajectory on the principal modes:

$$\mathbf{p}(t) = R^T M^{\frac{1}{2}}(\mathbf{x}(t) - \langle \mathbf{x} \rangle). \tag{2.36}$$

The eigenvalue λ_i is the mean square fluctuation of principal component i. To extract and visualize the collective motions, the trajectory can be filtered along one or more principal modes. This enables us to filter the global and often slow motions from the local, fast motions. Therefore, we are able to extract the functionally relevant motions from our simulations.

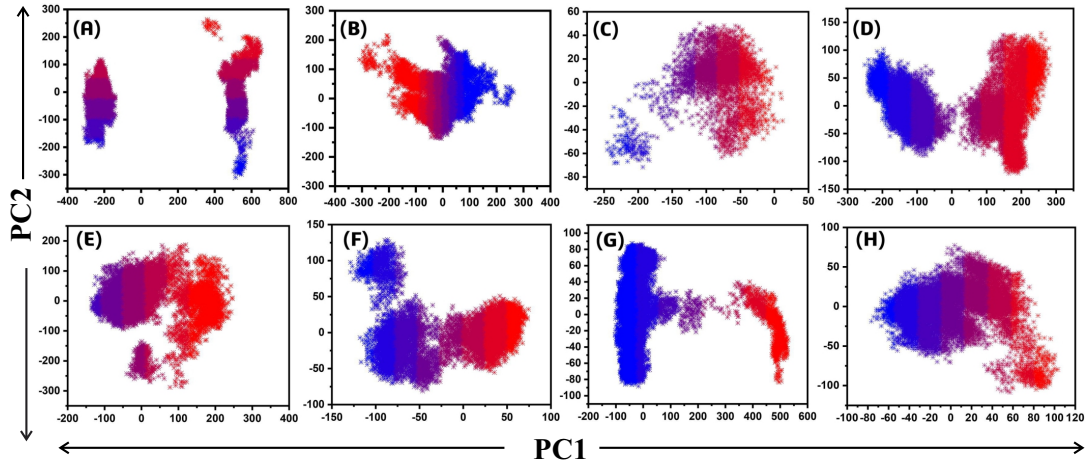


Figure 2.10.: Principal component analysis (PCA) of wild type (WT) and mutant N-terminal RNA binding domain (N-NTD) of SARS-CoV-2. (A) WT (B) T57A (C) H59A (D) S105A (E) R107A (F) G170A (G) F171A (H) Y172A. PC1 and PC2 from the PCA of the backbone carbon were used. Reprinted with permission from Ref. [118] under the CC BY-NC-ND 4.0 license (relabelled from original).

2.3.2. Markov State Models

Markov State Models (MSMs) is a kinetic network model for analyzing dynamical systems. This method provides a framework for predicting the long-timescale statistical conformational dynamics using data obtained from MD simulations. MSM approaches have been successfully used to address a wide variety of complex problems in biophysics. This includes the dynamics of intrinsically disordered peptides [120], protein folding [121], and protein-ligand binding processes [122].

An MSM consists of a set of microstates and a transition matrix. A microstate is set of related conformations which are clustered based on structural or kinetic similarity. In this work, I used the k-means clustering algorithm implemented in PyEMMA [123]. Firstly, the algorithm randomly places a predefined number of initial k cluster centroids. Each data point is then assigned to its nearest centroid. Next, the centroids are recalculated by minimizing the sum of squared deviations between the centroid and each of the observations in the cluster. These steps are iterated until convergence is reached.

The conformational state space has been discretized into trajectories, $s(t)$, jumping between n microstates. The molecular kinetics between these discrete states can be described by conditional transition probabilities:

$$P_{ij}(\tau) = \mathbb{P}(s(t+\tau) = j | s(t) = i) \tag{2.37}$$

Or in matrix form for longer time scales,

$$\mathbb{P}(s(t + k\tau) = j|s(t) = i) = [\mathbf{P}^k(\tau)]_{ij}. \tag{2.38}$$

Here, τ is the lagtime, k is an integer larger than 1, and i and j are arbitrary states. The system is "memoryless" and thus the lag time is considered Markovian. This means that the probability of a transition between two discrete states does not depend on where the system was before the initial state. Therefore, multiple short independent trajectories can be used.

For a Markov state model to be valid, two assumptions are made. Firstly, the MSM must be ergodic which means that the model does not have two or more dynamically disconnected states. As a result, for $t \rightarrow \infty$ each state in a trajectory is visited infinitely often. The second assumption is known as detailed balance. The system is in equilibrium and so the fraction of the system transitioning from i to j is equal to that of the reverse transition from j to i. Using the stationary probabilities of discrete states, π_i, MSM predicts the stationary distribution of $\mathbf{P}$ for any τ:

$$\pi^{\mathrm{T}} = \pi^{\mathrm{T}}\mathbf{P}(\tau). \tag{2.39}$$

Moreover, the accuracy of an MSM is highly dependent on the lag time. An appropriate lag time must be chosen such that the implied relaxation time scale is approximately constant within statistical uncertainty. The relation between the implied relaxation time scale t_i and lag time τ is given by the following expression:

$$t_i(\tau) = -\frac{\tau}{\ln|\lambda_i(\tau)|} \tag{2.40}$$

where $\lambda_i(\tau)$ is the ith largest eigenvalue at τ. The choice of lag time and hence the quality of the MSM, can be evaluated by the Chapman-Kolmogorov (CK) test [107]. The CK test compares the left- and right-hand sides of the Chapman-Kolmogorov equation,

$$\mathbf{P}(k\tau) = \mathbf{P}^k(\tau). \tag{2.41}$$

Here, $\mathbf{P}$ is the MSM transition matrix. The left-hand side of the equation corresponds to the MSM transition matrix at lag time $k\tau$, whereas the right-hand side of the equation is

an estimated MSM to the kth power. Markovianity is indicated by a good agreement of both sides of the equation within a statistical uncertainty.

2.3.2.1. Robust Perron Cluster Analysis (PCCA+)

An MSM that accurately approximates the statistical dynamics of a molecular system usually consists of hundreds to thousands of microstates. To obtain an interpretable kinetic model and thus gain a better intuition for a system, coarse-graining may be neccessary. In this context, coarse-graining refers to the process of "lumping" the MSM microstates into fewer metastable (or long-lived) states that capture the conformational changes.

Robust Perron Cluster Analysis (PCCA+) is a fuzzy spectral clustering algorithm, which can be used for coarse-graining a Markov state model [124, 125]. PCCA+ assigns the states $i \in \{1, ..., N\}$ to clusters $j \in \{1, ..., n_c\}$ each given a weighted membership $\chi_j(i) \in [0, 1]$. We refer to a set of vectors $\left\{\chi_j\right\}_{j=1}^{n_c}$ with $\chi_j \in \mathbb{R}^N$ as membership vectors that possess the following properties:

$$\chi_j(i) \geq 0 \quad \forall_i \in \{1, ..., N\}, \quad j \in \{1, ..., n_c\} \quad (positivity) \tag{2.42}$$

$$\sum_{j=1}^{n_c} \chi_j(i) = 1 \quad \forall_i \in \{1, ..., N\} \quad (partition\ of\ unity) \tag{2.43}$$

PCCA+ computes the non-singular transformation matrix $A \in R^{n_c \times n_c}$, which is related to the transformed vectors $\chi = \left[\chi_1, ..., \chi_j\right]$ by the following expression:

$$\chi = XA. \tag{2.44}$$

In this thesis, coarse-graining by PCCA+ was used to analyze the dual-basin MD simulations. The stationary probabilities of the metastable states (i.e. the two free energy basins) were calculated to deduce the likelihood of finding the system in the active or inactive states.

2.3.3. Other Quantitative Metrics

Root-mean-square deviation (RMSD)

Root-mean-square deviation (RMSD) is a frequently used quantitative measure of the similarity between two molecular conformations. After a least square fitting to the reference structure, the RMSD is calculated by:

$$RMSD(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|r_i(t) - r_{i,0}\right|^2}. \tag{2.45}$$

Here, N is the number of atoms, $r_i(t)$ is the position of atom i at time t, and $r_{i,0}$ is the position of atom i in the reference structure. In the context of proteins simulations, all of the atoms are not typically used in the RMSD calculations, nor for the least square fitting. Instead, proteins are more often fitted to and calculated for the atoms of the backbone (N, C_α, C) only. Monitoring the time evolution of the RMSD along the trajectory with respect to the initial conformation, gives an indication into the stability of the molecule. A limitation in the metric lies in the degeneracy at higher values. What this means is that many different structures can have the same RMSD from the same reference. Alternative RMSD-based metrics such as 2D RMSD, $\Delta RMSD$, and distance RMSD (dRMSD) can be used for more precision.

Single-structure RMSD and other RMSD-based metrics have also been applied in enhanced sampling methods to drive the starting conformations to a known target structure [126–128]. These methods enforce these transitions irrespective of the height of the energy barriers. I used umbrella sampling in combination with steered molecular dynamics (see sec. 2.4.1 and 2.4.2) to compute the activation pathways and free energy profiles of two kinases.

Radius of Gyration (R_g)

Another important quantity that is commonly used to analyze trajectories is the radius of gyration (R_g). The radius of gyration can be used to determine the compactness of a protein [129]. GROMACS determines R_g by the following equation [84]:

$$R_g = \left(\frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i} \right)^{\frac{1}{2}}. \tag{2.46}$$

Here, m_i is the mass of atom i and r_i is the position of atom i with respect to the molecule's center of mass. Small values of R_g indicate a compact conformation, whereas larger values suggest that the structures are less tightly packed. The allosteric mechanism in some proteins involve the switching between an "open" unligated state, and a "closed" conformation with a ligand bound. The available crystal structures for histidine kinases indicate that such a switching mechanism occurs as a response to ATP binding. Thus, I monitored the progress of the transitions using R_g.

2.4. Free Energy Calculations

Free energy calculations provide essential insights into the functions and structural properties of molecular systems. The free energy difference is the driving force for processes including large-scale conformational changes, protein-ligand binding processes and chemical reactions. There are two types of free energies which are dependent on the type of statistical ensemble used in the computer simulation.

The free energy associated with the canonical ensemble (NVT) of a system is known as Helmholtz free energy A. Helmholtz free energy is related to the canonical partition function Q_{NVT} by the following expression:

$$A(N, V, T) = -\frac{1}{\beta} \ln Q_{NVT}, \tag{2.47}$$

where,

$$Q_{NVT} = \int \exp\left[-\beta E(r)\right] dr. \tag{2.48}$$

Here, E is the potential energy, $\beta = 1/(k_B T)$, k_B is the Boltzmann's constant, T is the absolute temperature, and N is the number of degrees of freedom. If the pressure is constant rather than the volume, we have an isothermal-isobaric ensemble (NPT). Similarly to A, Gibbs free energy G is the free energy related to the isothermal-isobaric partition function Q_{NPT} by,

$$G(N,P,T) = -\frac{1}{\beta}\ln Q_{NPT}, \tag{2.49}$$

$$Q_{NPT} = \int Q_{NVT}\exp\left[-\beta PV\right] dV. \tag{2.50}$$

Despite the change in ensemble, ΔA and ΔG are numerically very similar in the condensed phase, which is relevant in most applications. In biomolecular simulations, however, it is more convenient to calculate the potential of mean force (PMF). The PMF of a system is the free energy surface along a chosen reaction coordinate or collective variable (ξ). A collective variable (CV) is a continuous parameter that ideally provides a clear distinction between thermodynamic states, and hence describes the transition of interest.

The probability distribution along a defined CV, can be calculated by integrating out all degrees of freedom,

$$Q(\xi) = \frac{\int \delta\left[\xi(r) - \xi\right]\exp(-\beta E)d^N r}{\int \exp(-\beta E)d^N r}, \tag{2.51}$$

where $\delta\left[\xi(r) - \xi\right]$ is the Dirac delta function. This direct phase-space integral is impossible to calculate with computer simulations. However, for an ergodic system where all possible configurations are sampled, $Q(\xi)$ is equivalent to,

$$P(\xi) = \lim_{t\to\infty}\frac{1}{t}\int_0^t \rho\left[\xi(t')\right] dt'. \tag{2.52}$$

In other words, the ensemble average $Q(\xi)$ and time average $P(\xi)$ become equal for infinite sampling in an ergodic system. In equation 2.52, t denotes the time and ρ counts the occurrence of ξ in a given interval. In principle, the changes in free energy can be obtained by monitoring $P(\xi)$ in an MD simulation. The Helmholtz free energy can be reformulated as a function of $P(\xi)$ to obtain the change in PMF between arbitrary states a and b:

$$\Delta A = A(\xi_b) - A(\xi_a) = -\frac{1}{\beta}\ln\frac{P(\xi_b)}{P(\xi_a)}. \tag{2.53}$$

However, MD simulations are run for a finite amount of time and so some regions in configuration space are left unexplored. Typically, regions around the energy minimum are

well-sampled, whereas higher energy regions are under-sampled. Insufficient sampling and hence unconverged biomolecular simulations result in inaccurate free energy calculations. To circumvent this problem, many enhanced sampling techniques have been developed. These approaches can be divided into two categories: one type is based on the addition of bias potentials along the predefined CVs of the system (CV-based), while the other approaches are not dependent on a predefined CV (CV-free). Widely used CV-based enhanced sampling methods include umbrella sampling (US) [130], steered molecular dynamics (SMD) [131], metadynamics (MetaD) [132] and temperature accelerated MD (TAMD) [133]. If the CV is well selected, the convergence and accuracy of free energy calculations are largely improved. However, choosing an appropriate CV is not a trivial task for many complex systems. In this case, CV-free enhanced sampling methods such as replica exchange (RE) [134], accelerated molecular dynamics (aMD) [135], or self-guided molecular Langevin dynamics (SGLD) [136] could be used instead.

In this work, I used the CV-based enhanced sampling methods umbrella sampling and steered MD, as we have some priori knowledge of our system. Specifically, the crystal structures of the inactive and activate states of histidine kinase are available in the Protein Data Bank. The initial atomic positions taken from the crystal structures of the end states can be used to drive the transition with an RMSD-based biasing potential.

2.4.1. Steered Molecular Dynamics

Steered molecular dynamics (SMD) is an enhanced sampling method inspired by atomic force microscopy (AFM) experiments [131]. The system is driven from an intial configuration to a desire state by introducing a time-dependent, harmonic restraint. In SMD, the external bias potential ω_{SMD} with respect to the collective variable ξ is,

$$\omega_{SMD}(\xi) = \frac{k(t)}{2}\left[\xi(t) - \xi_0 - vt\right]^2, \tag{2.54}$$

where $k(t)$ is the force constant, v is the pulling velocity and t is time. As work is being applied to induce the conformational change from one state to another, the system is no longer in equilibrium. To obtain the equilibrium free energy difference, one could use the Jarzynski relation [137]. The Jarzynski equation is,

$$\Delta G_{AB} = -k_B T log \left\langle e^{-\beta W_{AB}} \right\rangle_A \tag{2.55}$$

where W is the work performed to force the transition from state A to B and the angular bracket denotes averaging over a canonical ensemble of the initial state. In order to obtain an accurate free energy profile, it is necessary perform the non-equilibrium SMD simulations with the same CV multiple times. An alternative protocol is to first perform an SMD simulation followed by umbrella sampling. SMD is used to generate intermediate configurations along the transition pathway. Each of these structures are then used as the initial configurations for independent US simulations. Further details are found in the next section.

2.4.2. Umbrella Sampling

Umbrella sampling (US) is one of the most widely used enhanced sampling methods for overcoming large free energy barriers [130]. The conformational transitions of a system is modelled by a series of independent simulations along a predefined CV ξ, as depicted in Fig. 2.11. These independent simulations are also known as windows. In each window, a bias potential that is solely dependent on the CV is applied to the system. The addition of a harmonic bias potential ω_i to window i results in a new biased potential energy $E^b(r)$:

$$E^b(r) = E^u(r) + \omega_i(\xi) \tag{2.56}$$

$$\omega_i(\xi) = \frac{k}{2}(\xi - \xi_i^{ref})^2 \tag{2.57}$$

where, $E^u(r)$ is the true unbiased potential energy, k is the force constant, and ξ_i^{ref} is the reference ξ value for window i. To ensure all regions of ξ are well sampled, an appropriate force constant must be chosen. The force constant must be large enough to overcome the energetic barrier, however, when k is too large, it results in very narrow distributions. Sufficient overlap in the probability distributions from umbrella windows is required for free energy calculations with the weighted histogram analysis method (WHAM). Therefore, it would be necessary to run additional windows to ensure overlap making this even more computationally expensive.

To obtain the unbiased free energy $A_i(\xi)$, we must first define the unbiased distribution P_i^u:

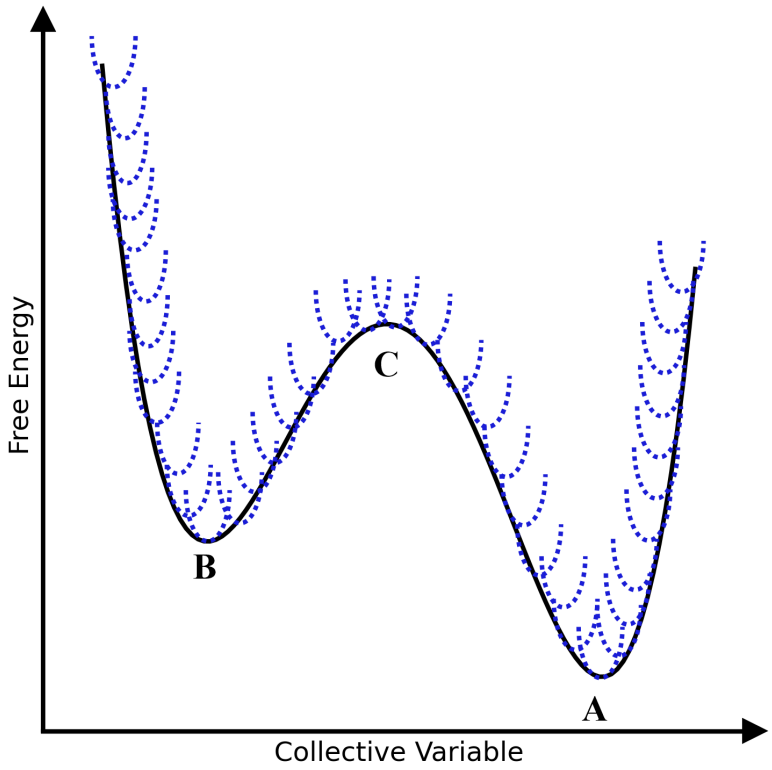


Figure 2.11.: Schematic illustration of the umbrella sampling method. The blue dotted lines represent the harmonic bias potentials that are added to the system's Hamiltonian at different windows along the collective variable space.

$$P_i^u(\xi) = \frac{\int \exp\left[-\beta E(r)\right] \delta\left[\xi'(r) - \xi\right] d^N r}{\int \exp\left[-\beta E(r)\right] d^N r}. \tag{2.58}$$

US simulations of an ergodic system provides the biased distribution P_i^b along the collective variable ξ,

$$P_i^b(\xi) = \frac{\int \exp\left\{-\beta\left[E(r) + \omega_i(\xi'(r))\right]\right\} \delta\left[\xi'(r) - \xi\right] d^N r}{\int \exp\left\{-\beta\left[E(r) + \omega_i(\xi'(r))\right]\right\} d^N r}. \tag{2.59}$$

The harmonic bias depends solely on ξ and thus the biased distribution can be reformulated as,

$$P_i^b(\xi) = \exp\left[-\beta\omega_i(\xi)\right] \times \frac{\int \exp\left[-\beta E(r)\right] \delta\left[\xi'(r) - \xi\right] d^N r}{\int \exp\left\{-\beta\left[E(r) + \omega_i(\xi'(r))\right]\right\} d^N r}. \tag{2.60}$$

Using equation 2.58, the unbiased distribution can be rewritten as,

$$\begin{aligned} P_i^u(\xi) =& P_i^b(\xi)\ \exp\left[\beta\omega_i(\xi)\right] \times \frac{\int \exp\left\{-\beta\left[E(r)+\omega_i(\xi(r))\right]\right\} d^N r}{\int \exp\left[-\beta E(r)\right] d^N r} \\ =& P_i^b(\xi)\ \exp\left[\beta\omega_i(\xi)\right] \times \frac{\int \exp\left[-\beta E(r)\right] \exp\left\{-\beta\omega_i[\xi(r)]\right\} d^N r}{\int \exp\left[-\beta E(r)\right] d^N r}. \\ =& P_i^b(\xi)\ \exp\left[\beta\omega_i(\xi)\right] \left\langle \exp\left[-\beta\omega_i(\xi)\right]\right\rangle \end{aligned} \tag{2.61}$$

This equation links the unbiased distribution with the biased distribution obtained from an MD simulation. Therefore, the unbiased free energy $A_i(\xi)$ can be evaluated by,

$$A_i(\xi) = -\frac{1}{\beta} \ln P_i^b(\xi) - \omega_i(\xi) + F_i \tag{2.62}$$

$$F_i = -\frac{1}{\beta} \ln \left\langle \exp\left[-\beta\omega_i(\xi)\right]\right\rangle . \tag{2.63}$$

The final step is to combine the independent umbrella windows to obtain the free energy landscape.

2.4.3. Weighted Histogram Analysis Method (WHAM)

The weighted histogram analysis method (WHAM) is a method for combining multiple independent umbrella sampling simulations [138, 139]. The global unbiased distribution P^u is calculated by a weighted average of the unbiased distributions of the individual windows P_i^u:

$$P^u(\xi) = \sum_i^{windows} p_i(\xi) P_i^u(\xi). \tag{2.64}$$

Here, p_i are weights with values chosen to minimize the statistical error of the global distribution:

$$\frac{\partial \sigma^2(P^u)}{\partial p_i} = 0 \tag{2.65}$$

under the condition $\sum p_i = 1$. This results in,

$$p_i = \frac{a_i}{\sum_j a_j}, \; a_i(\xi) = N_i \exp\left[-\beta\omega_i(\xi) + \beta F_i\right] \tag{2.66}$$

where N_i is the total number of steps sampled for window i. The F_i terms are calculated by:

$$\exp(-\beta F_i) = \int P^u(\xi) \exp\left[-\beta\omega_i(\xi)\right] d\xi. \tag{2.67}$$

Equations 2.66 and 2.67 are deemed the WHAM equations. The WHAM algorithm iterates these two equations until convergence is reached. The accuracy of the free energy profiles obtained with WHAM can be evaluated by Monte Carlo bootstrap error analysis.

3. Dual-basin Structure-based Model of WalK Histidine Kinase

3.1. Introduction

Despite the rapid increase in computing power, classical atomistic molecular dynamics simulations of biological systems are still limited to relatively small systems and biologically short time scales. To circumvent this limitation, simplified representations of complex systems with their interactions described by simplified force fields have been developed [63–66, 69, 70]. These are known as coarse-grained models. CG models have since been used to study important biological processes such as protein folding [75, 140, 141], conformational changes [76, 142], and allostery [73, 74, 143]. Native structure-based models (SBMs), a class of CG models based on the energy landscape theory and the principle of minimal frustration [67, 68], are used in this work.

In recent years, multiple-basin CG models that integrate structural information of more than one protein conformation have been introduced [86–92]. Kinases, such as adenylate [88–90], Src [93] and cyclin-dependent kinase [96], have been extensively studied using dual-basin CG models. In this chapter, a dual-basin structure-based model of a histidine kinase (HK) is introduced for the first time.

WalK is an essential homodimeric HK involved in the regulation of cell growth and division in low GC-content Gram-positive bacteria, such as *Bacillus subtilis* [144] and *Staphylococcus aureus* [145]. The crystal structures of the inactive and active conformations of the kinase core are available in the protein data bank (see Fig. 3.1). The kinase core consists of the catalytic ATP-binding (CA) domains and the dimerization and histidine phosphotransfer (DHp) domains. Following signal detection, a series of transient conformational changes occur resulting in an autophosphorylation reaction at the kinase core. The conformational

changes during HK activation, as well as the reverse transition, can be addressed using the dual-basin SBM.

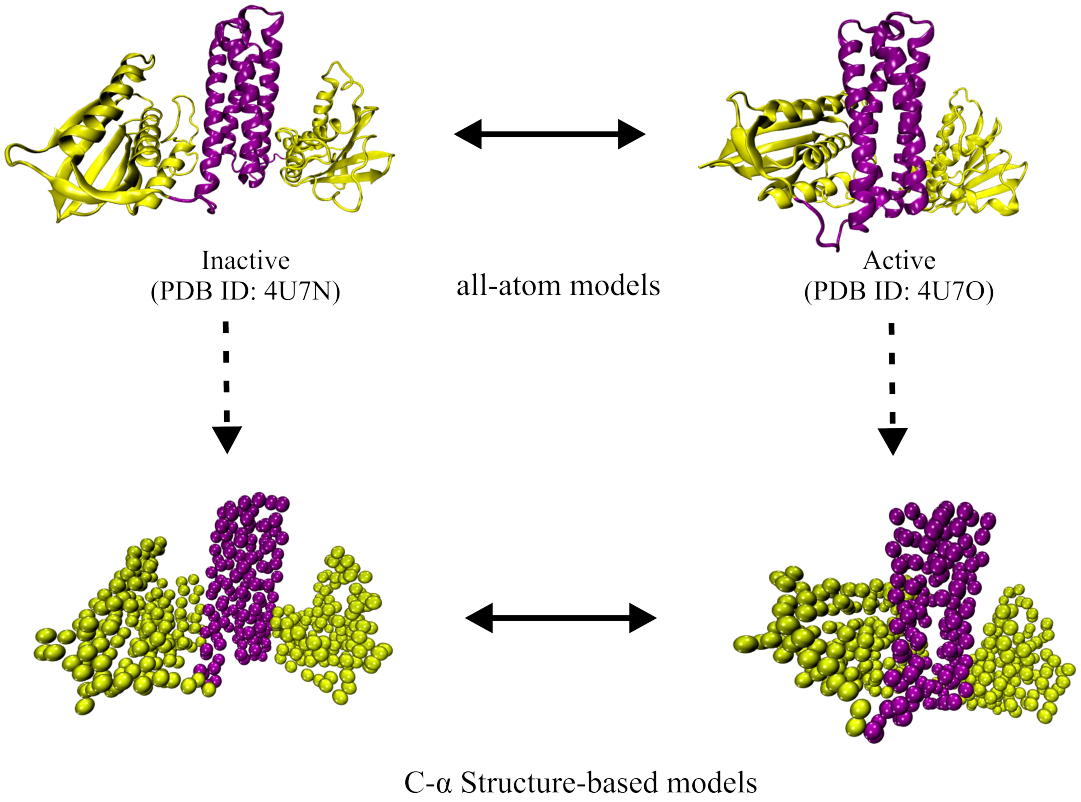


Figure 3.1.: Protein models of WalK histidine kinase. **Top:** Crystal structures of WalK in the inactive (PDB ID: 4U7N [31]) and active (PDB ID: 4U7O [31]). The catalytic ATP-binding (CA) domains and the dimerization and histidine phosphotransfer (DHp) domains are colored in yellow and purple, respectively. **Bottom:** Structured-based models of the inactive and active states generated using the SMOG server [81]. Each amino acid is represented by a single bead at the α-carbon position. The double headed arrow denotes the reversible conformational transition from one state to the other.

3.2. Methods

3.2.1. Dual-basin Structure-based Model Construction

The inactive and active states of histidine kinase WalK, an essential HK from Lactobacillus plantarum, were obtained from the protein data bank (PDB ID: 4U7N and 4U7O, respectively) [31]. The non-terminal missing residues were modelled using Chimera's [146] interface to MODELLER version 10.0 [147]. Structure-based models (SBMs) with each amino acid residue represented by a single bead at the alpha-carbon position were generated using the SMOG server (see Fig. 3.1) [81]. The functional form of the single-basin C-α

Hamiltonian was introduced in equation 2.23. Native contact maps were defined using the shadow contact map algorithm with a maximum contact distance of 0.6 nm and a shadowing radius of 0.1 nm.

Table 3.1.: Native contact analysis of the active and inactive states of histidine kinase

PDB ID/chain	Total number of native contacts	Shared contacts	Unique native contacts			
			Total	DHp-CA	DHp-DHp	CA-CA
4U7N/B	708	506	202	60	16	126
4U7O/B	682		176	79	20	77

The single-basin potential of the inactive state was perturbed by the addition of native contacts unique to the active state. The dual-basin structure-based potential is,

$$V_{DB} = V_{backbone} + V_{LJ}^{mixed}, \tag{3.1}$$

where,

$$V_{backbone} = \sum_{bonds}^{\alpha} \varepsilon_r^{\alpha} \left(r^{\alpha} - r_0^{\alpha}\right)^2 + \sum_{angles}^{\alpha} \varepsilon_{\theta}^{\alpha} \left(\theta^{\alpha} - \theta_0^{\alpha}\right)^2 + \sum_{backbone}^{\alpha} \varepsilon_D^{\alpha} F_D^{\alpha}\left(\phi\right) \tag{3.2}$$

$$\begin{aligned} V_{LJ}^{mixed} = & \sum_{contacts}^{\alpha} \varepsilon_C^{\alpha} \left[5 \left(\frac{\sigma_{ij}^{\alpha}}{r^{\alpha}} \right)^{12} - 6 \cdot \lambda_{\alpha} \left(\frac{\sigma_{ij}^{\alpha}}{r^{\alpha}} \right)^{10} \right] \\ & + \sum_{contacts}^{\beta} \varepsilon_C^{\beta} \left[5 \left(\frac{\sigma_{ij}^{\beta}}{r^{\beta}} \right)^{12} - 6 \cdot \lambda_{\beta} \left(\frac{\sigma_{ij}^{\beta}}{r^{\beta}} \right)^{10} \right] \\ & + \sum_{shared}^{\alpha\beta} \varepsilon_C^{\alpha} \left[5 \left(\frac{\sigma_{ij}^{\alpha}}{r^{\alpha}} \right)^{12} - 6 \left(\frac{\sigma_{ij}^{\alpha}}{r^{\alpha}} \right)^{10} \right] \\ & + \sum_{non-contacts}^{\alpha} \varepsilon_{NC}^{\alpha} \left(\frac{\sigma_{NC}^{\alpha}}{r^{\alpha}} \right)^{12}. \end{aligned} \tag{3.3}$$

The subscripts and superscripts 'α' and 'β' denote the inactive and active state quantities, respectively. The energetic weights, ε (in reduced units), are ε_r =20000, ε_θ = 40, ε_D = 1, ε_{NC} = 1, and ε_C = 1. The reference bond lengths, angles and proper dihedral angles are r_0, θ_0 and ϕ_0, respectively. A total of 176 native contact energies specific to the active state were added (see Table 3.1). As shown in equations 3.2 and 3.3, the shared native contacts and bonded interactions are given the weights and ground state values of the inactive state. To obtain a new potential that yields a reversible transition between the two states, two weighting parameters, λ_α and λ_β were introduced.

The optimal values of λ_α and λ_β were determined by running various SBM simulations with different combinations of these parameters (see Table 3.2). All SBM simulations were run with GROMACS 2019 [148] with stochastic dynamics, Langevin temperature coupling (coupling constant of 1.0), time steps of 0.002 (reduced units) and at GROMACS temperature 100. GROMACS temperatures are specified in 0.00831451 (i.e. the Boltzmann's constant, k_B) reduced units (see section 2.2.1.3). Therefore, GROMACS temperature 100 corresponds to 0.83 reduced units. When the optimal set of contact parameters were found, addition SBM simulations with GROMACS temperatures 98, 102, 104 and 106 were performed.

Table 3.2.: The combinations of native contact strengths used in the test simulations

Inactive native contacts weight, λ_α	Active native contacts weight, λ_β
0.8	1.0
0.8	0.9
0.8	0.85
0.8	0.825
0.8	0.8
0.8	0.5
0.825	0.825
0.825	0.85
0.85	0.85

The active-state-specific native contacts were further divided based on the region at which the interacting residues are located. Contact pairs can either be formed between atoms that are both located in the DHp domain, both in the CA domain or formed between one atom in the CA domain and the other in the DHp domain. Instead of giving all the active contact pairs the same weight, λ_β, each was given λ_{DHp}, λ_{CA} or λ_{CA-DHp}. Further SBM simulations were run with various combinations of λ_{DHp}, λ_{CA} and λ_{CA-DHp} at GROMACS temperature 98 (see Table 3.3).

The two-dimensional free energy profiles were computed as a function of RMSD with the inactive and active as the reference structures. Markov state models were estimated from the SBM simulations, using k-means clustering to assign the configurations into structurally similar microstates. Using PCCA+, the microstates were assigned to the two energetic basins (i.e. the metastable states) and the stationary probability of each basin was calculated [124]. MSM estimation and coarse-graining with PCCA+ were carried out using PyEMMA [123].

Table 3.3.: The combinations of native contact strengths of the different subgroups tested.

CA-DHp native contacts weight, λ_{CA-DHp}	CA-CA native contacts weight, λ_{CA}	DHp-DHp native contacts weight, λ_{DHp}
1.0	0.85	0.85
0.85	**1.0**	0.85
0.85	0.85	**1.0**
0.8	0.85	0.85
0.85	**0.8**	0.85
0.85	0.85	**0.8**
0.75	0.85	0.85
0.85	**0.75**	0.85
0.85	0.85	**0.75**

3.2.2. Reconstruction of All-Atom Models

Two structures were extracted from the dual-basin molecular dynamics simulation with the optimal parameters determined in the previous stage. One structure was in the active state while the other was in the inactive state. These structures only have the alpha carbons present and must be reconstructed via a two-stage process. Firstly, the backbone atoms were reconstructed using PD2 ca2main web-server [98], which uses a fragment library of 528 backbone fragments, obtained using Gaussian mixture models (GMM). Energy minimization was performed on the backbone-only model by the server. Next, the side chain and hydrogen atoms were added using SCWRL4 [103].

3.2.3. Umbrella Sampling of the Reconstructed All-Atom Models

3.2.3.1. System Setup

All MD simulations of the two reconstructed all-atom models were conducted at 300 K using GROMACS 2020.4 [84] patched with PLUMED v2.7 [149], with AMBER99SB-ILDN force-field [59], and with a time step of 2 fs. Each HK model was first solvated with the TIP3P water model [150]. The electrostatic interactions were calculated using the particle mesh Ewald (PME) method [62], with a direct cutoff of 1.0 nm and a grid spacing of 0.16 nm. A cutoff of 1.0 nm was used for the van der Waals interactions. All bonds involving hydrogen were constrained by the LINCS algorithm [49]. Each HK-water system was first minimised using the steepest decent algorithm for 1000 steps. This was followed by NVT equilibration with the velocity-rescaling thermostat [54] for 1 ns and then NPT equilibration with the Parrinello-Rahman barostat [55] for 1 ns.

3.2.3.2. Steered Molecular Dynamics

In order to obtain intermediate configurations along the pathway between the two states, steered MD was used. A time-dependent, geometrical constraint was applied to the inactive structure. Specifically, the root-mean-square deviation (RMSD) relative to the active state was used as a collective variable. The functional form of the potential is:

$$V_{SMD} = \frac{k}{2} \left(RMSD(t) - RMSD^* \right)^2 \tag{3.4}$$

where k is the spring constant, $RMSD(t)$ is the instantaneous RMSD between the current coordinates and the reference structure. RMSD* evolves linearly from the initial RMSD at the first steered MD step to the final target RMSD. Histidine kinase activation to the target structure was guided by the means of steering forces with a spring constant of 5000 $\mathrm{kJmol^{-1}nm^{-2}}$ applied to the α-carbons of the DHp domain of both chains and the β-sheet regions of the active protomer's CA domain. Starting from the equilibrated inactive HK in explicit TIP3P system, a steered MD simulation was run for 2ns.

3.2.3.3. Umbrella Sampling

A total of 32 conformations along the collective variable (i.e. RMSD) were extracted from the steered MD trajectory. The RMSDs ranged between 0.225 to 0.985 nm. Each of these conformations were used as the starting structure of an independent umbrella sampling simulation known as an umbrella window. The following harmonic bias was applied to each umbrella window i:

$$V_i = \frac{k}{2} \left(RMSD - RMSD_i^{ref} \right)^2 . \tag{3.5}$$

Here, k is the spring constant, $RMSD$ is the RMSD of the current coordinates and $RMSD_i^{ref}$ is the reference RMSD of the umbrella window i. The alpha carbons of both DHp protomers and the beta regions of the CA domain of the active protomer were considered in the RMSD calculations. A spring constant of 10000 $\mathrm{kJmol^{-1}nm^{-2}}$ was applied to each umbrella window. Each umbrella window was simulated for 30 ns. The free energy profile was obtained using the weighted histogram analysis method (WHAM) and the statistical errors were computed by Monte Carlo bootstrap error analysis (50 bootstrap samples) implemented in the WHAM code [151].

3.2.4. Umbrella Sampling of the Crystal All-Atom Models

The crystal structures of the inactive and active states of histidine kinase WalK were obtained from the protein data bank (PDB ID: 4U7N and 4U7O respectively) [31]. MD simulations of each of the two states were performed using GROMACS 2020.4 [50] patched with PLUMED v2.7 [149], with AMBER99SB-ILDN force-field at 300K [59]. Each HK conformation was solvated, minimized and equilibrated using the same protocol as detailed in section 3.2.3.1.

Histidine kinase was driven from the inactive state to the active state by steered MD and the intermediate structures were extracted from the trajectory. The procedure is the same as detailed in section 3.2.3.2 for the reconstructed all-atom models. The harmonic bias in equation 3.5 was implemented in the umbrella sampling simulations. A total of 26 intermediates which have roughly a 0.025nm spacing were used as the starting structures for the umbrella windows. An additional window centred at 0.66 nm was simulated to ensure sufficient sampling in that region of the conformational space. Each umbrella window was simulated for 30 ns with a spring constant of 10000 $\text{kJmol}^{-1}\text{nm}^{-2}$. Exceptions are the windows centred at 0.275 nm and 0.675 nm which had a spring constant of 15000 $\text{kJmol}^{-1}\text{nm}^{-2}$. This was to speed up the time of convergence. The free energy profiles were obtained using WHAM and the statistical errors were computed by Monte Carlo bootstrap error analysis (50 points) implemented in the WHAM code [151].

3.3. Results and Discussions

3.3.1. Determining the Dual-basin Structure-based Potential Parameters

Two x-ray crystal structures of the kinase core region of WalK histidine kinase are available. One of the kinases is in the inactive symmetrical conformation, while the other is in the active asymmetrical conformation. To study the conformational transitions between the two states, a dual-basin structure-based model was constructed using a micro-mixing of approach. Firstly, a single-basin SBM for each of the two known conformation was defined. In these coarse-grained models, each amino acid is represented as a single bead at the α-carbon position. This reduction in the number of degrees of freedom accelerates the computation speed. Furthermore, the smoothing of the protein's energy landscape

is achieved by categorizing all non-bonded interaction pairs as native or non-native contacts. Native contact pair interactions are favourable, whereas non-native contact are unfavourable and described by a repulsive term. By comparing the native contact pairs list of the two SBMs, it was found that there are 176 contact pairs unique to the active state (see Table 3.1). I introduced these active-state-specific native contact energies as perturbations into the single-basin structure-based potential of the inactive state.

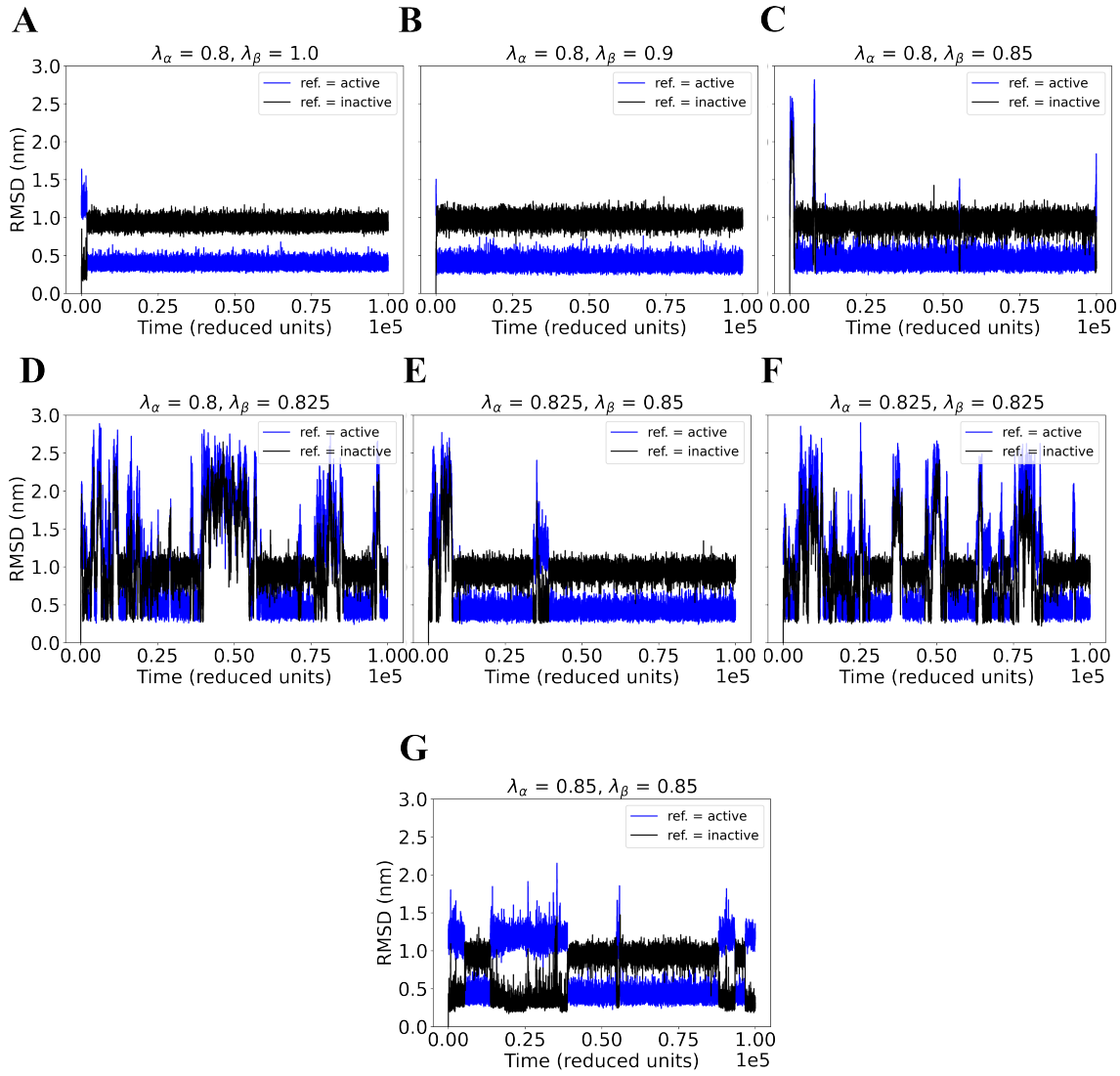


Figure 3.2.: Time evolution of the root mean square deviation (RMSD) along the different test simulations. RMSD was calculated with the inactive (black) and active (blue) α-carbons as the reference atoms.

The first objective was to determine the appropriate relative weights of the native contact energies of the two conformations, in order to induce a reversible conformational transition. I performed nine independent SBM simulations with different combinations of λ_α and λ_β (see Table 3.2). To monitor the progress of the simulations, I calculated the RMSD with

respect to the α-carbon atom positions of the active and inactive states (see Fig. 3.2). In seven of these simulations, HK had been activated as indicated by the decrease in RMSD with respect to the active state. A single transition occurred almost immediately at the beginning of test simulations A (λ_α = 0.8 and λ_β = 1.0, Fig. 3.2A) and B (λ_α = 0.8 and λ_β = 0.9, Fig. 3.2B). The system then stabilises in the active state. As we cannot gain insight into the reverse transitions, these are not the appropriate set of parameters.

For the parameter sets shown in Figures 3.2C-F, a reversible transition occurred in the simulation period. However, the system only remained in the inactive state momentarily, and so the system strongly favours the active state. Furthermore, several unstable intermediates, i.e. conformations with RMSD values over 2.0 nm, were observed in simulation D (λ_α = 0.8 and λ_β = 0.825) and F (λ_α = 0.825 and λ_β = 0.825). When λ_α and λ_β were both set to 0.85, the time spent in the two states became more balanced. Therefore, these are the optimal weights for studying WalK histidine kinase activation and inactivation.

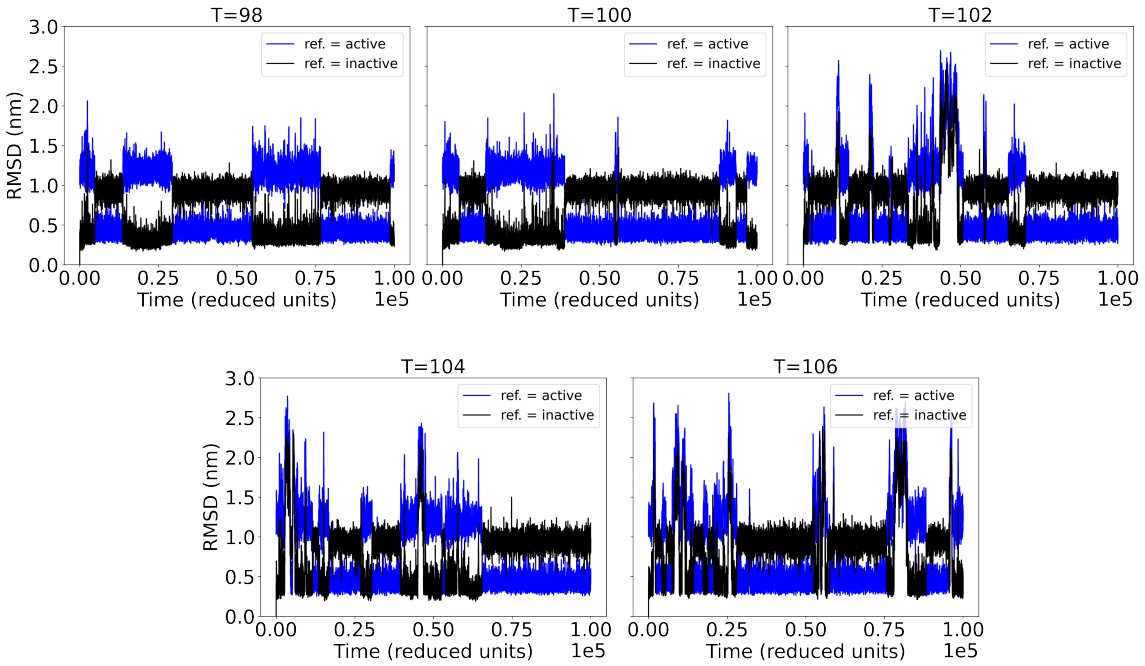


Figure 3.3.: Time evolution of the root mean square deviation (RMSD) along SBM simulations at five different temperatures. In these simulations, λ_α and λ_β are both equal to 0.85. RMSD was calculated with the inactive (black) and active (blue) α-carbons as the reference atoms.

Next, the optimal temperature for studying these transitions was determined. I performed constant temperature SBM simulations at GROMACS temperatures 98, 100, 102, 104 and 106 with the weighting parameters λ_α and λ_β set to 0.85. As shown in Figure 3.3, a preference for the active conformation arises at higher temperatures. Additionally, more unstable intermediates that are dissimilar to the end states are formed. At temperature 98,

the time spent in each state was roughly equal and the transitions occurred multiple times in the simulation period. Therefore, the ideal dual-basin SBM parameters for WalK HK are λ_α and λ_β equals 0.85, and 98 GROMACS temperature.

3.3.2. Conformational dynamics of the transitions

Figure 3.4 shows the changes in the radius of gyration over the simulation time period. It shows that the histidine kinase stabilizes at around 2.8-3.0 nm in the inactive conformation, and around 2.45-2.7 nm in the active conformation. This reflects the changes in the compactness of the protein. During HK activation, the catalytic ATP-binding domain is brought closer to the DHp domain resulting in the more compact, closed conformation. This is necessary to bring the ATP-lid closer to the phosphorylatable histidine in the DHp domain.

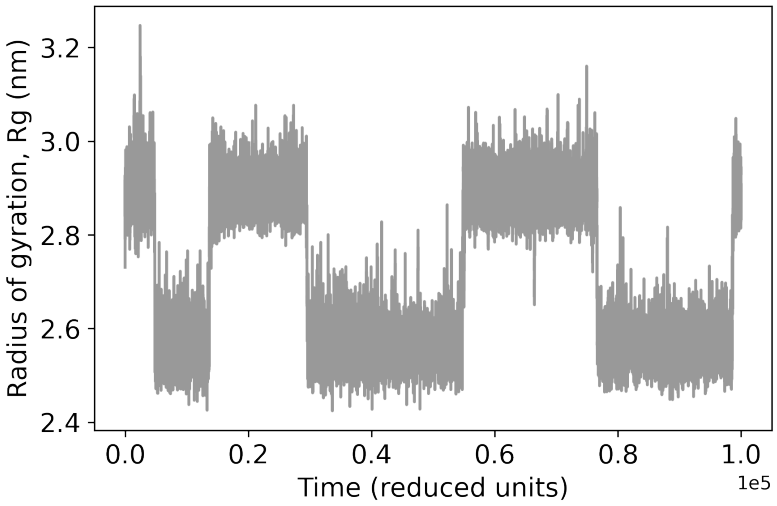


Figure 3.4.: Time evolution of the radius of gyration of histidine kinase. In these SBM simulations, λ_α and λ_β are both equal to 0.85.

Principal component analysis (PCA) is a technique that can be used to extract and analyze the collective motions from MD simulations. I used PCA to filtered the functionally relevant and often slow, collective motions from the fast, local fluctuations. A set of eigenvalues was obtained by the diagonalizing the covariance matrix of the α-carbon atomic fluctuations. As shown in Figure 3.5A, the proportion of variance decreases with the eigenvector indices. Here, we can see that the first eigenvector captures about 80 % of the total motion which indicates that this eigenvector defines the essential conformational subspace. The proportion of variance decays rapidly from eigenvector index 2 onward

which suggests that these represent rapid, localized fluctuations do not play a major role in the large-scale conformational changes in HK. The 2D PCA projection of the trajectory on the first two principal eigenvectors shows two distinct basins that display minimal overlap (see Fig. 3.5B). To visualize this conformational change, three different conformations were filtered from the PCA projection along the first eigenvector. The first structure is in the initial inactive conformation, the second is an intermediate, and the third structure is an active conformation of HK. As shown in Figure 3.5C, the slowest collective motion is the rotation of the catalytic ATP-binding domain with respect to the helical bundle of the DHp domain.

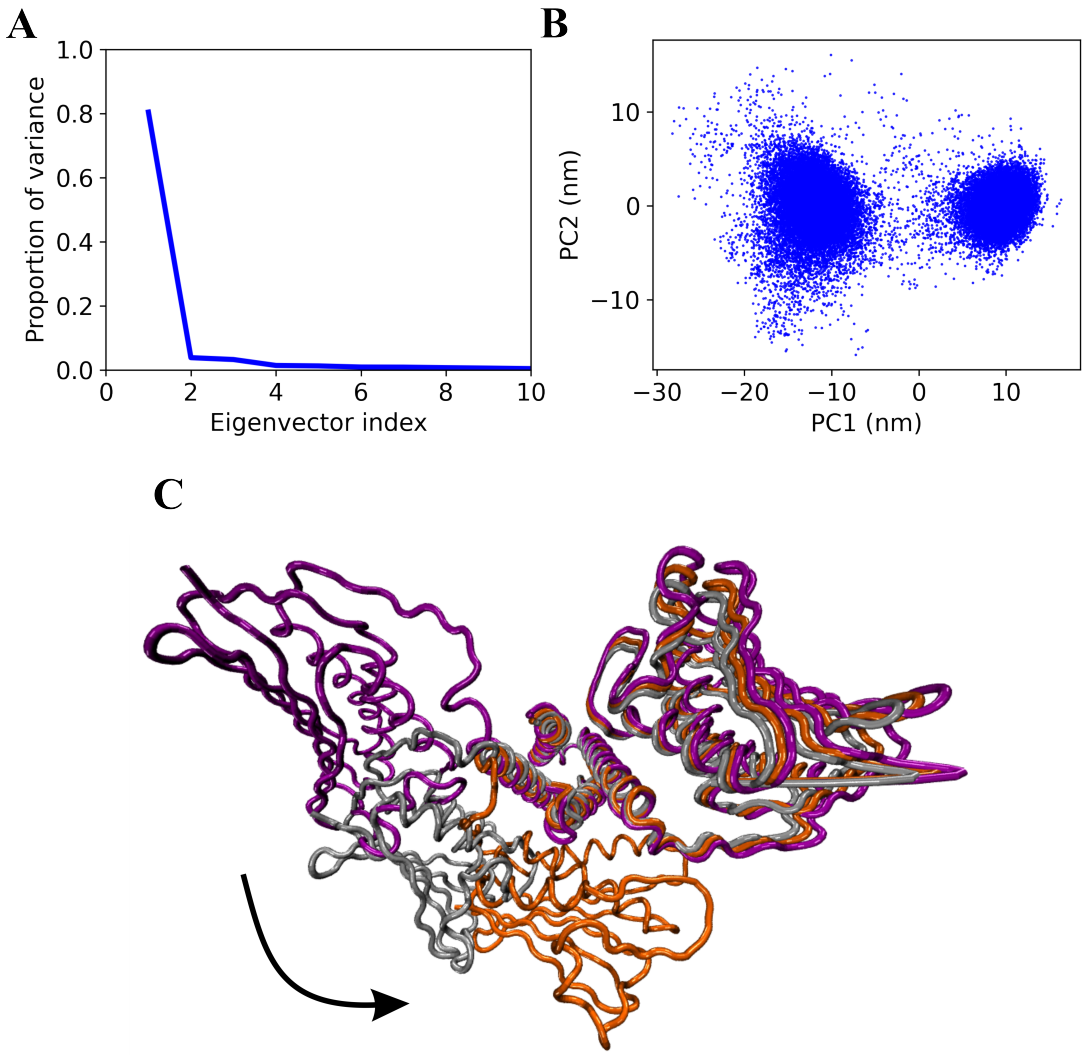


Figure 3.5.: Principal component analysis (PCA) of the dual-basin histidine kinase simulation. A) The proportion of variation of each eigenvector. B) Two-dimensional PCA projection along eigenvector 1 and eigenvector 2. C) Three representative structures extracted along PC1 are superimposed to illustrate the rotational motion of the catalytic-ATP binding domain.

Recently, Olivieri et al. [152] proposed two activation mechanisms for WalK histidine kinase. In the first mechanism, the CA domain detaches and rebinds to the DHp in the active conformation. In the second activation mechanism, CA-DHp contacts simultaneously form and break as the CA domain 'walks' along the DHp. Due to the lower energy barrier, the authors reported that the walking mechanism is the most probable pathway. The results from the PCA of the dual-basin SBM trajectory supports this hypothesis.

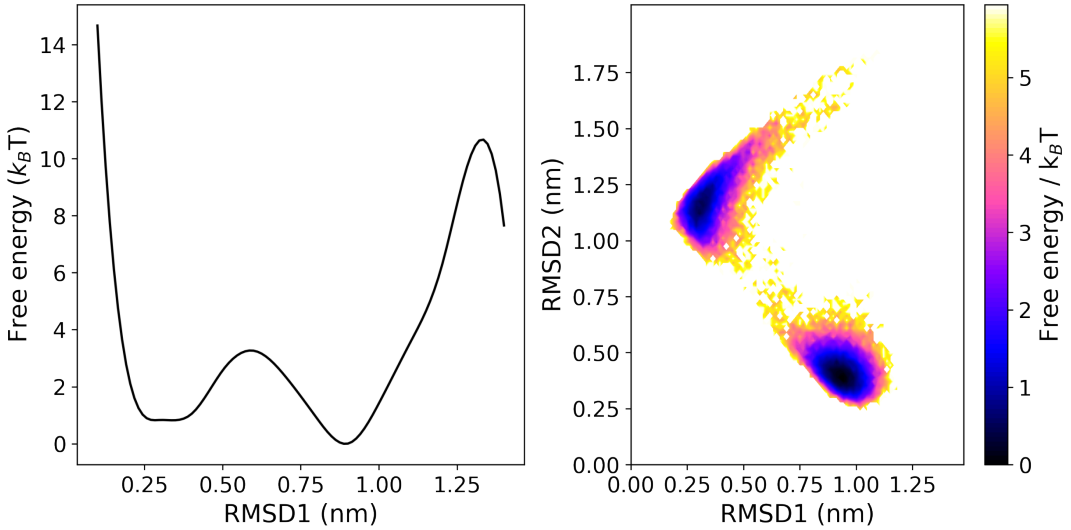


Figure 3.6.: Free energy profiles computed from the dual-basin SBM simulation of WalK histidine kinase (HK). **Left:** 1D free energy as a function of RMSD from the inactive conformation. **Right:** The 2D free energy profile as a function of RMSD1 and RMSD2, which are the α-carbon RMSDs from the inactive (PDB ID: 4U7N) and active (PDB ID: 4U7O) states, respectively [31].

The free energy profile was calculated from the 2D histogram of the RMSD values of the conformations along the dual-basin SBM trajectory (see Fig. 3.6). In the 2D free energy surface, there are two free energy minima, or basins, with the global minimum at the active state. This preference is perhaps more clearly illustrated by the 1D free energy profile, which was computed from the RMSD values with the inactive conformation as the reference structure. The energetic barrier for the forward transition is 2.4 k_BT, whereas the energetic barrier for the reverse transition is 3.3 k_BT. Therefore, the system favours the active state, albeit only by a 0.9 k_BT difference.

3.3.3. Identifying the Essential Contacts

The next objective was to identify which of the 176 native contact pairs are the most essential for HK activation. These contact pairs were divided based on the domains that

the residues are located in (see Table 3.1). The three subgroups are DHp-DHp, CA-CA and CA-DHp contacts. There are 79 interdomain contacts, 20 contact pairs within the DHp domain and 77 pairs within the CA domain. The weight λ_x for each subgroup x was varied whilst keeping the other subgroups at 0.85, which was previously determined as the optimum λ_β value. Figure 3.7 shows the 2D free energy landscapes for the simulations at the different parameters.

Modifications to the interdomain native contact pairs had the largest impact on whether the HK conformational transitions occur. When the contact strength λ_{CA-DHp} was increased to 1.0, HK activation occurred rapidly at the start of the simulation. The system stabilized in the active conformation, and thus the reverse transition did not occur. Decreasing the contact strength to 0.75 prevented the HK activation from occurring completely. In addition to the stable inactive state, high energy, unfavourable states were also seen in the trajectory. Similar observations were made when λ_{CA-DHp} was set to 0.8, although in this case, the system very briefly visits the active state. In contrast, decreasing λ_{CA} or λ_{DHp} below 0.85 did not have a significant impact on the probability of finding the system in either conformations (see Table 3.4). However, relative to the case where λ_β of all subgroup were 0.85, more high energy states were also visited. Increasing these parameters to 1.0 resulted in larger stationary probabilities at the active conformation basin of 0.78 and 0.81 for λ_{CA} and λ_{DHp}, respectively. The stationary probabilities of the metastable states were calculated using PCCA+ [124] from MSMs built from the SBM simulations. The stationary probabilities for different sets of weighting parameters are summarized in Table 3.4.

Table 3.4.: Summary from the PCCA+ for each set of SBM parameters.

#	λ_{CA-DHp}	λ_{CA}	λ_{DHp}	k-centers	Lag time (steps)	Stationary probabilities	
						Inactive	Active
1	0.85	0.85	0.85	100	10	0.46	0.54
*2	**1**	0.85	0.85	-	-	-	-
3	0.85	**1**	0.85	100	20	0.22	0.78
4	0.85	0.85	**1**	100	50	0.19	0.81
*5	**0.8**	0.85	0.85	-	-	-	-
6	0.85	**0.8**	0.85	100	50	0.49	0.51
7	0.85	0.85	**0.8**	100	20	0.47	0.53
*8	**0.75**	0.85	0.85	-	-	-	-
9	0.85	**0.75**	0.85	100	20	0.48	0.52
10	0.85	0.85	**0.75**	100	50	0.40	0.60

* Since there was only one metastable state, PCCA+ was not performed.

An advantage of using Markov State Models is that an MSM for the dynamics can be built from multiple trajectories. The stationary probabilities were therefore estimated from four repeat simulations of the same parameters to ensure sufficient sampling and confirm reproducibility of the results. The chosen lag time and hence the quality of the MSM, were further evaluated by the Chapman-Kolmogorov (CK) test (see Fig. A.1-A.3) [107]. For all cases, the probabilities predicted from MSMs have small deviations from the probability counts from the MD simulations. Therefore, the chosen lag time for each case are indeed appropriate.

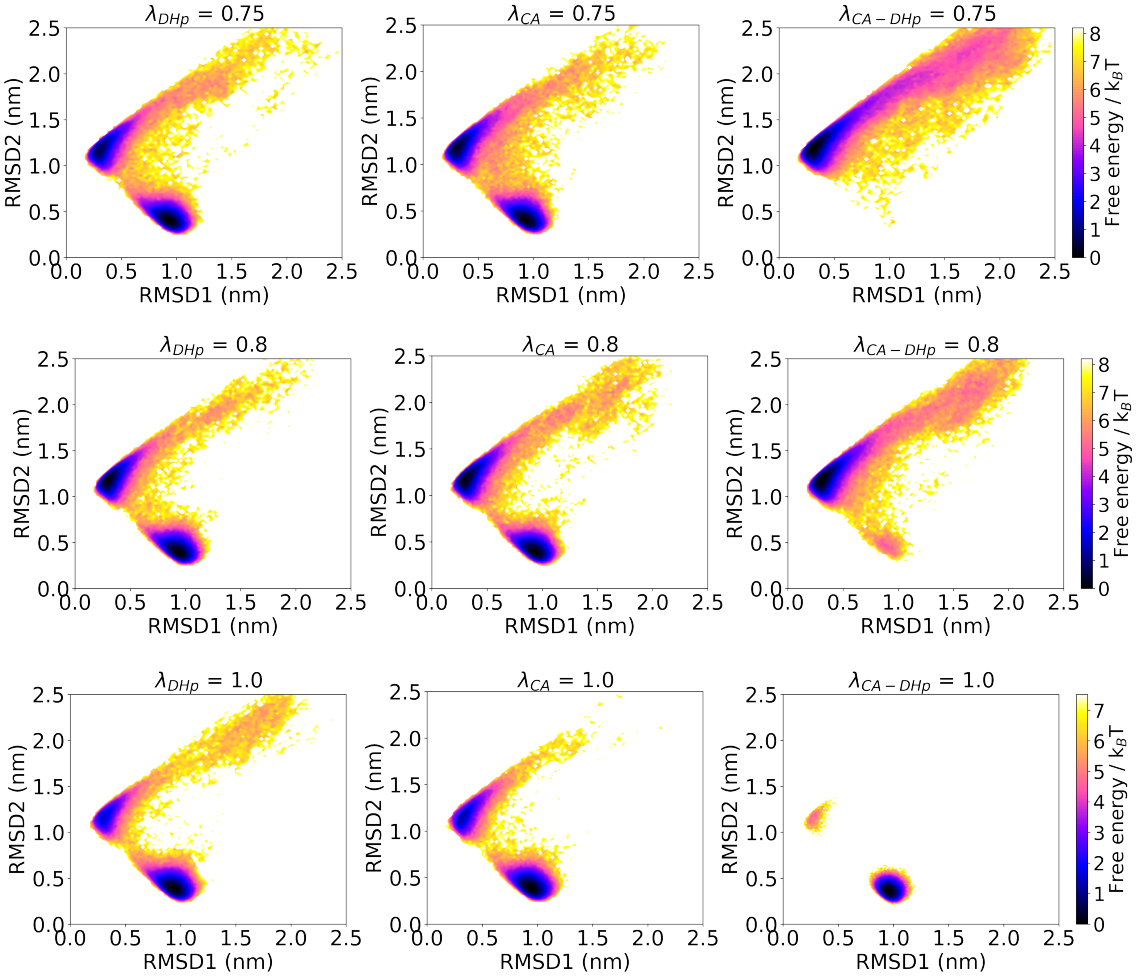


Figure 3.7.: Free energy landscapes of histidine kinase for different strengths of interdomain and intradomain contacts. One subgroup of contacts were modified at a time, while the other subgroups were given weights 0.85. The free energy profiles are plotted with respect to RMSD1 and RMSD2, which are the α-carbon root-mean-square deviations from reference structures inactive (PDB ID: 4U7N [31]) and active (PDB ID: 4U7O [31]), respectively.

3.3.4. Evaluation of Dual-basin SBM Parameters

In this chapter, I demonstrated that the large-scale reversible conformational transitions in HK can be simulated with a dual-basin SBM at a low computational cost. I have also shown that by altering certain contact energies, one can gain an intuition for the non-bonded interactions that drives the transition. An effective CG model should be able to reproduce experimental data or behavior observed in high resolution atomistic simulations. To assess the dual-basin structure-based potential determined in section 3.3.1, the free energy landscape was recalculate using the more detailed all-atom representation and AMBER99SB-ILDN force field [59]. Two conformations that are representative of the end states were extracted from the SBM trajectory. These two structures only contain the α-carbons of each amino acid, and so the corresponding all-atom models were reconstructed via a two step process. Firstly, the positions of the backbone atoms were predicted and the resultant backbone-only structure was energy minimized. In the final step, the side chain and hydrogen atoms were inserted.

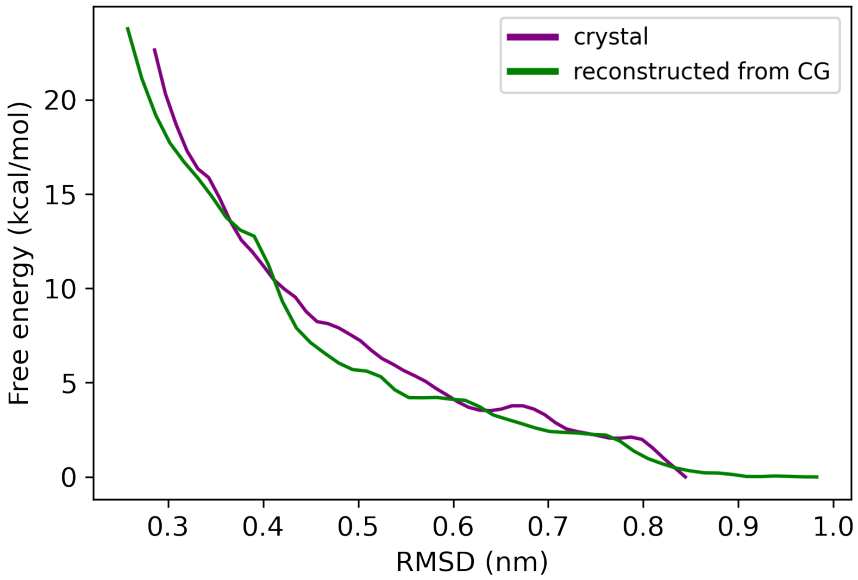


Figure 3.8.: Free energy profiles of WalK histidine kinase obtained using the weighted histogram analysis method for the crystal (purple) and reconstructed from coarse-grained (green) models. The active and inactive crystal structures were taken from PDB IDs 4U7O and 4U7N, respectively. The RMSDs were calculated with respect to the alpha carbons of both DHp protomers and the beta regions of the CA domain of the active protomer.

To obtain the free energy profile for conformational transitions between the inactive to active conformations, I used steered MD and umbrella sampling. Firstly, a time-dependent

harmonic bias drove the system from the remodelled inactive to active conformation. In this steered MD simulation, the harmonic bias was applied to the RMSD with respect to the remodelled active state. This results in a series of intermediate conformations along the activation pathway with decreasing RMSD values. A total of 32 conformations that are representative of different points along the collective variable (i.e. RMSD) were selected. Each of these conformations were used as the starting structures for the umbrella windows.

The free energy profile was obtained by combining 32 umbrella windows using WHAM [151]. The histograms from the 32 umbrella windows overlap which indicates that the phase space is properly sampled (see Fig. A.4b). To evaluate whether convergence was reached, the free energies were calculated using data from different time intervals (see Fig. A.4a). The first 10 ns were therefore discarded as equilibration phase, and the remaining 20 ns were used to construct the free energy profile. As shown in Figure 3.8, the transition pathway from the open to closed conformation of HK has a steep free energy increase with a large energetic barrier equivalent to ΔG. This free energy profile is almost in agreement with the crystal WalK HK simulations, which was computed using the same methodology. The values for ΔG obtained with the crystal and reconstructed from CG models were 23 ± 0.64 and 24 ± 0.53 kcal/mol respectively. It is worth noting, a limitation of using a single RMSD bias is the difficulty in distinguishing between conformations with large RMSD values. These are conformations that are significantly dissimilar from the reference structure. This could explain the small deviations in the intermediate states. For a more detailed analysis on the free energy landscape of WalK histidine kinase, see chapter 4.

3.4. Conclusion

Multiple basin force-fields have been previously developed to study the conformational transitions of kinases, namely adenylate [88–90] and Src kinase [93]. In this work, a dual-basin structure-based potential for studying histidine kinase conformational dynamics has been reported for the first time. This modified structure-based potential incorporates structural information of both the active and inactive conformations of HK, and thus enabling 'basin-hopping' to occur. The dual-basin potential was constructed by the addition of native contact energies that stabilise the active state into the single-basin potential of the inactive state. Weighting parameters, λ_α and λ_β, were introduced to modulate the

energetic contributions of the native contacts of the two states. Unlike in other micro-mixing implementations, the energetic contributions of the shared contact pairs were left unperturbed. It was found that by simply modifying λ_α and λ_β, we can induce the reversible transitions in HK. The optimal λ_α and λ_β values for obtaining an MD trajectory with approximately equal probability of the finding the system in either state was 0.85. The dual-basin SBM successfully reproduced the free energy change of activation predicted by the more computationally expensive enhanced sampling method with only small deviations. This confirms the robustness of the CG model.

In addition, it was determined that the interdomain contact pairs have a significantly greater influence than the intradomain contacts, in driving the forward transition. In future work, one could further investigate which interdomain contact pairs are the most essential. For example, one could explore the impact of the contacts around the ATP-binding pocket in the CA domain and around the phosphorylatable histidine of the DHp domain. One could also explore the possibility of integrating more complexity, for instance, electrostatic interactions into the existing dual-basin model. Recently, electrostatic interactions have been incorporated into alpha-carbon SBMs to investigating their effects in protein stability [153, 154], protein-DNA interactions [155], and protein folding [140, 141].

Overall, the dual-basin SBM provides a means for exploring the conformational pathways of histidine kinase at a remarkably low computational cost. Principal component analysis of the MD simulations with this CG model support the claim that the kinase core activates via a 'walking' mechanism. This forms the basis of the work in the following chapters.

4. Activation Pathways of WalK Histidine Kinase

This chapter is in part reproduced with permission from: F. Idiris, M. Kansari, H. Szurmant, T. Kubař and A. Schug. "Activation and Autophosphorylation of Histidine Kinases" (**submitted**).

4.1. Introduction

Histidine kinases (HK) and their role in the signal transduction cascade in bacteria has been widely studied in the recent years [31, 33, 156–160]. As HKs are large multi-domain transmembrane proteins, full-length characterization of the conformations poses a great challenge. In this work, I study the conformational dynamics of WalK histidine kinase. WalK HK is essential in the regulation of cell growth and division in low GC-content Gram-positive bacteria, including *Bacillus subtilis* [144] and *Staphylococcus aureus* [145]. The kinase core region of WalK, which is the site of an autophosphorylation reaction, has been characterized in the open inactive and closed active conformations (see Fig. 4.1A). This region is homodimeric and consists of the catalytic ATP-binding (CA) domains and the dimerization and histidine phosphotransfer (DHp) domains. While the structural details of the endpoints are known, we still have limited knowledge on the dynamics of the transition, the coupling between the domains, and the relationship between the conformation changes and the subsequent phosphoryl transfer reaction.

Figure 4.1B shows the superimposition of the two known crystal structures of WalK. We observe that two large-scale conformational changes likely occur at the kinase core during activation. The α-helices of the DHp bend, and the CA domain of one protomer rotates by 57° with respect to the DHp. The conformational dynamics from the inactive to active WalK HK transition has been analyzed in recent work by Olivieri et al [152]. The authors

argue that the activation pathway is likely via a "walking" mechanism in which the CA domain gradually rotates relative to the DHp domain through the simultaneous breaking and formation of inter-domain contacts. They also proposed an alternative "release and rebind" activation pathway which had a significantly higher energy barrier and is therefore unfavorable [152]. In chapter 3, MD simulations with the dual-basin SBM confirmed the walking mechanism. The coupling between the conformational changes of the individual domains however, have not been fully addressed and thus, is the focus of this work.

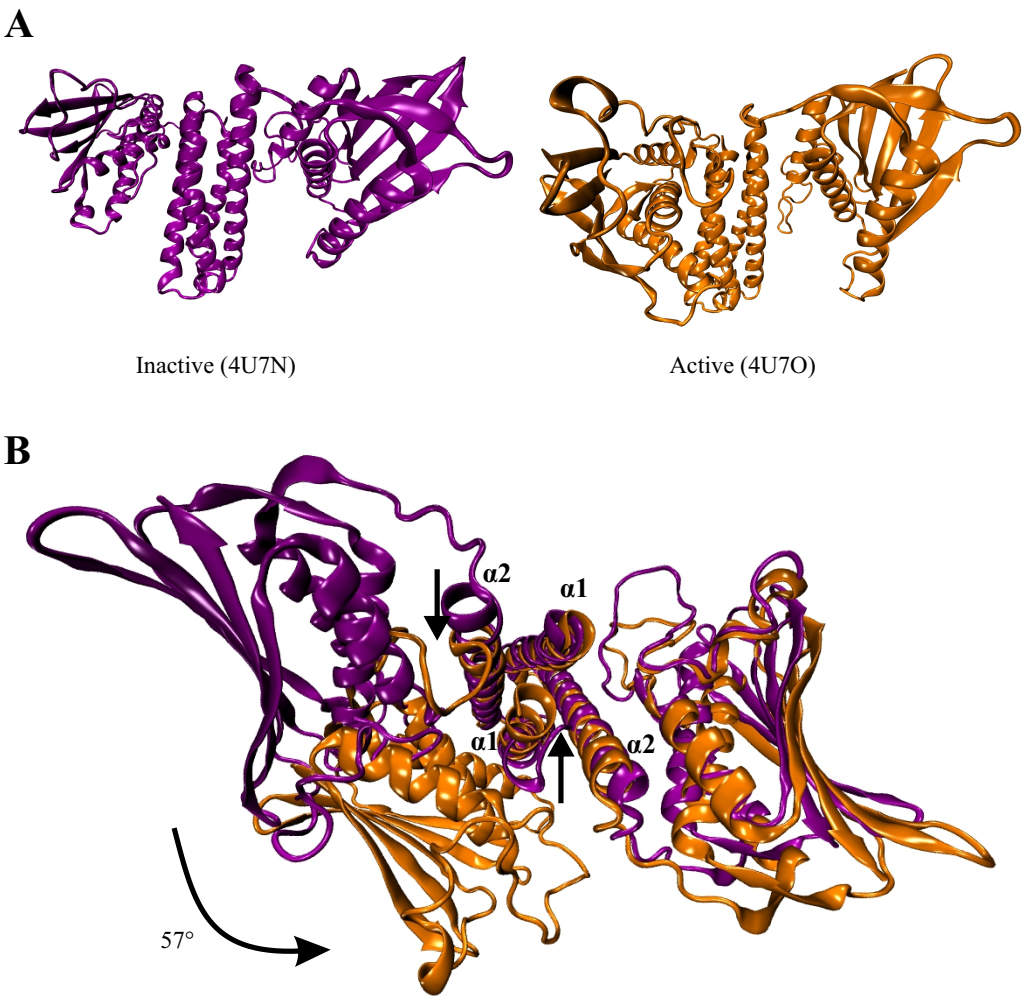


Figure 4.1.: The crystal structures of WalK histidine kinase. A) On the left is the inactive (PDB ID: 4U7N [31]) state and on the right is the active (PDB ID: 4U7O [31]) state of the kinase core. B) The two states are aligned to highlight the structural differences. The rotation of the CA domain is indicated with a curved arrow. The helical bending of the DHp domain at the α-1 and α-2 helices are each indicated with a straight arrow.

The aim of this work is to explore the relationship between the conformational changes of the DHp and CA domains. Here, I present two possible activation pathways of WalK HK using enhanced sampling molecular dynamics simulations. In the first activation

pathway, helical bending of the DHp domain and the rotation of the CA domain occur simultaneously. In the second activation pathway, helical bending of the DHp is induced before the rotation of the CA domain. The rationale behind this stems from the direction at which the signal travels. The signal first arrives at the kinase core from the N-terminus of the DHp domain, and then travels along the alpha-helix bundle before reaching the CA domain at the C-terminus.

4.2. Methods

4.2.1. Starting Materials

The inactive and active states of histidine kinase WalK, an essential HK from *Lactobacillus plantarum*, were obtained from the protein data bank (PDB ID: 4U7N and 4U7O, respectively) [31]. The non-terminal missing residues were modelled using Chimera's [146] interface to MODELLER version 10.0 [147].

All MD simulations of WalK HK were conducted at 300 K using GROMACS 2020.4 [84] patched with PLUMED v2.7 [149], with AMBER99SB-ILDN force-field [59], and with a time step of 2 fs. Each HK state was solvated with the TIP3P water model [150]. Electrostatic interactions were calculated using the particle mesh Ewald (PME) method [62], with a direct cutoff of 1.0 nm and a grid spacing of 0.16 nm. A cutoff of 1.0 nm was used for the van der Waals interactions. All bonds involving hydrogen were constrained by the LINCS algorithm [49]. Each HK-water system was first minimised using the steepest descent algorithm for 1000 steps. This was followed by NVT equilibration with the velocity-rescaling thermostat [54] for 1 ns and then NPT equilibration with the Parrinello-Rahman barostat [55] for 1 ns.

4.2.2. Concerted Activation Mechanism

4.2.2.1. Steered Molecular Dynamics

To obtain the starting structures for the individual umbrella windows, I used steered MD. The following time-dependent harmonic bias potential was applied to the difference in root-mean-square deviation (RMSD) between the two reference states:

$$\Delta RMSD = RMSD_{inactive} - RMSD_{active} \tag{4.1}$$

$$V_{SMD} = \frac{k}{2} \left(\Delta RMSD(t) - \Delta RMSD^*\right)^2 . \tag{4.2}$$

Here, k is the spring constant, $\Delta RMSD(t)$ is the instantaneous $\Delta RMSD$ between the current coordinates and the reference structures. $\Delta RMSD^*$ is a value that evolves linearly from the initial $\Delta RMSD$ at the first steered MD step to the final target $\Delta RMSD$. The alpha carbons of the DHp domain of both protomers and the beta regions of the active protomer's CA domain were considered for the RMSD calculations. To ensure that the system was fully equilibrated, 1 ns of MD without steering forces was performed. Then, histidine kinase activation was guided by a spring constant of 5000 $kJmol^{-1}nm^{-2}$ for 3 ns.

4.2.2.2. Umbrella Sampling

A total of 37 intermediate conformations along the collective variable $\Delta RMSD$ were extracted from the steered MD trajectory. The values of $\Delta RMSD$ for each of these structures ranged between -0.66 to 0.50 nm. Each of these conformations were used as the starting structure of an independent umbrella sampling simulation known as an umbrella window. The following harmonic bias was applied to each umbrella window i:

$$V_i = \frac{k}{2} \left(\Delta RMSD - \Delta RMSD_i^{ref}\right)^2 . \tag{4.3}$$

Here, $\Delta RMSD_i^{ref}$ is the reference $\Delta RMSD$ for the umbrella window i. Each window was simulated for 30 ns with a spring constant of 10000 $kJmol^{-1}nm^{-2}$. The free energy profile was obtained by combining the umbrella windows using the weighted histogram analysis method [151]. The first 10 ns was discarded as equilibration phase. The statistical errors were computed by Monte Carlo bootstrap error analysis (50 bootstrap samples) implemented in the WHAM code [151].

4.2.3. Step-wise Activation Pathway

Using the equilibrated inactive WalK histidine kinase in explicit water system as the initial structure, a time-dependent harmonic bias was applied to the alpha carbon atoms of the DHp domain of both protomers.

$$V_{DHp} = \frac{k}{2} \left(RMSD(t) - RMSD^{*} \right)^2 . \tag{4.4}$$

Here, k is the spring constant, $RMSD(t)$ is the instantaneous RMSD between the current coordinates and the reference structure. $RMSD^{*}$ is a value that evolves linearly from the initial RMSD at the first steered MD step to the final target RMSD. At each time step, the structure was first aligned with the alpha carbons of the DHp domain and the beta regions of the active CA domain. Then, the RMSD was calculated for only the alpha carbons of the DHp domain. This steered MD simulation was run for 2 ns with a spring constant of 5000 $\mathrm{kJmol^{-1}nm^{-2}}$.

4.3. Results and Discussions

4.3.1. Concerted Activation Pathway of WalK Histidine Kinase

I investigated two possible transition pathways for WalK histidine kinase activation at the kinase core. The first activation mechanism follows the assumption that the conformational changes of the individual subdomains of HK occur simultaneously. Activation of the kinase core results in helical bending of the DHp domain, and one of the CA domains rotate. The conformational transitions were driven by steered MD with a time-dependent harmonic bias along the collective variable, $\Delta RMSD$. The atoms used in the RMSD calculations were situated in both the CA and DHp domains which ensured that both domain motions occurred concurrently. As shown in Figure 4.2, the $\Delta RMSD$ value plateaus after around 2 ns at about 0.5 nm. This corresponds to an RMSD value with respect to the active conformation of around 0.38 nm.

During this conformational transition, the catalytic ATP-binding domain gradually rotates by 57° with respect to the DHp helical bundle. As shown in Figure 4.3A, the center-of-mass (COM) distance between the DHp and CA domains decreased from approximately 2.55 to 1.92 nm by the end of the simulation. This brings the phosphorylatable histidine (His391) in the DHp domain closer to the ATP-lid (Leu545-Leu577) in the CA domain to form the asymmetrical active state. As shown in Figure 4.3B, the backbone radius of gyration decreased from about 2.9 to 2.65 nm during the transition. This indicates that the overall compactness of the protein has increased and further confirms the formation of the closed, active conformation.

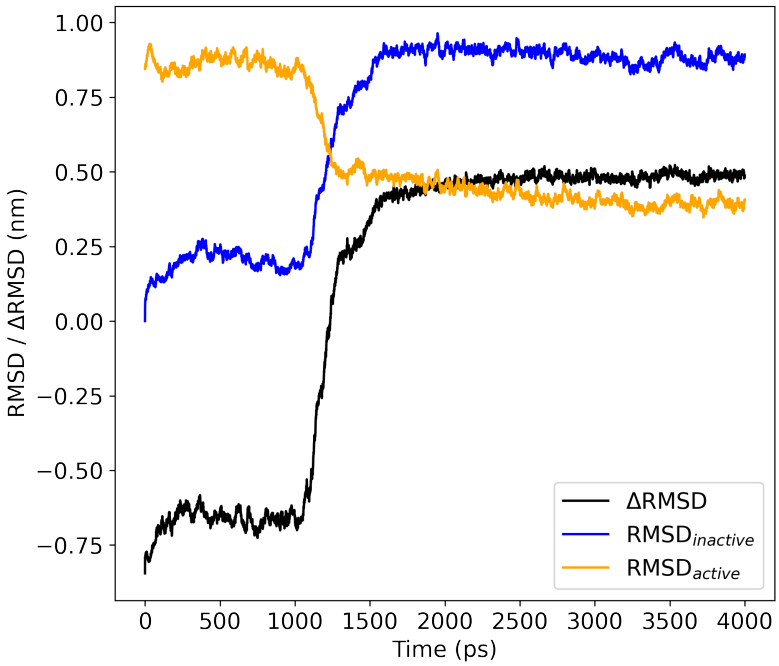


Figure 4.2.: Time evolution of RMSD and $\Delta RMSD$ during the steered molecular dynamics simulation of WalK histidine kinase. The RMSDs were calculated from the α-carbons of the DHp domains and beta regions of the CA domain of the inactive and active conformations of the HK.

I selected a total of 37 conformations from the steered MD trajectory that were sampled along the transition pathway. Each of these conformations were used as the initial structures for the independent US simulations. Although it is possible to estimate a free energy profile solely using steered MD, this would necessitate performing the simulation several times with the same parameters. Especially in the case of large complex systems such as proteins, where convergence is not easily reached, this potentially has a large computation cost and statistical error. Therefore, it is often more efficient to use steered MD as a preliminary step before umbrella sampling. The histograms from the 37 umbrella windows overlap which indicates that the phase space is properly sampled (see Fig. B.1b). To ensure that the simulations were properly converged, I calculated the free energy using different subsets of data along the simulations (see Fig. B.1a). The free energy converged from 10 ns onward, and so the simulation data from the first 10 ns were discarded from the free energy calculations with WHAM.

In the free energy profile, there is initially a small increase followed by a steep rise in the free energy from around $\Delta RMSD = 0$ nm onward (see Fig. 4.4). To rationalize this observation, I estimated the formation and breaking of contacts in these two stages. Contacts are defined as non-bonded interactions formed between residues within a cutoff

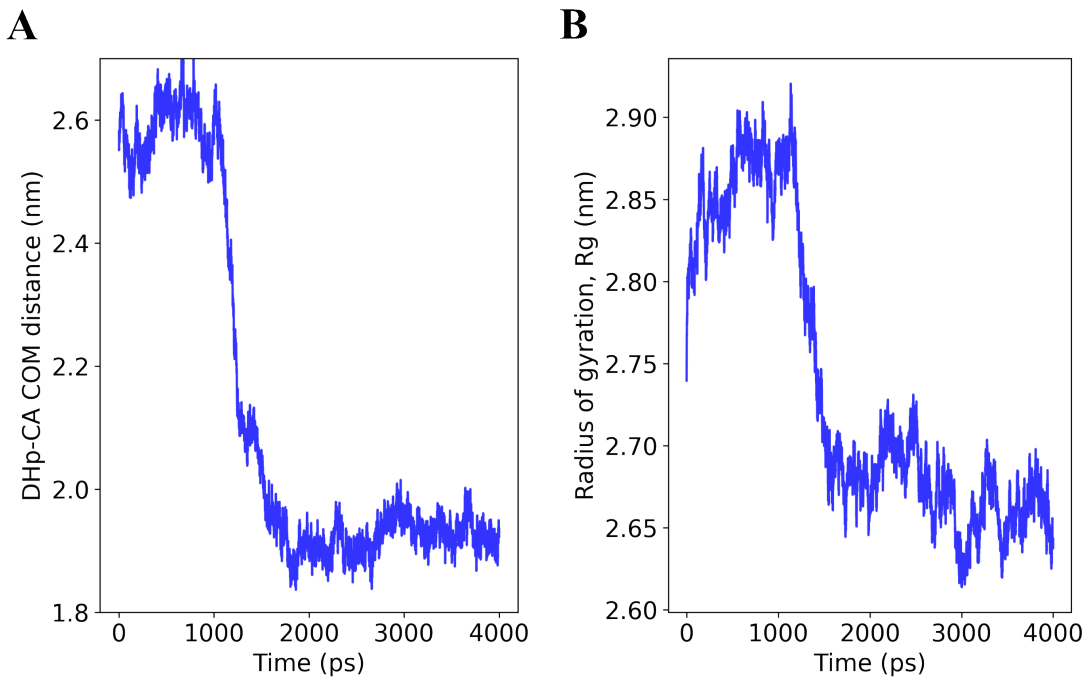


Figure 4.3.: A) Time evolution of the backbone radius of gyration and B) the changes in the center-of-mass (COM) distance between the DHp and CA domains during the steered molecular dynamics simulations of histidine kinase.

distance of 0.45 nm. Contact frequencies are the fraction of frames during which the contact was present. I first calculated the contact frequencies from the US trajectories of umbrella windows $\Delta RMSD$ = -0.66, 0 and 0.5 nm. The umbrella window centered at $\Delta RMSD$ = -0.66 nm is in the inactive conformation, $\Delta RMSD$ = 0 nm is an intermediate with equal structural similarity to both end states, and $\Delta RMSD$ = 0.5 nm is the active state (see Fig. 4.5A). I then calculated the difference in contact frequencies to obtain the contact maps shown in Figure 4.5B.

The free energy increases by about 6 kcal/mol in the transition from $\Delta RMSD = -0.66$ to 0 nm. At this stage, new contacts between the DHp and CA domain including Arg434–Asp502, Gln570–Thr395 and Gly392–Thr573 are formed as the CA domain rotates by 30°. In the second stage from $\Delta RMSD$ = 0 to 0.5 nm, there is a much larger increase in free energy of about 20 kcal/mol. Relative to the initial phase, more CA–DHp contacts are formed. Three regions in the difference contact frequency maps where contact formation is favorable, have been identified and highlighted. The number of CA–DHp contacts between residues Thr395–Gly410 of the DHp and residues Lys475–Pro480 in the CA alpha helix (see Fig. 4.5B colored in red) increased significantly. In this step, the CA domain rotates by 27°, bringing these two regions together in an ideal position for non-bonded interactions to form. Contacts between DHp residues Leu422–Asp430 and CA residues

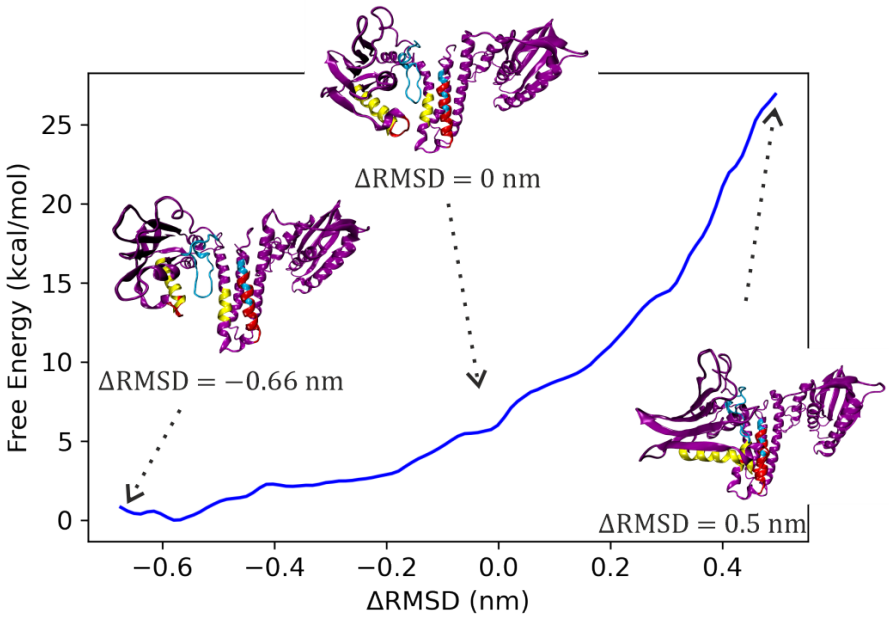


Figure 4.4.: The free energy profile with respect to the difference in RMSDs ($\Delta RMSD$) to the inactive and active reference structures. The initial structures for 3 of the windows are shown. Regions colored in red, yellow and blue are the areas where CA-DHp contact pair interactions were favored in the active state

Asn460–Lys476 formed in step 1 with some residue interactions slightly shifted in the formation of the active state in step 2. For example, the interaction between Val424-Lys476 broke in order for Val424–Arg469, Val424–Met472 and Ser402–Lys476 interactions to form. The final region of interest is between DHp residues Glu392–Ser402 and CA loop region residues Phe559–Ile579. Contact pair formation in this region became increasingly more favorable as the two domain were brought closer together.

Overall, the energetic barrier for WalK HK activation was 26.9 ± 0.96 kcal/mol, which is in reasonable agreement with the previous study [152]. In spite of the increase in interdomain contact pairs, the final conformation had a larger free energy than the inactive state. Many of the amino acids in the contact pairs are hydrophillic due to the presence of a polar or charged side chain. These interactions are weaker than hydrophobic interactions due to their tendency to also interact with the water molecules surrounding the protein. This could possibly explain the instability of the final active conformation.

Enhanced sampling techniques provide a means for exploring large conformational changes that would otherwise have been infeasible in a single classical MD simulation. I chose umbrella sampling as the endpoint crystal structures are available in the protein data bank, and we already had an intuition of the transition pathway (see chapter 3). While enhanced sampling techniques are great tools for predicting the FEP of conformational

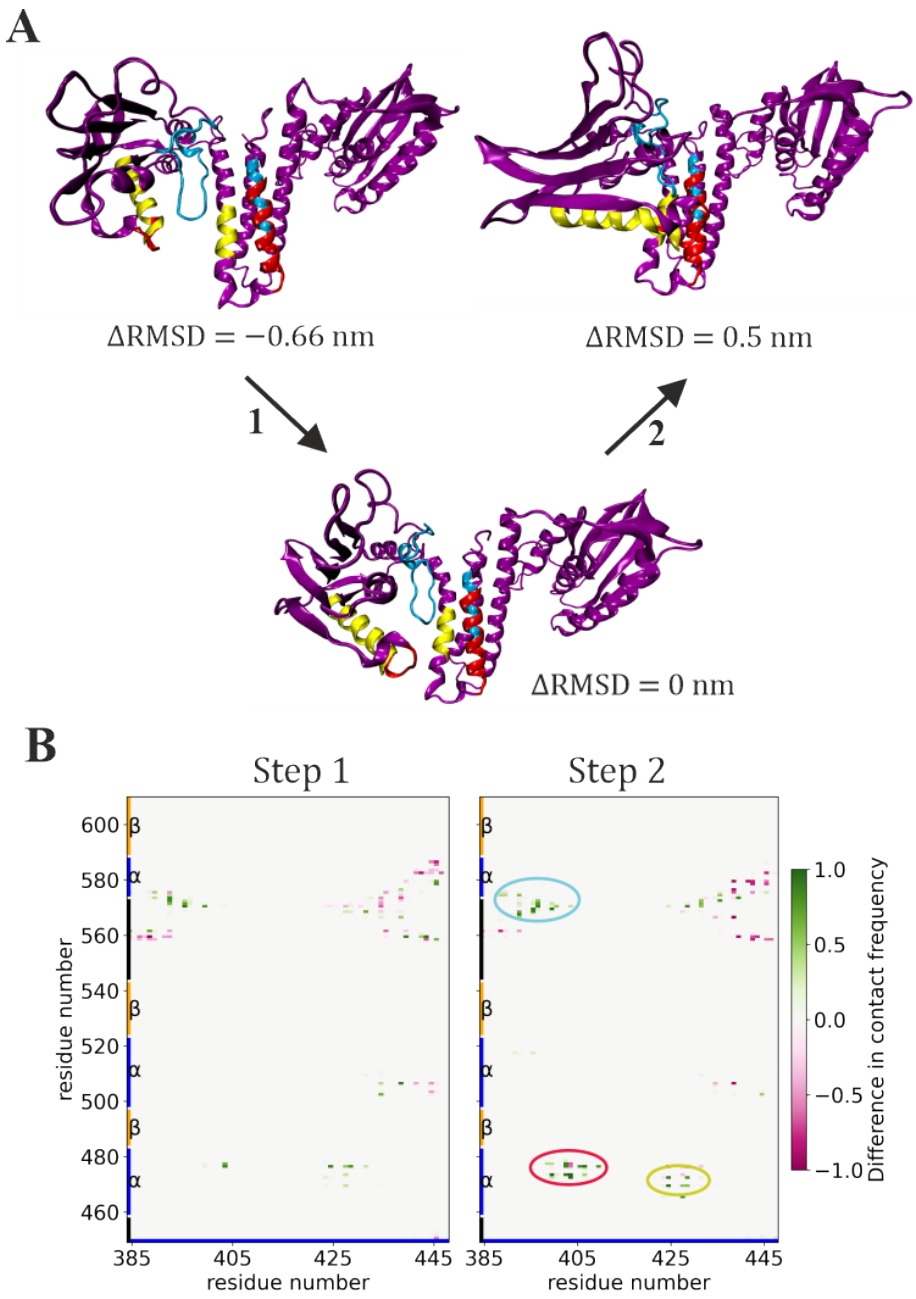


Figure 4.5.: Histidine kinase activation pathway split into two steps. A) Three structures from the transition pathway. Step 1 is the transition from the inactive state ($\Delta RMSD$ = -0.66 nm) to an intermediate ($\Delta RMSD$ = 0 nm) and step 2 is from this intermediate to the active state ($\Delta RMSD$ = 0.5 nm). Regions colored in red, yellow and blue are the areas where interdomain contact pair interactions were favored in the active state. B) Difference in contact frequency maps for steps 1 and 2. Three regions where contact pair formation was highly favorable are circled in red, yellow and blue as seen in A.

transitions in proteins, it is important to note that the intrinsic free energy values are possibly slightly overestimated by these methods [161]. To minimize this error, I ensured that the conformation space was sufficiently sampled as indicated by the overlap in the

histograms from the 37 umbrella windows (see Fig. B.1b). I also ensured that the simulations had reached convergence by comparing the free energy calculations from different subsets of the simulation data (see Fig. B.1a). What we can indeed interpret from the estimated free energy profile is that the active conformation of WalK is highly unstable relative to the inactive conformation.

4.3.2. Step-wise Activation Pathway

In section 4.3.1, I presented and investigated a transition pathway in which WalK HK activation occurs via a concerted mechanism. As the biasing $\Delta RMSD$ potential was applied to the α-carbon atoms situated in both domains, the helical bending of the DHp domain and the rotation of the CA domain necessary for activation occurred simultaneously. Here, I present an alternative activation pathway to further explore the relationship between the conformational changes of the individual domains. In this transition mechanism, the domains of the kinase core undergo their respective conformational changes independently in a step-wise manner. When an external stressor is detected at the sensor domain of HK, the signal first arrives at the N-terminus of the DHp domain and then travels along to the CA domain. I performed a steered MD simulation with an RMSD harmonic biasing potential applied solely to the α-carbons of the DHp domain. This means that the CA domain was left unperturbed and the structural transition was only induced on the DHp domains by the local perturbation. As shown in Figure 4.6, the RMSD quickly converges at around 0.28 nm.

Surprisingly, the steered MD trajectory illustrates that the DHp helical bending lead to the rearrangement of the protein's 3D structure, such that the CA domain rotates concurrently. The CA domain rotates by 50.1° which is just 7° less than what had occurred when the RMSD bias was applied onto both domains. Additionally, the center of mass distance between the domains decreased from 2.55 to 1.95 nm, as shown in Figure 4.7A. As a result, the radius of gyration of the protein's backbone decreased from approximately 2.8 nm in the initial open conformation to 2.65 nm in the final conformation (see Fig. 4.7B). This indicates the formation of a more compact HK conformation. Both the DHp-CA COM distance and R_g values of the end state is in agreement with the values obtained in the concerted mechanism. This means that the final conformations from both pathways have roughly the same degree of compactness. This also suggests that the phosphorylatable histidine in the DHp domain is in an ideal position from the ATP lid in the CA domain for the phosphorylation reaction to occur. Furthermore, the final conformation had a

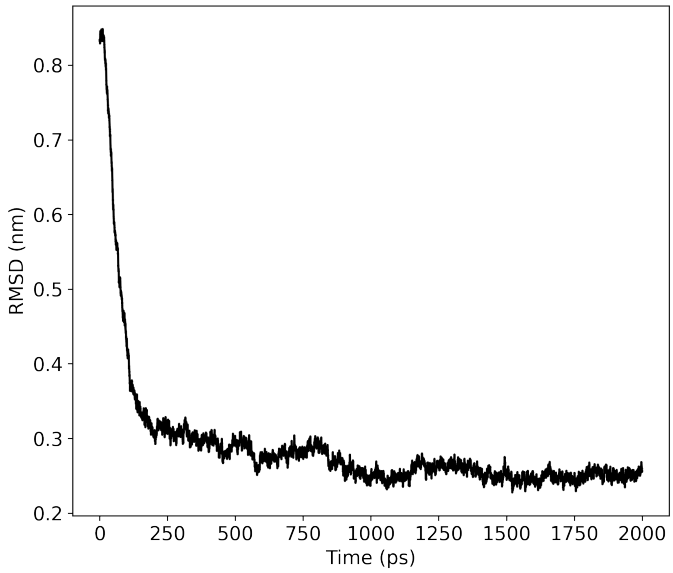


Figure 4.6.: Time evolution of RMSD during the steered molecular dynamics simulation of histidine kinase. The RMSDs were calculated from the α-carbons of the DHp domains of the active conformations of HK.

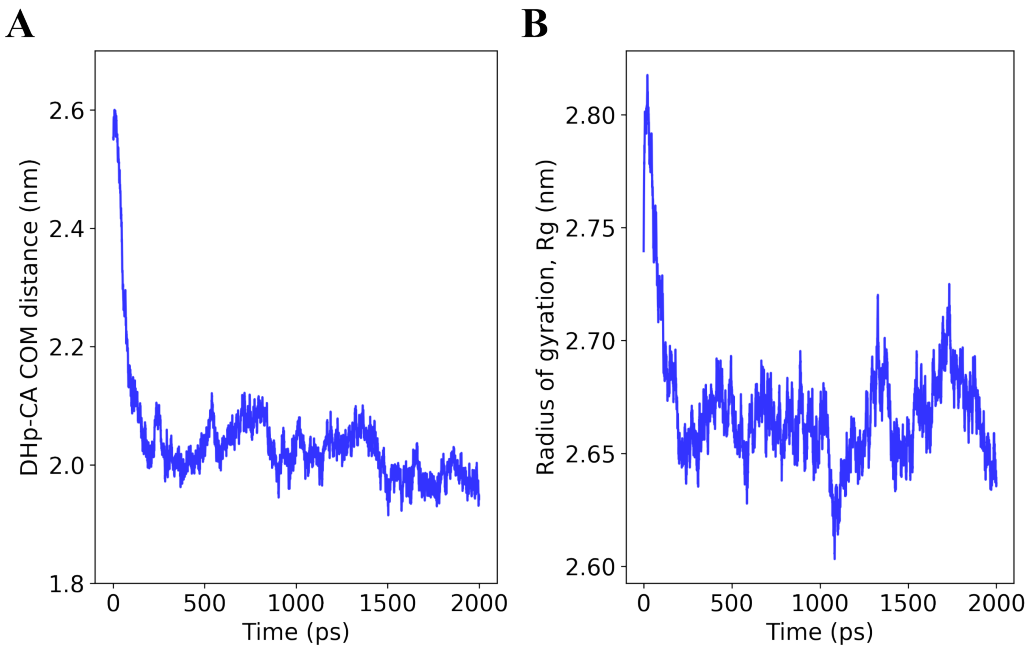


Figure 4.7.: A) Time evolution of the backbone radius of gyration and B) the changes in the center-of-mass (COM) distance between the DHp and CA domains during the steered molecular dynamics simulations of histidine kinase.

backbone RMSD of 0.45 nm with respect to the crystal structure of activated WalK (PDB ID 4U7O [31]), indicating a large degree of structural similarity.

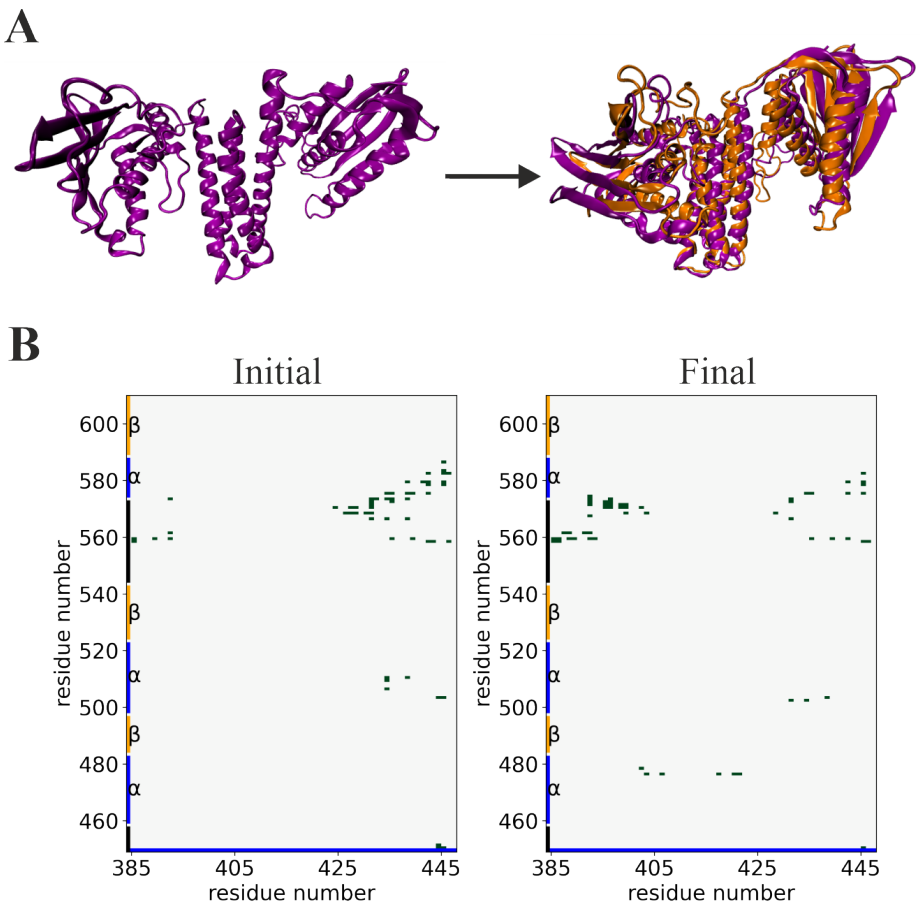


Figure 4.8.: A) The initial and final conformations from steered MD. The final state (right; purple) is superimposed with the crystal active WalK histidine kinase (PDB ID 4U7O; orange). B) Contact maps of CA-DHp interactions of the initial (left) and final (right) states. Each point on the contact map represents the presence (green square) or absence (no marking) of contacts between residues of the CA and DHp domains of chain B. Contacts are defined as non-bonded interactions between heavy atoms within a cut-off distance of 0.45 nm.

In addition, important CA–DHp contacts including Arg568–Glu428, Arg434–Asp502 and Phe559–Asn388 were formed. Overall, there is a larger number of CA-DHp non-bonded interactions in the final state relative to the starting structure (see Fig. 4.8B). For these reasons we can infer that the final structure is in the appropriate conformation for phosphorylation reaction in WalK HK to occur. This suggests that the large-scale conformational transitions of the DHp and CA domains involved in HK activation are tightly coupled and hence essentially a single step process.

In contrast, Marsico et al [162] demonstrated that while HAMP-DHp activation in CpxA histidine kinase induces CA domain reorientation, only approximately one third of the full transition is achieved. A key difference between CpxA and WalK histidine kinase lies in the directionality of the autophosphorylation reaction. Specifically, WalK HK undergoes cis-phosphorylation whereas CpxA HK undergoes trans-phosphorylation. For histidine

kinases that phosphorylate in cis, each monomer phosphorylates itself. Conversely, trans-phosphorylation occurs when each monomer in the dimer phosphorylates the other subunit. It is believed that whether phosphorylation occurs in cis or trans is dependent on the handedness of the hairpin loop that connects the individual DHp helices [35]. Moreover, the WalK HK model did not include HAMP as the crystal structure was not readily available. Therefore, the implications of the presence of a linker domain such as HAMP can not be determined.

4.4. Conclusions

Histidine kinases undergo a series of transient conformational changes when a signal is detected at the sensor domain. A specific focus of this work was the kinase core, which is the region at which a phosphorylation reaction occurs. I examined the relationship between the conformational changes that occur at each of the subdomains of the kinase core using enhanced sampling techniques. I presented and analyzed two possible activation pathways of a *cis*-phosphorylating histidine kinase known as WalK.

In the first activation mechanism, helical bending of the DHp domains and the rotation of one of the CA domains occurs simultaneously. The domains are brought closer together as the CA domain rotates gradually by 57° with respect to the DHp helical bundle. Contact frequency maps were calculated for umbrella windows of three different points along the collective variable, $\Delta RMSD$. The difference in contact frequencies has shown that weakly held interdomain non-bonded interactions were simultaneously broken and formed as the system activates. The free energy profile of the conformational changes revealed that the active state of WalK is highly unstable relative to the inactive state.

An alternative step-wise activation pathway was also investigated. In this transition pathway, local perturbations were placed only on the atoms of the DHp domain. Interestingly, bending of the helical bundle of the DHp domain had induced the rotation of the CA domain by 50.1°. While the rotational angle was somewhat smaller than as seen in the concerted MD simulation, it is possible that an active conformation has been obtained. The changes in the backbone radius of gyration were comparable to the concerted mechanism and both reflect the formation of a closed compact conformation. Additionally, the center-of-mass distance between the DHp and the CA domains by the end of the steered MD simulations of both pathways was approximately 1.9 nm. This suggests that the phosphorylatable histidine is in close enough proximity to the ATP for the phosphorylation reaction to occur.

All in all, the large-scale conformational changes of the subdomains at the kinase core during activation are tightly coupled and essentially a single-step process.

5. Activation Pathway of CpxA Histidine Kinase

5.1. Introduction

CpxA histidine kinase is an envelope stress senor protein found in *Escherichia coli* [163]. The two-component system for regulating the response to protein misfolding in the periplasm is comprised of CpxA HK and a response regulator protein, CpxR. The homodimeric kinase CpxA consists of a periplasmic sensor domain, a transmembrane domain, a HAMP domain, a dimerization and histidine phosphotransfer (DHp) and a catalytic ATP-binding (CA) domain (see Fig. 5.1A) [156].

In chapters 3 and 4, another type of histidine kinase known as WalK was examined. Although WalK and CpxA possess structural similarities, they differ in the directionality of the phosphorylation reaction that occurs at the kinase core. Both HKs are homodimeric however, CpxA undergoes trans-phosphorylation whereas WalK undergoes cis-phosphorylation. This is due to the handedness of the hairpin loop that connects the individual DHp helices [35]. I used the crystallographic structures of the inactive and active conformations to gain insight into the signalling mechanism of CpxA (see Fig. 5.1B). The inactive homodimer possesses a symmetrical structure with the ATP positioned far away from the phosphorylatable histidine (His248). Two crystal forms (trigonal and hexagonal space groups) of the asymmetric active homodimer are known. These two active conformations differ solely in the direction of the DHp-HAMP helical bending.

In this chapter, I explored the large-scale conformational changes involved in driving the system from the inactive state to the hexagonal hemiphosphorylated active states of CpxA (see Fig. 5.1B). In a previous study, Marsico et al. [162] simulated possible transition pathways for CpxA activation using steered MD with a coarse-grained model. The authors also reported the free energy profile for the autophosphorylation reaction, however,

they did not address the thermodynamic properties associated with the conformational changes [162]. To fill this gap in our knowledge, I report the free energy profile of the conformational transitions which I obtained using umbrella sampling. Additionally, I identified the functionally relevant collective motions using principal component analysis, and analyzed the formation and breaking of non-bonded interactions. I then drew a comparison between the conformational dynamics of the kinase core of CpxA and WalK.

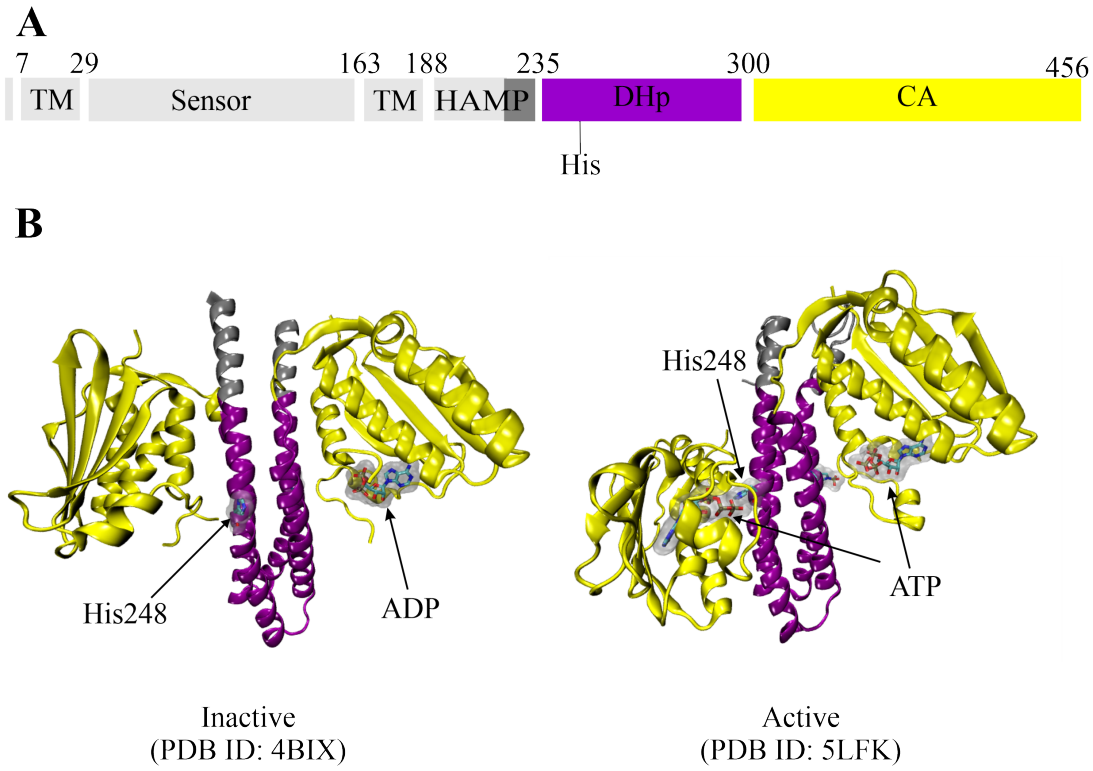


Figure 5.1.: A) Schematic of the full length structure of CpxA histidine kinase. The domains and the amino acid residue numbers are displayed. B) Crystallographic structures of two conformations of CpxA hisitidine kinase used in the molecular dynamics simulations. The kinase core region consisting of the catalytic ATP-binding domains (CA) and the dimerization and histidine phosphotransfer domain (DHp) are colored in yellow and purple, respectively. The HAMP domain (in grey) is also partially present in these crystal structures. In the inactive state (PDB ID: 4BIX [156]), HAMP residues Ala220-Arg234 in chain A and residues Ala223-Arg234 in chain B are present. In the active state (PDB ID: 5LFK [36]), HAMP residues Ala213-Arg234 in chain A and residues Gln216-Arg234 in chain B are present. The phosphorylatable histidine (His248) and ATP/ADP are displayed and highlighted in light grey.

5.2. Methods

5.2.1. Starting Materials and System Preparation

The crystal structures of the inactive and active states of CpxA histidine kinase from *Escherichia coli* were obtained from the protein data bank (PDB ID: 4BIX and 5LFK respectively) [36, 156]. The non-terminal missing residues were modelled using Chimera's [146] interface to MODELLER version 10.0 [147]. As shown in Figure 5.1B, the CA (residues Asn301-Arg456) and DHp (residues Met235-Lys300) domains are almost fully modelled, while only a few residues of the HAMP domain are present. In the inactive crystal structure, HAMP residues Ala220-Arg234 in chain A and residues Ala223-Arg234 in chain B are present. In the active conformation, HAMP residues Ala213-Arg234 in chain A and residues Gln216-Arg234 in chain B are present.

All MD simulations of CpxA histidine kinase were conducted at 300 K using GROMACS 2020.4 [84] patched with PLUMED v2.7 [149], with AMBER99SB-ILDN force-field [59], and with a time step of 2 fs. Each CpxA HK conformation was first solvated with the TIP3P water model [150]. The electrostatic interactions were calculated using the particle mesh Ewald (PME) method [62], with a direct cutoff of 1.0 nm and a grid spacing of 0.16 nm. A cutoff of 1.0 nm was used for the van der Waals interactions. All bonds involving hydrogen were constrained by the LINCS algorithm [49]. Each HK-water system had undergone energy minimisation using the steepest descent algorithm for 1000 steps. Energy minimisation was followed by NVT equilibration with the velocity-rescaling thermostat [54] for 1 ns and then NPT equilibration with the Parrinello-Rahman barostat [55] for 1 ns.

5.2.2. Steered Molecular Dynamics Simulations of CpxA Histidine Kinase

To study the transition between the two conformational states of CpxA steered MD was performed using two RMSDs simultaneously as the collective variables. The following time-dependent harmonic bias potentials were applied to the system:

$$V_{RMSD1} = \frac{k}{2}\left(RMSD(t) - RMSD^{*}_{inactive}\right)^2, \tag{5.1}$$

$$V_{RMSD2} = \frac{k}{2}\left(RMSD(t) - RMSD^{*}_{active}\right)^2. \tag{5.2}$$

Here, k is the spring constant, $RMSD(t)$ is the instantaneous RMSDs between the current coordinates and the reference structure. $RMSD^{*}_{inactive}$ and $RMSD^{*}_{active}$ are values that evolve linearly from the initial RMSD at the first steered MD step to the final target RMSD. The reference structures for equations 5.1 and 5.2 were the inactive and active states, respectively. The atoms accounted for in the RMSD calculations were the alpha carbons of the DHp domain of both protomers and the beta regions of chain B's CA domain. Firstly, 1 ns of equilibration MD was performed, then CpxA activation was guided by the means of the double RMSD harmonic bias with a spring constant of 5000 kJmol^{-1}nm^{-2} for an additional 3 ns. After the simulation, a new variable, $\Delta RMSD$, was defined and assigned to each of the conformations along the steered MD trajectory. $\Delta RMSD$ was defined as:

$$\Delta RMSD = RMSD_{inactive} - RMSD_{active}. \tag{5.3}$$

5.2.3. Umbrella Sampling of CpxA Histidine Kinase

A total of 42 intermediate conformations along the transition pathway in the steered MD trajectory were chosen and extracted. Each conformation had a $\Delta RMSD$ value within the range of -0.85 to 0.65 nm. Each of these conformations were used as the initial structure of an umbrella window. The harmonic bias potential defined in Chapter 4 equation 4.3 was used. Each window was restrained using the $\Delta RMSD$ collective variable for 20 ns with a spring constant k of 10000 kJmol^{-1}nm^{-2}. The free energy profile was obtained using Weighted Histogram Analysis Method and the statistical errors were computed by Monte Carlo bootstrap error analysis (50 bootstrap samples) which is implemented in the WHAM code [151].

5.3. Results and Discussions

5.3.1. Steered Molecular Dynamics Analysis

Firstly, I sampled the conformational space along the transition pathway using steered MD. Starting from the inactive state, I introduced a double RMSD harmonic bias after the first 1 ns (i.e. the equilibration phase). As shown in Fig. 5.2, the RMSD with respect to the inactive gradually increased while the RMSD with respect to the active state decreased. In other words, the system was gradually driven towards the active conformation and away from the inactive conformation. The RMSD_{active} value plateaus after around 3 ns at around 0.4 nm. I defined a new CV, $\Delta RMSD$, which I used for the umbrella sampling simulations.

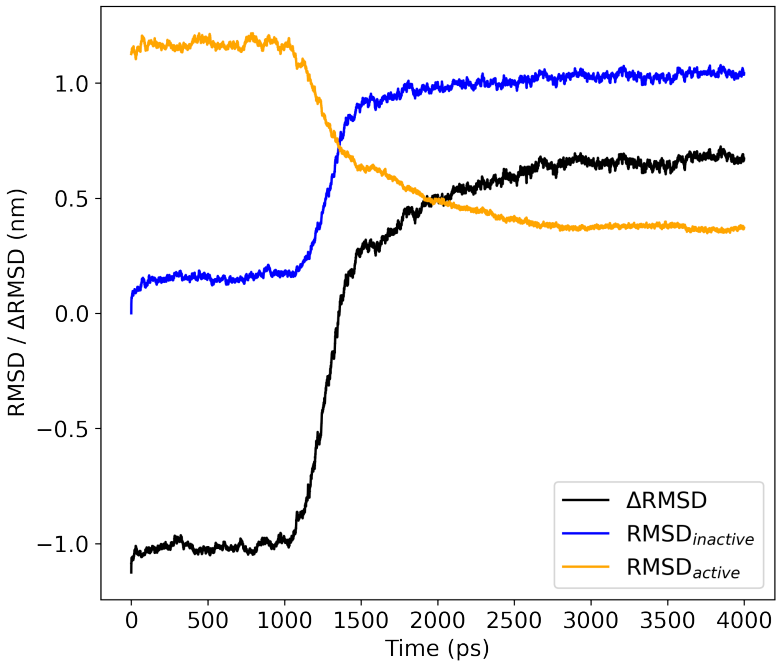


Figure 5.2.: Time evolution of RMSD and $\Delta RMSD$ during the steered molecular dynamics simulation of CpxA histidine kinase. The RMSDs were calculated from the α-carbons of the DHp domains and beta regions of the CA domain of the inactive and active conformations of HK.

As with other kinases, CpxA switches between an open and closed conformation following activation at the kinase core. A phosphorylatable histidine (His248) is contained in the DHp α-helix of chain A, and the ATP-binding pocket is situated in the CA domain of chain B. To assess whether an active conformation with an appropriate DHp-CA orientation and compactness was obtained, I monitored the DHp-CA center-of-mass distance over

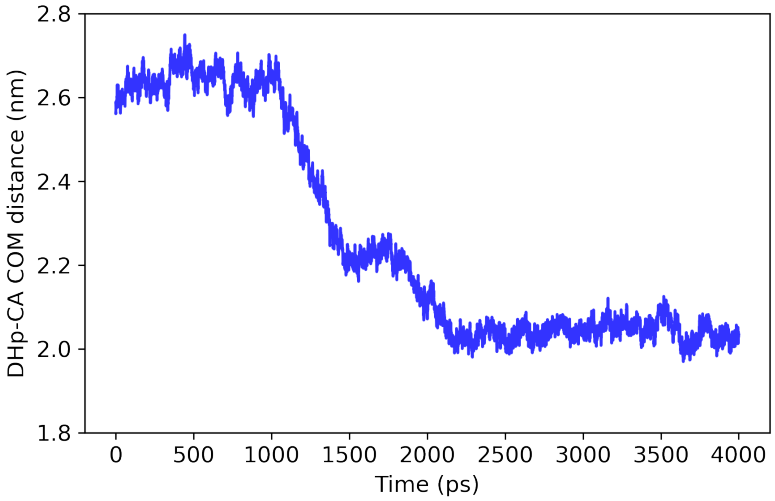


Figure 5.3.: The changes in the center-of-mass (COM) distance between the DHp and CA domains during the steered molecular dynamics simulation of CpxA histidine kinase. The COM of the DHp was calculated from the atoms of the DHp strands in close proximity to the CA domain (i.e. chain A residues Met235-Glu270 and chain B residues Glu270-Lys300).

the simulation period (see Fig. 5.3). The COM distance began at approximately 2.6 nm, progressively decreased and stabilized at 1.9-2.1 nm. This is comparable to the DHp-CA COM distance in the active closed crystal structure (PDB ID: 5LFK) which is 1.89 nm.

5.3.1.1. Principal component analysis

I used principal component analysis to identify the functionally relevant collective motions of the transition pathway from the steered MD trajectory. By diagonalizing the covariance matrix of the backbone atomic fluctuations, a set of eigenvalues was obtained. As shown in Figure 5.4A, the first eigenvector captures about 90% of the total motion and the proportion of variance rapidly decays with increasing eigenvector index. Therefore, the first eigenvector defines the essential conformational subspace and the other eigenvectors represent rapid, localized fluctuations. Projection along the first eigenvector shows the rotational motion of the CA domain relative to the DHp domain by 68° (see Fig. 5.4C). The twisting motion of the DHp helical bundle facilitated this rearrangement in the CA domain with respective to the DHp. Simultaneously, alpha helix segments Ala220-Ala223 (chain A) and Ala223-Ala231 (chain B) of the HAMP at the termini bent away from the rotating CA domain.

The second eigenvector represents another local helical bending motion along the dimeric HAMP-DHp. Another kink forms at Asn226 at chain A, causing the Ala220-Asn226

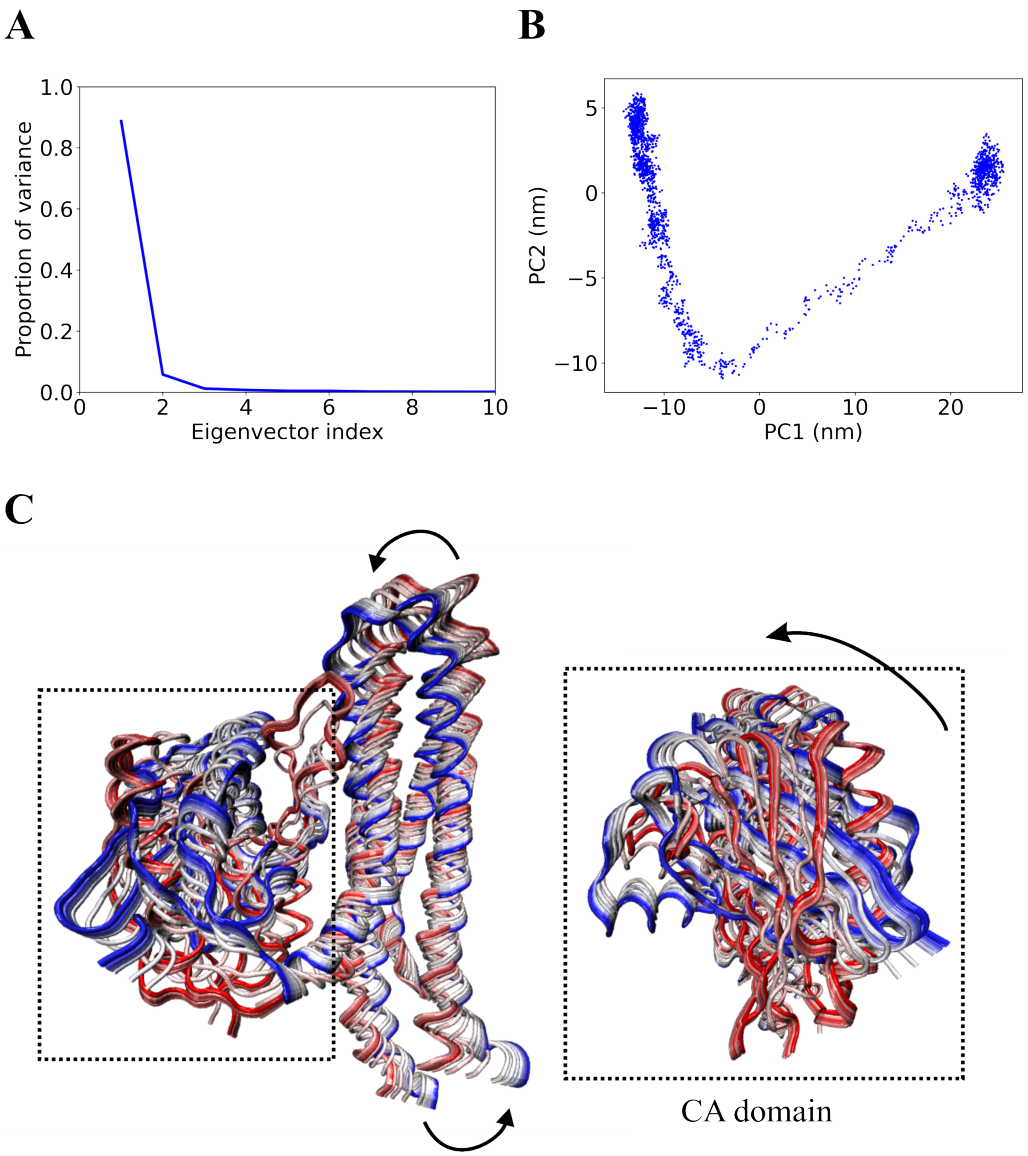


Figure 5.4.: Principal component analysis (PCA) of the steered molecular dynamics simulation of CpxA histidine kinase. A) The proportion of variation of each eigenvector. B) Two-dimensional PCA projection along eigenvector 1 and eigenvector 2. C) The superimposition of conformers along principal component 1 (PC1) depicting the global motion. The CA domain of chain B and the DHp domains of both protomers are shown on the left and a close-up of the CA domain is shown on the right. A red-white-blue color gradient is used to show the progression of the simulation, with red being the initial conformation and blue being the final conformation.

segment at the terminus to bend away from the activating CA domain. The 2D projection along the trajectory with respect to the first two eigenvectors, PC1 and PC2, is shown in Figure 5.4B. The initial values for PC1 and PC2 were 22.3 and 1.4 nm, respectively. Firstly, the collective motions associated with PC1 occurs as indicated by the decrease to ~-9 nm. In parallel, the helical bending at Asn226 occurs as shown by the concurrent decrease in PC2 to ~-10 nm. At this stage, the CA domain is now at an ideal position

for the autophosphorylation reaction occur. Interestingly, after activation, the numerical value of PC2 increases to approximately 5 nm. This is due to the Ala220-Asn226 alpha helix segment reverting back into the unbent state.

5.3.2. Free Energy Profile of CpxA Hisitidine Kinase Activation

The free energy profile of the conformational changes associated with CpxA histidine kinase activation was predicted using umbrella sampling (see Fig. 5.5). I selected a total of 42 conformations from the steered MD trajectory based on their $\Delta RMSD$ values. These are representative structures along the activation pathway of CpxA, and each served as the starting structure for an umbrella window. I calculated the unbiased free energies by combining the umbrella windows using WHAM. The histograms from the umbrella windows overlap indicating that the phase space between the reference structures has been sufficiently sampled (see Fig. C.1b). Additionally, to check that the simulations were fully converged, I calculated the free energies at various time intervals along the umbrella trajectories. As seen in Figure C.1a, the free energies converged after 10 ns, and so the simulation data from the first 10 ns were discarded and deemed the equilibration phase. The free energy profile with respect to $\Delta RMSD$ in Figure 5.5 shows a steep increase in free energy as the system transitions from the inactive to active states. Overall, CpxA HK activation has an energetic barrier of 31.1 ± 1.09 kcal/mol. Marsico et. al [162] reported a free energy barrier of 16 kcal/mol for the autophosphorylation reaction that occurs in the activated CpxA. This suggests that conformational changes represent the rate-limiting step of the signalling mechanism at the kinase core.

Principal component analysis of the steered MD trajectory has showed that the rotation motion of the CA domain relative to the DHp-HAMP helical bundle is the dominant motion in the overall activation pathway (see Fig. 5.4). Therefore, we can rationalize the changes in free energy in relation to the changes in the interactions that form between the CA and DHp domains.

The notable instability of the final active conformation could possibly be attributed to the net loss in CA-DHp contacts. CA-DHp contacts were defined as nonbonded interactions formed between heavy atoms in the activating CA domain (residues Asn301-Lys455 in chain B) and atoms of the DHp strands in close proximity to the CA domain (residues Met235-Glu270 in chain A and residues Glu270-Lys300 in chain B). A cut-off distance of 0.45 nm was set. As the CA domain rotates, CA-DHp nonbonded interactions are

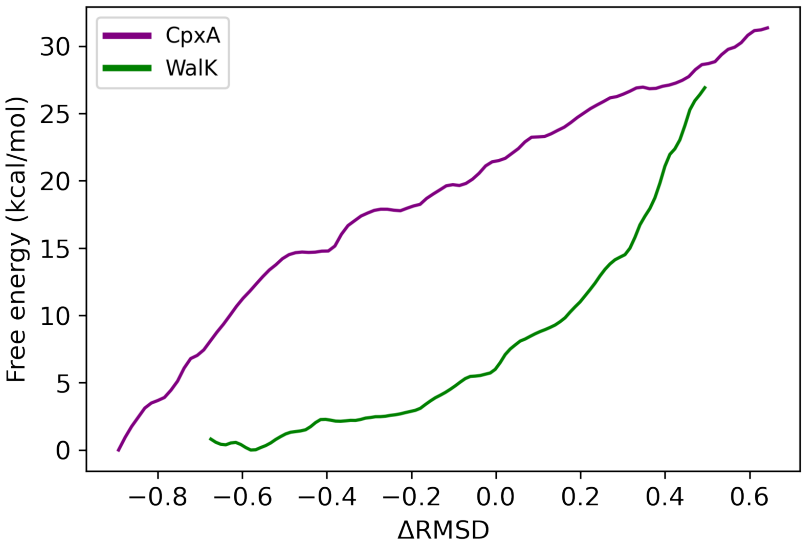


Figure 5.5.: Free energy profiles of CpxA (purple) and WalK (green) histidine kinase with respect to the difference in RMSDs ($\Delta RMSD$) to the inactive and active reference structures. See Chapter 4 for more detail on the free energy profile of WalK.

simultaneously forming and breaking. The CA-DHp contact frequencies from umbrella windows $\Delta RMSD$ -0.85 nm (inactive) and $\Delta RMSD$ 0.65 nm (active) were compared (see Fig. 5.6). There were 12 contact pairs with a difference in contact frequency greater than 0.75 as summarized in Table 5.1. These can be interpreted as new contacts that have formed. Conversely, there were 38 contact pairs with a difference in contact frequency of less than -0.75. These contact pairs could be interpreted as nonbonded interactions that have broken. This overall loss in nonbonded interactions results in a highly unstable product with a larger Gibbs free energy.

In comparison to WalK, CpxA's difference in free energy was 4 kcal/mol larger (see Chapter 4 and Fig. 5.5). Both free energies were predicted using umbrella sampling with a $\Delta RMSD$ harmonic bias potential and under the same conditions. The distance between the γ-phosphate of ATP in the CA domain and the phosphorylatable histidine in the DHp domain is larger in CpxA. From the crystal structure of the hemiphosphorylated CpxA (PDB ID: 5LFK), this distance decreases by 1.73 nm. Similarly in activated WalK (PDB ID: 4U7O), the distance between β-phosphorous and His391 decreases by just 1.28 nm. Note that the distance was calculated from the β-phosphorous of AMP phosphoramidate as a crystallographic structure with ATP bound is not currently available for WalK. Moreover, we observe that the CA domain of CpxA rotates by 68° whereas in WalK, the CA domain only rotates by 56.5° to achieve the ideal conformation for the phosphorylation reaction to

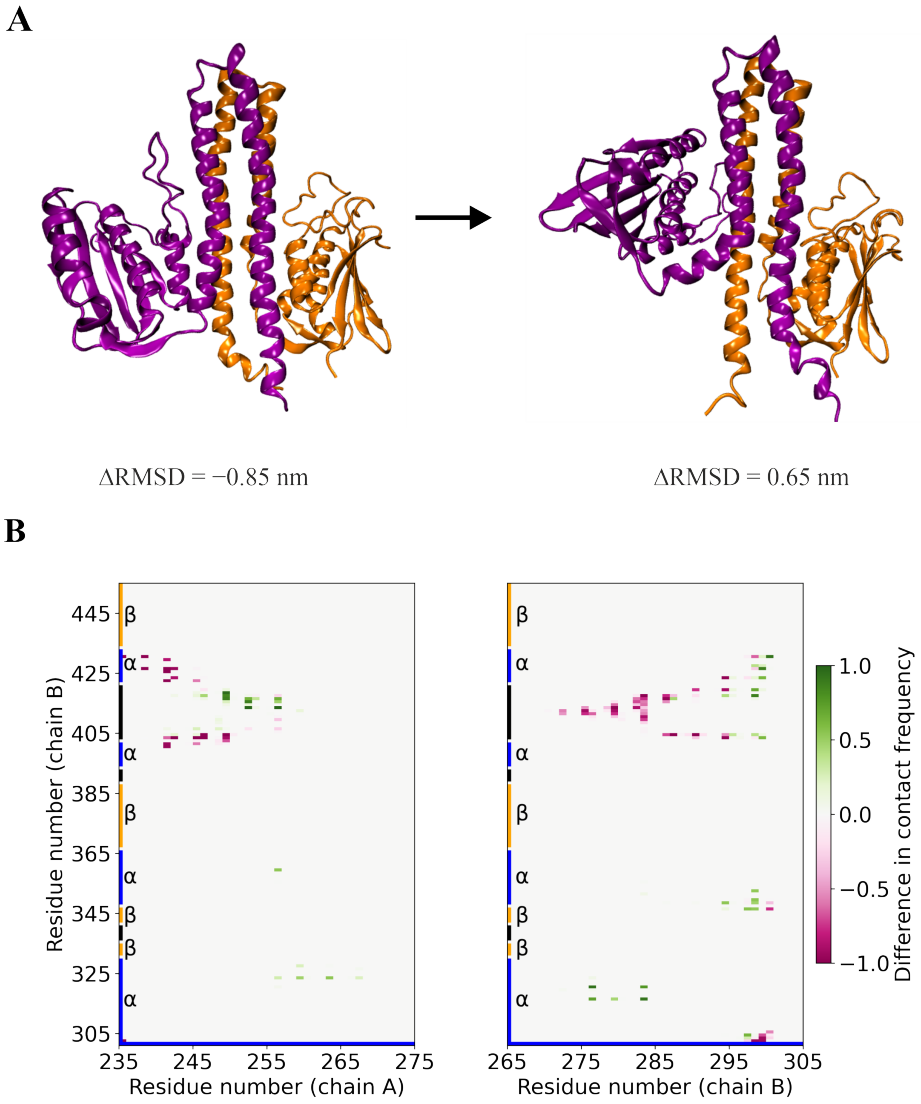


Figure 5.6.: (A) The cartoon representations of the initial structures from umbrella windows, Δ*RMSD* -0.85 nm (inactive) and Δ*RMSD* 0.65 nm (active). Chains A and B are labelled in orange and purple, respectively. (B) Difference in contact frequency maps of CpxA histidine kinase. The contact frequencies were taken from umbrella windows Δ*RMSD* -0.85 nm (inactive) and Δ*RMSD* 0.65 nm (active) with cut-off distance 0.45 nm between DHp (chain A residues Met235-Glu270 and chain B residues Glu270-Lys300) and CA (chain B residues Asn301-Lys455) heavy atoms.

occur. Furthermore, the active closed conformation of WalK is partially stabilized by the net gain of weakly held nonbonded interdomain interactions. CpxA, on the other hand, experiences a net loss in CA-DHp contact pairs.

However, caution should be taken when making a direct comparison between the free energy profiles. In the PCA projection of the SMD trajectory along the first eigenvector, one could also observe the local helical bending at the HAMP termini. In the crystal

Table 5.1.: Important CA-DHp Contact Pairs in CpxA Histidine Kinase's Active State.

Contact pair (CA-DHp)	Difference in contact frequency (active - inactive)
Glu413–Thr252	1.000
Glu413–Arg256	1.000
Gly418–Glu249	0.965
Asn320–Arg276	0.945
Glu316–Arg283	0.929
Gln430–Lys300	0.926
Thr417–Gln298	0.882
Thr426–Gln299	0.878
Gly416–Glu249	0.877
Gly415–Thr252	0.876
Thr417–Glu249	0.844
Asn320–Arg283	0.776

structure of CpxA only residues Ala220-Arg234 in chain A and residues Ala223-Arg234 in chain B of the HAMP domain were present. Although the HAMP domain was only partially present, it is possible that the formation of kinks at Ala223 in chain A and Ala231 in chain B could have contributed to the free energy calculations. The crystal structure of WalK only contained the DHp and CA domains, and thus the conformational changes associated with these domains contributed to the free energies.

5.4. Conclusions

I applied enhanced sampling techniques to examine the conformational dynamics of CpxA, an envelope stress sensor histidine kinase [163]. The conformational transition between two known states of the kinase core was simulated. One of which is in the symmetrical inactive conformation, and the other is in the hemiphosphorylated asymmetrical active conformation. Principal component analysis of the steered MD trajectory has enabled us to determine and visualize the relative importance of the different motions. The projection along the PC1, i.e., the slowest collective motion, showed that the counter-clockwise rotation of the CA domain and simultaneous helical bending in DHp-HAMP termini accounts for about 90% of the total motion. This highlights the tight coupling between the motions of the subdomains.

The free energy profile with respect to $\Delta RMSD$ of the activation pathway of CpxA was obtained using umbrella sampling and WHAM. Similarly to WalK histidine kinase, a steep rise in free energy is observed as the kinase core activates. However, the difference in Gibbs free energy was 4.5 kcal/mol larger than that of WalK activation. The lower stability could possibly be due to the net loss in interdomain nonbonded interactions in CpxA, while WalK experiences a net gain in CA-DHp contacts. Additionally, a greater rotation motion is required to reorient the CA domain to bring the γ-phosphate of ATP in close proximity to the phosphorylable histidine in the DHp domain. This is necessary for the autophosphorylation reaction to occur. It is important to acknowledge that the HAMP domain was partially present in the CpxA simulation model, and so the motions at this region also contributed to the free energies. In fact, the local helical bending at HAMP termini was accounted for in PC1 and PC2. As these are rather short segments however, their contributions to the difference in free energy are probably minor.

All in all, the free energy barrier associated with the conformation changes of CpxA activation is significantly larger than that of the autophosphorylation reaction reported in a previous study [162]. This suggests that the large-scale conformational changes is the rate-limiting step of the signalling mechanism at the kinase core.

6. Summary

In this work, various computational approaches have been used to study the conformational dynamics that are essential in histidine kinase activity. The two test systems were WalK and CpxA which are HKs that phosphorylate in *cis* and *trans*, respectively.

In chapter 3, a novel dual-basin structure-based model of WalK histidine kinase was developed to simulate the conformational pathways between two known states. The dual-basin SBM was constructed by perturbing the single-basin SBM of the open, inactive conformation of HK with active-state-specific native contact energies. Simulations with this modified coarse-grained force-field provided insights into the back-and-forth transitions between the two states at a remarkably low computation cost. Principal component analysis enabled us to filter the global, collective motions from the fast, local motions. It was found that the rotational motion of the catalytic ATP-binding domain with respect to the DHp helical bundle was the dominant collective motion. This supports the hypothesis that the kinase core activates by a "walking" mechanism, whereby CA-DHp non-bonded interactions are formed and broken simultaneously as the CA domain "walks" along the DHp domain [152, 162].

The active-state-specific native contact pairs were further categorized into subgroups based on which domain each atom in the pair belongs to. To assess their impact on the transitions, the relative contact strengths of each subgroup was varied. Markov state modelling and PCCA+ of the SBM simulations with the different sets of parameters revealed the changes to the stabilities of the two basins (i.e. the metastable states). Modifications to the interdomain contact energies lead to pronounced changes in the conformational transitions. When the interdomain contact energies were decreased, only a single metastable state was observed. Decreasing the intradomain contact energies, on the other hand, had minimal impact on stationary probabilities of the metastable states. This further highlights the significance of the formation of interdomain contacts.

Conformations representative of the end states were extracted from the dual-basin SBM simulations and the all-atom models were reconstructed. The free energy profile calculated

using these models was in agreement with that obtained with the crystal structures. Obtaining quantitatively similar results starting from the coarser representation confirms the robustness of this model. Despite comprising only α-carbon atoms, the presented model enabled us the gain an intuition into the conformational dynamics of WalK's kinase core. In future work, more complexity (e.g. electrostatics or hydrogen bonding) could be incorporated into the force-field for a more accurate representation of the dynamics. When an x-ray structure that contains both the kinase core and the linker domain becomes available, this model could be further extended.

After building an intuition with the CG model in chapter 3, I used the more detailed, all-atom representation of the systems in the proceeding chapters. This allowed us to gain more detailed insights into the signalling mechanisms of HKs.

In chapter 4, I explored the relationship between the conformational motions of the DHp helical bundle and the CA domain. I proposed and analyzed two possible activation pathways of WalK HK. In the first pathway, the large-scale conformational changes necessary for activation at the DHp and CA domains are induced simultaneously. The free energy profile revealed that the asymmetrical, active conformation is highly unstable relative to the symmetrical, inactive conformation. The free energy barrier obtained is in agreement with a previous study by Olivieri et al [152]. Detailed contact map analysis along the transition pathway demonstrated that weakly held interdomain non-bonded interactions were forming and breaking in a "walking" mechanism.

In the second activation mechanism, a biasing potential was placed exclusively on the atoms in the DHp helical bundle. The steered MD trajectory with the DHp-only biasing revealed that the structural rearrangement of the helical bundle, concurrently induced the rotation of the CA domain. Although the rotational angle was 7° less than the value calculated with the crystal structures, the phosphorylatable histidine was brought into an ideal distance to the ATP-lid. Therefore, the phosphoryl transfer reaction essential for signal transduction can occur. This suggests that the conformational dynamics of the individual domains of the kinase core are tightly coupled.

In chapter 5, the activation mechanism of CpxA, a *trans*-phosphorylating HK was examined. In a previous study, the conformational transitions were simulated with a coarse-grained model [162]. In this work, I performed enhanced sampling MD simulations with the atomistic models to provide a more detailed analysis into the activation mechanism. PCA of the trajectory showed that the slowest collective motion was the counter-clockwise rotation of the CA domain, and the simultaneous helical bending in DHp-HAMP termini.

Despite having different modes of phosphorylation, these structural rearrangements are similar to that observed in WalK. Moreover, I reported the free energy profile of the conformational transitions. As seen in the case of WalK, the active conformation was highly unstable. However, the energetic barrier was slightly higher for CpxA which is possibly due to the net loss in CA-DHp interactions. Furthermore, the estimated free energy barrier associated with the conformation changes from the US simulations is considerably larger than that of the autophosphorylation reaction reported by Marsico et al [162]. This strongly suggests that the large-scale conformational changes is the rate-limiting step of the signalling mechanism at the kinase core.

Bibliography

[1] L. E. Ulrich and I. B. Zhulin. "The MiST2 database: a comprehensive genomics resource on microbial signal transduction". *Nucleic Acids Research* 38 (2010), D401–D407.

[2] P. Ortet, D. E. Whitworth, C. Santaella, W. Achouak, and M. Barakat. "P2CS: updates of the prokaryotic two-component systems database". *Nucleic Acids Research* 43 (2015), D536–D541.

[3] M. Igarashi et al. "Waldiomycin, a novel WalK-histidine kinase inhibitor from Streptomyces sp. MK844-mF10". *The Journal of Antibiotics* 66 (2013), 459–464.

[4] T. Watanabe, A. Okada, Y. Gotoh, and R. Utsumi. "Inhibitors targeting two-component signal transduction". *Bacterial Signal Transduction: Networks and Drug Targets* (2008), 229–236.

[5] Y. Gotoh et al. "Novel antibacterial compounds specifically targeting the essential WalR response regulator". *The Journal of Antibiotics* 63 (2010), 127–134.

[6] Y. Gotoh, Y. Eguchi, T. Watanabe, S. Okamoto, A. Doi, and R. Utsumi. "Two-component signal transduction as potential drug targets in pathogenic bacteria". *Current Opinion in Microbiology* 13 (2010), 232–239.

[7] A. Okada et al. "Walkmycin B targets WalK (YycG), a histidine kinase essential for bacterial cell growth". *The Journal of Antibiotics* 63 (2010), 89–94.

[8] A. E. Dago, A. Schug, A. Procaccini, J. A. Hoch, M. Weigt, and H. Szurmant. "Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis". *Proceedings of the National Academy of Sciences* 109 (2012), E1733–E1742.

[9] M. M. Igo, A. J. Ninfa, J. B. Stock, and T. J. Silhavy. "Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor." *Genes & Development* 3 (1989), 1725–1734.

[10] Y. Liu et al. “A pH-gated conformational switch regulates the phosphatase activity of bifunctional HisKA-family histidine kinases”. *Nature Communications* 8 (2017), 1–10.

[11] G. Rivera-Cancel, W. Ko, D. R. Tomchick, F. Correa, and K. H. Gardner. “Full-length structure of a monomeric histidine kinase reveals basis for sensory regulation”. *Proceedings of the National Academy of Sciences* 111 (2014), 17839–17844.

[12] I. Dikiy et al. “Insights into histidine kinase activation mechanisms from the monomeric blue light sensor EL346”. *Proceedings of the National Academy of Sciences* 116 (2019), 4963–4972.

[13] D. Albanesi et al. “Structural plasticity and catalysis regulation of a thermosensor histidine kinase”. *Proceedings of the National Academy of Sciences* 106 (2009), 16185–16190.

[14] L. C. Wang, L. K. Morgan, P. Godakumbura, L. J. Kenney, and G. S. Anand. “The inner membrane histidine kinase EnvZ senses osmolality via helix-coil transitions in the cytoplasm”. *The EMBO Journal* 31 (2012), 2648–2659.

[15] A. M. Bilwes, L. A. Alex, B. R. Crane, and M. I. Simon. “Structure of CheA, a signal-transducing histidine kinase”. *Cell* 96 (1999), 131–141.

[16] A. R. Greenswag, A. Muok, X. Li, and B. R. Crane. “Conformational transitions that enable histidine kinase autophosphorylation and receptor array integration”. *Journal of Molecular Biology* 427 (2015), 3890–3907.

[17] J. Cheung and W. A. Hendrickson. “Sensor domains of two-component regulatory systems”. *Current Opinion in Microbiology* 13 (2010), 116–123.

[18] J. Cheung and W. A. Hendrickson. “Structural analysis of ligand stimulation of the histidine kinase NarX”. *Structure* 17 (2009), 190–201.

[19] J. O. Moore and W. A. Hendrickson. “An asymmetry-to-symmetry switch in signal transmission by the histidine kinase receptor for TMAO”. *Structure* 20 (2012), 729–741.

[20] J. Cheung, C. A. Bingman, M. Reyngold, W. A. Hendrickson, and C. D. Waldburger. “Crystal structure of a functional dimer of the PhoQ sensor domain”. *Journal of Biological Chemistry* 283 (2008), 13762–13770.

[21] M. Sevvana et al. “A ligand-induced switch in the periplasmic domain of sensor histidine kinase CitA”. *Journal of Molecular Biology* 377 (2008), 512–523.

[22] L. Aravind and C. P. Ponting. “The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins”. *FEMS Microbiology Letters* 176 (1999), 111–116.

[23] H. U. Ferris et al. “The mechanisms of HAMP-mediated signaling in transmembrane receptors”. *Structure* 19 (2011), 378–385.

[24] C. P. Zschiedrich, V. Keidel, and H. Szurmant. “Molecular mechanisms of two-component signal transduction”. *Journal of Molecular Biology* 428 (2016), 3752–3775.

[25] A. Möglich, R. A. Ayers, and K. Moffat. “Structure and signaling mechanism of Per-ARNT-Sim domains”. *Structure* 17 (2009), 1282–1294.

[26] Y. J. Ho, L. M. Burden, and J. H. Hurley. “Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor”. *The EMBO Journal* 19 (2000), 5288–5299.

[27] J. D. Batchelor, P. S. Lee, A. C. Wang, M. Doucleff, and D. E. Wemmer. “Structural mechanism of GAF-regulated σ54 activators from Aquifex aeolicus”. *Journal of Molecular Biology* 425 (2013), 156–170.

[28] C. Wang et al. “Mechanistic insights revealed by the crystal structure of a histidine kinase with signal transducer and sensor domains”. *PLoS Biology* 11 (2013), e1001493.

[29] M. Hulko et al. “The HAMP domain structure implies helix rotation in transmembrane signaling”. *Cell* 126 (2006), 929–940.

[30] M. V. Airola, K. J. Watts, A. M. Bilwes, and B. R. Crane. “Structure of concatenated HAMP domains provides a mechanism for signal transduction”. *Structure* 18 (2010), 436–448.

[31] Y. Cai et al. “Conformational dynamics of the essential sensor histidine kinase WalK”. *Acta Crystallographica Section D: Structural Biology* 73 (2017), 793–803.

[32] D. Kim and S. Forst. “Genomic analysis of the histidine kinase family in bacteria and archaea”. *Microbiology* 147 (2001), 1197–1212.

[33] P. Casino, L. Miguel-Romero, and A. Marina. “Visualizing autophosphorylation in histidine kinases”. *Nature Communications* 5 (2014), 1–12.

[34] M. P. Bhate, K. S. Molnar, M. Goulian, and W. F. DeGrado. “Signal transduction in histidine kinases: insights from new structures”. *Structure* 23 (2015), 981–994.

[35] O. Ashenberg, A. E. Keating, and M. T. Laub. "Helix bundle loops determine whether histidine kinases autophosphorylate in cis or in trans". *Journal of Molecular Biology* 425 (2013), 1198–1209.

[36] A. E. Mechaly, S. S. Diaz, N. Sassoon, A. Buschiazzo, J.-M. Betton, and P. M. Alzari. "Structural coupling between autokinase and phosphotransferase reactions in a bacterial histidine kinase". *Structure* 25 (2017), 939–944.

[37] R. Gao, S. Bouillet, and A. M. Stock. "Structural basis of response regulator function". *Annual Review of Microbiology* 73 (2019), 175–197.

[38] T. E. Quax et al. "Structure and function of the archaeal response regulator CheY". *Proceedings of the National Academy of Sciences* 115 (2018), E1259–E1268.

[39] J. Zapf, U. Sen, J. A. Hoch, K. I. Varughese, et al. "A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction". *Structure* 8 (2000), 851–862.

[40] M. Solà, F. X. Gomis-Rüth, L. Serrano, A. González, and M. Coll. "Three-dimensional crystal structure of the transcription factor PhoB receiver domain". *Journal of Molecular Biology* 285 (1999), 675–687.

[41] A. G. Blanco, M. Sola, F. X. Gomis-Rüth, and M. Coll. "Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator". *Structure* 10 (2002), 701–713.

[42] S. Chakraborty, J. Sivaraman, K. Y. Leung, and Y. Mok. "Two-component PhoB-PhoR regulatory system and ferric uptake regulator sense phosphate and iron to control virulence genes in type III and VI secretion systems of Edwardsiella tarda". *Journal of Biological Chemistry* 286 (2011), 39417–39430.

[43] Y.-L. Tzeng and J. A. Hoch. "Molecular recognition in signal transduction: the interaction surfaces of the Spo0F response regulator with its cognate phosphorelay proteins revealed by alanine scanning mutagenesis". *Journal of Molecular Biology* 272 (1997), 200–212.

[44] K. Stephenson and J. A. Hoch. "Evolution of signalling in the sporulation phosphorelay". *Molecular Microbiology* 46 (2002), 297–304.

[45] K. I. Varughese, X. Z. Zhou, J. M. Whiteley, J. A. Hoch, et al. "Formation of a novel four-helix bundle and molecular recognition sites by dimerization of a response regulator phosphotransferase". *Molecular Cell* 2 (1998), 485–493.

[46] J. A. Hoch and K. Varughese. “Keeping signals straight in phosphorelay signal transduction”. *Journal of Bacteriology* 183 (2001), 4941–4949.

[47] A. Buschiazzo and F. Trajtenberg. “Two-component sensing and regulation: how do histidine kinases talk with response regulators at the molecular level?” *Annual Review of Microbiology* 73 (2019), 507–528.

[48] P. Casino, V. Rubio, and A. Marina. “Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction”. *Cell* 139 (2009), 325–336.

[49] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije. “LINCS: a linear constraint solver for molecular simulations”. *Journal of Computational Chemistry* 18 (1997), 1463–1472.

[50] E. Lindahl, M. Abraham, B. Hess, and D. van der Spoel. *GROMACS 2020 Source code*. Version 2020. Jan. 2020. DOI: 10.5281/zenodo.3562495. URL: https://doi.org/10.5281/zenodo.3562495.

[51] R. W. Hockney, S. Goel, and J. Eastwood. “Quiet high-resolution computer models of a plasma”. *Journal of Computational Physics* 14 (1974), 148–158.

[52] L. Verlet. “Computer “experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules”. *Physical Review* 159 (1967), 98.

[53] T. Schlick. *Molecular modeling and simulation: an interdisciplinary guide*. Vol. 2. Springer, 2010.

[54] G. Bussi, D. Donadio, and M. Parrinello. “Canonical sampling through velocity rescaling”. *The Journal of Chemical Physics* 126 (2007), 014101.

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

[56] H. Berendsen. “Transport properties computed by linear response through weak coupling to a bath”. *Computer Simulation in Materials Science*. Springer, 1991, 139–155.

[57] S. Nosé. “A molecular dynamics method for simulations in the canonical ensemble”. *Molecular Physics* 52 (1984), 255–268.

[58] A. D. MacKerell Jr. et al. “All-atom empirical potential for molecular modeling and dynamics studies of proteins”. *The Journal of Physical Chemistry B* 102 (1998), 3586–3616.

[59] K. Lindorff-Larsen et al. “Improved side-chain torsion potentials for the Amber ff99SB protein force field”. *Proteins: Structure, Function, and Bioinformatics* 78 (2010), 1950–1958.

[60] H. M. Berman et al. “The protein data bank”. *Nucleic Acids Research* 28 (2000), 235–242.

[61] M. Patra, M. Karttunen, M. T. Hyvönen, E. Falck, P. Lindqvist, and I. Vattulainen. “Molecular dynamics simulations of lipid bilayers: major artifacts due to truncating electrostatic interactions”. *Biophysical Journal* 84 (2003), 3636–3645.

[62] 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 (1993), 10089–10092.

[63] A. Kolinski, L. Jaroszewski, P. Rotkiewicz, and J. Skolnick. “An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side group centers of mass”. *The Journal of Physical Chemistry B* 102 (1998), 4628–4637.

[64] A. Koliński et al. “Protein modeling and structure prediction with a reduced representation”. *Acta Biochimica Polonica* 51 (2004).

[65] A. Liwo et al. “A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions”. *Journal of Molecular Modeling* 20 (2014), 1–15.

[66] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski. “Coarse-grained protein models and their applications”. *Chemical Reviews* 116 (2016), 7898–7936.

[67] Y. Ueda, H. Taketomi, and N. Gō. “Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme”. *Biopolymers: Original Research on Biomolecules* 17 (1978), 1531–1548.

[68] J. Karanicolas and C. L. Brooks III. “Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions”. *Journal of Molecular Biology* 334 (2003), 309–325.

[69] L. Brand and M. L. Johnson. *Numerical Computer Methods, Part D*. Elsevier, 2004.

[70] M. Levitt and A. Warshel. “Computer simulation of protein folding”. *Nature* 253 (1975), 694–698.

[71] P. E. Leopold, M. Montal, and J. N. Onuchic. “Protein folding funnels: a kinetic approach to the sequence-structure relationship.” *Proceedings of the National Academy of Sciences* 89 (1992), 8721–8725.

[72] A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant. “High-resolution protein complexes from integrating genomic information with molecular simulation”. *Proceedings of the National Academy of Sciences* 106 (2009), 22124–22129.

[73] E. L. Baxter, P. A. Jennings, and J. N. Onuchic. “Strand swapping regulates the iron-sulfur cluster in the diabetes drug target mitoNEET”. *Proceedings of the National Academy of Sciences* 109 (2012), 1955–1960.

[74] V. H. Giri Rao and S. Gosavi. “Structural Perturbations Present in the Folding Cores of Interleukin-33 and Interleukin-1β Correlate to Differences in Their Function”. *The Journal of Physical Chemistry B* 119 (2015), 11203–11214.

[75] M. B. Borgia et al. “Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins”. *Nature* 474 (2011), 662–665.

[76] P. C. Whitford and K. Y. Sanbonmatsu. “Simulating movement of tRNA through the ribosome during hybrid-state formation”. *The Journal of Chemical Physics* 139 (2013), 09B619_1.

[77] L. L. Chavez, J. N. Onuchic, and C. Clementi. “Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates”. *Journal of the American Chemical Society* 126 (2004), 8426–8432.

[78] C. Clementi, H. Nymeyer, and J. N. Onuchic. “Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? An investigation for small globular proteins”. *Journal of Molecular Biology* 298 (2000), 937–953.

[79] M. F. Rey-Stolle, M. Enciso, and A. Rey. “Topology-based models and NMR structures in protein folding simulations”. *Journal of Computational Chemistry* 30 (2009), 1212–1219.

[80] I. Reinartz et al. “Simulation of FRET dyes allows quantitative comparison against experimental data”. *The Journal of Chemical Physics* 148 (2018), 123321.

[81] J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic. “SMOG@ ctbp: simplified deployment of structure-based models in GROMACS”. *Nucleic Acids Research* 38 (2010), W657–W661.

[82] J. K. Noel, P. C. Whitford, and J. N. Onuchic. "The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function". *The Journal of Physical Chemistry B* 116 (2012), 8692–8702.

[83] C. McPhalen and M. James. "Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds". *Biochemistry* 26 (1987), 261–269.

[84] E. Lindahl, M. Abraham, B. Hess, and D. van der Spoel. *GROMACS 2020 Manual*. Version 2020. Jan. 2020. DOI: 10.5281/zenodo.3562512. URL: https://doi.org/10.5281/zenodo.3562512.

[85] J. Jackson, K. Nguyen, and P. C. Whitford. "Exploring the balance between folding and functional dynamics in proteins and RNA". *International Journal of Molecular Sciences* 16 (2015), 6868–6889.

[86] A. Shinobu, C. Kobayashi, Y. Matsunaga, and Y. Sugita. "Building a macro-mixing dual-basin Gō model using the Multistate Bennett Acceptance Ratio". *Biophysics and physicobiology* 16 (2019), 310–321.

[87] A. Shinobu, C. Kobayashi, Y. Matsunaga, and Y. Sugita. "Coarse-Grained Modeling of Multiple Pathways in Conformational Transitions of Multi-Domain Proteins". *Journal of Chemical Information and Modeling* 61 (2021), 2427–2443.

[88] C. Kobayashi, Y. Matsunaga, R. Koike, M. Ota, and Y. Sugita. "Domain motion enhanced (DoME) model for efficient conformational sampling of multidomain proteins". *The Journal of Physical Chemistry B* 119 (2015), 14584–14593.

[89] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic. "Conformational transitions of adenylate kinase: switching by cracking". *Journal of Molecular Biology* 366 (2007), 1661–1671.

[90] P. C. Whitford, S. Gosavi, and J. N. Onuchic. "Conformational transitions in adenylate kinase: allosteric communication reduces misligation". *Journal of Biological Chemistry* 283 (2008), 2042–2048.

[91] K.-i. Okazaki and S. Takada. "Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms". *Proceedings of the National Academy of Sciences* 105 (2008), 11182–11187.

[92] M. D. Daily, G. N. Phillips Jr, and Q. Cui. "Many local motions cooperate to produce the adenylate kinase conformational transition". *Journal of Molecular Biology* 400 (2010), 618–631.

[93] S. Yang and B. Roux. "Src kinase conformational activation: thermodynamics, pathways, and mechanisms". *PLoS Computational Biology* 4 (2008), e1000047.

[94] R. B. Best, Y.-G. Chen, and G. Hummer. "Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of arc repressor". *Structure* 13 (2005), 1755–1763.

[95] K.-i. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes. "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations". *Proceedings of the National Academy of Sciences* 103 (2006), 11844–11849.

[96] H. Huang, R. Zhao, B. M. Dickson, R. D. Skeel, and C. B. Post. "αC helix as a switch in the conformational transition of Src/CDK-like kinase domains". *The Journal of Physical Chemistry B* 116 (2012), 4465–4475.

[97] J. P. Singh, P. C. Whitford, N. Hayre, J. Onuchic, and D. L. Cox. "Massive conformation change in the prion protein: Using dual-basin structure-based models to find misfolding pathways". *Proteins: Structure, Function, and Bioinformatics* 80 (2012), 1299–1307.

[98] B. L. Moore, L. A. Kelley, J. Barber, J. W. Murray, and J. T. MacDonald. "High–quality protein backbone reconstruction from alpha carbons using Gaussian mixture models". *Journal of Computational Chemistry* 34 (2013), 1881–1889.

[99] S. A. Adcock. "Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield". *Journal of Computational Chemistry* 25 (2004), 16–27.

[100] Y. Iwata, A. Kasuya, and S. Miyamoto. "An efficient method for reconstructing protein backbones from α-carbon coordinates". *Journal of Molecular Graphics and Modelling* 21 (2002), 119–128.

[101] P. Rotkiewicz and J. Skolnick. "Fast procedure for reconstruction of full-atom protein models from reduced representations". *Journal of Computational Chemistry* 29 (2008), 1460–1465.

[102] S. Subramaniam and A. Senes. "Backbone dependency further improves side chain prediction efficiency in the Energy-based Conformer Library (bEBL)". *Proteins: Structure, Function, and Bioinformatics* 82 (2014), 3177–3187.

[103] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack Jr. "Improved prediction of protein side-chain conformations with SCWRL4". *Proteins: Structure, Function, and Bioinformatics* 77 (2009), 778–795.

[104] Y. Naritomi and S. Fuchigami. "Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis". *The Journal of Chemical Physics* 139 (2013), 12B605_1.

[105] F. Yao, J. Coquery, and K.-A. Lê Cao. "Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets". *BMC Bioinformatics* 13 (2012), 1–15.

[106] H. Wan and V. A. Voelz. "Adaptive Markov state model estimation using short reseeding trajectories". *The Journal of Chemical Physics* 152 (2020), 024103.

[107] J.-H. Prinz et al. "Markov models of molecular kinetics: Generation and validation". *The Journal of Chemical Physics* 134 (2011), 174105.

[108] I. T. Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[109] C. C. David and D. J. Jacobs. "Principal component analysis: a method for determining the essential dynamics of proteins". *Protein Dynamics*. Springer, 2014, 193–226.

[110] I. Daidone and A. Amadei. "Essential dynamics: foundation and applications". *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2 (2012), 762–770.

[111] N. A. Jonniya, M. F. Sk, and P. Kar. "Investigating phosphorylation-induced conformational changes in WNK1 kinase by molecular dynamics simulations". *American Chemical Society Omega* 4 (2019), 17404–17416.

[112] J. Chen. "Functional roles of magnesium binding to extracellular signal-regulated kinase 2 explored by molecular dynamics simulations and principal component analysis". *Journal of Biomolecular Structure and Dynamics* 36 (2018), 351–361.

[113] P. Mishra and S. Günther. "New insights into the structural dynamics of the kinase JNK3". *Scientific Reports* 8 (2018), 1–13.

[114] M. Martinez, N. Duclert-Savatier, J.-M. Betton, P. M. Alzari, M. Nilges, and T. E. Malliavin. "Modification in hydrophobic packing of HAMP domain induces a destabilization of the auto-phosphorylation site in the histidine kinase CpxA". *Biopolymers* 105 (2016), 670–682.

[115] G. Novikov, V. Sivozhelezov, and K. Shaitan. "Influence of orthosteric ligand binding on the conformational dynamics of the β-2-adrenergic receptor via essential dynamics sampling simulation". *Molecular Biology* 48 (2014), 399–413.

[116] D. Narzi, I. Daidone, A. Amadei, and A. Di Nola. "Protein folding pathways revealed by essential dynamics sampling". *Journal of Chemical Theory and Computation* 4 (2008), 1940–1948.

[117] G. G. Maisuradze, A. Liwo, and H. A. Scheraga. "Principal component analysis for protein folding dynamics". *Journal of Molecular Biology* 385 (2009), 312–329.

[118] A. Khan et al. “Structural insights into the mechanism of RNA recognition by the N-terminal RNA-binding domain of the SARS-CoV-2 nucleocapsid phosphoprotein”. *Computational and Structural Biotechnology Journal* 18 (2020), 2174–2184.

[119] A. Amadei, A. B. Linssen, and H. J. Berendsen. “Essential dynamics of proteins”. *Proteins: Structure, Function, and Bioinformatics* 17 (1993), 412–425.

[120] Q. Qiao, G. R. Bowman, and X. Huang. “Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation”. *Journal of the American Chemical Society* 135 (2013), 16092–16101.

[121] V. A. Voelz et al. “Markov state models of millisecond folder ACBP reveals new views of the folding reaction”. *Biophysical Journal* 100 (2011), 515a.

[122] D. Huang and A. Caflisch. “The free energy landscape of small molecule unbinding”. *PLoS Computational Biology* 7 (2011), e1002002.

[123] M. K. Scherer et al. “PyEMMA 2: A software package for estimation, validation, and analysis of Markov models”. *Journal of Chemical Theory and Computation* 11 (2015), 5525–5542.

[124] S. Röblitz and M. Weber. “Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification”. *Advances in Data Analysis and Classification* 7 (2013), 147–179.

[125] P. Deuflhard and M. Weber. “Robust Perron cluster analysis in conformation dynamics”. *Linear Algebra and its Applications* 398 (2005), 161–184.

[126] K. Arora and C. L. Brooks. “Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism”. *Proceedings of the National Academy of Sciences* 104 (2007), 18496–18501.

[127] B. Bouvier, K. Zakrzewska, and R. Lavery. “Protein–DNA recognition triggered by a DNA conformational switch”. *Angewandte Chemie International Edition* 50 (2011), 6516–6518.

[128] H. Yang et al. “Simulations of cellulose synthesis initiation and termination in bacteria”. *The Journal of Physical Chemistry B* 123 (2019), 3699–3705.

[129] M. Y. Lobanov, N. Bogatyreva, and O. Galzitskaya. “Radius of gyration as an indicator of protein structure compactness”. *Molecular Biology* 42 (2008), 623–628.

[130] G. M. Torrie and J. P. Valleau. “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. *Journal of Computational Physics* 23 (1977), 187–199.

[131] H. Lu and K. Schulten. “Steered molecular dynamics simulations of force-induced protein domain unfolding”. *Proteins: Structure, Function, and Bioinformatics* 35 (1999), 453–463.

[132] A. Laio and M. Parrinello. “Escaping free-energy minima”. *Proceedings of the National Academy of Sciences* 99 (2002), 12562–12566.

[133] L. Maragliano and E. Vanden-Eijnden. “A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations”. *Chemical Physics Letters* 426 (2006), 168–175.

[134] Y. Sugita and Y. Okamoto. “Replica-exchange molecular dynamics method for protein folding”. *Chemical Physics Letters* 314 (1999), 141–151.

[135] D. Hamelberg, J. Mongan, and J. A. McCammon. “Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules”. *The Journal of Chemical Physics* 120 (2004), 11919–11929.

[136] X. Wu and S. Wang. “Self-guided molecular dynamics simulation for efficient conformational search”. *The Journal of Physical Chemistry B* 102 (1998), 7238–7250.

[137] C. Jarzynski. “Nonequilibrium equality for free energy differences”. *Physical Review Letters* 78 (1997), 2690.

[138] M. Souaille and B. Roux. “Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations”. *Computer Physics Communications* 135 (2001), 40–57.

[139] 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 (1992), 1011–1021.

[140] P. H. B. Ferreira, F. C. Freitas, M. E. McCully, G. G. Slade, and R. J. de Oliveira. “The role of electrostatics and folding kinetics on the thermostability of homologous cold shock proteins”. *Journal of Chemical Information and Modeling* 60 (2020), 546–561.

[141] V. G. Contessoto, V. M. De Oliveira, S. J. De Carvalho, L. C. Oliveira, and V. B. Leite. “NTL9 folding at constant pH: the importance of electrostatic interaction and pH dependence”. *Journal of Chemical Theory and Computation* 12 (2016), 3270–3277.

[142] A. Okuda et al. “Solution structure of multi-domain protein ER-60 studied by aggregation-free SAXS and coarse-grained-MD simulation”. *Scientific Reports* 11 (2021), 1–13.

[143] Q. Wang et al. “Probing the Allosteric Inhibition Mechanism of a Spike Protein Using Molecular Dynamics Simulations and Active Compound Identifications”. *Journal of Medicinal Chemistry* (2021).

[144] C. Fabret and J. A. Hoch. “A two-component signal transduction system essential for growth of Bacillus subtilis: implications for anti-infective therapy”. *Journal of Bacteriology* 180 (1998), 6375–6383.

[145] P. K. Martin, T. Li, D. Sun, D. P. Biek, and M. B. Schmid. “Role in cell permeability of an essential two-component system in Staphylococcus aureus”. *Journal of Bacteriology* 181 (1999), 3666–3673.

[146] E. F. Pettersen et al. “UCSF Chimera—a visualization system for exploratory research and analysis”. *Journal of Computational Chemistry* 25 (2004), 1605–1612.

[147] B. Webb and A. Sali. “Comparative protein structure modeling using MODELLER”. *Current Protocols in Bioinformatics* 54 (2016), 5–6.

[148] E. Lindahl, M. Abraham, B. Hess, and D. van der Spoel. *GROMACS 2019.2 Manual*. Version 2019.2. Apr. 2019. DOI: 10.5281/zenodo.2636383. URL: https://doi.org/10.5281/zenodo.2636383.

[149] M. Bonomi. “Promoting transparency and reproducibility in enhanced molecular simulations”. *Nature Methods* 16 (2019), 670–673.

[150] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. “Comparison of simple potential functions for simulating liquid water”. *The Journal of Chemical Physics* 79 (1983), 926–935.

[151] A. Grossfield. *WHAM: the weighted histogram analysis method*. http://membrane.urmc.rochester.edu/wordpress/?page_id=126. Accessed: 28-02-2022.

[152] F. A. Olivieri et al. “Conformational and Reaction Dynamic Coupling in Histidine Kinases: Insights from Hybrid QM/MM Simulations”. *Journal of Chemical Information and Modeling* 60 (2020), 833–842.

[153] V. M. de Oliveira, V. de Godoi Contessoto, F. B. da Silva, D. L. Z. Caetano, S. J. de Carvalho, and V. B. P. Leite. “Effects of pH and salt concentration on stability of a protein G variant using coarse-grained models”. *Biophysical Journal* 114 (2018), 65–75.

[154] F. O. Tzul, K. L. Schweiker, and G. I. Makhatadze. “Modulation of folding energy landscape by charge–charge interactions: Linking experiments with computational modeling”. *Proceedings of the National Academy of Sciences* 112 (2015), E259–E266.

[155] X. Chu et al. “Dynamic conformational change regulates the protein-DNA recognition: an investigation on binding of a Y-family polymerase to its target DNA”. *PLoS Computational Biology* 10 (2014), e1003804.

[156] A. E. Mechaly, N. Sassoon, J.-M. Betton, and P. M. Alzari. “Segmental helical motions and dynamical asymmetry modulate histidine kinase autophosphorylation”. *PLoS Biology* 12 (2014), e1001776.

[157] I. Gushchin, P. Orekhov, I. Melnikov, V. Polovinkin, A. Yuzhakova, and V. Gordeliy. “Sensor histidine kinase NarQ activates via helical rotation, diagonal scissoring, and eventually piston-like shifts”. *International Journal of Molecular Sciences* 21 (2020), 3110.

[158] P. K. Párraga Solórzano, A. C. Shupe, and T. E. Kehl-Fie. “The Sensor Histidine Kinase ArlS Is Necessary for Staphylococcus aureus To Activate ArlR in Response to Nutrient Availability”. *Journal of Bacteriology* 203 (2021), e00422–21.

[159] A. Koh, M. J. Gibbon, M. W. Van der Kamp, C. R. Pudney, and S. Gebhard. “Conformation control of the histidine kinase BceS of Bacillus subtilis by its cognate ABC-transporter facilitates need-based activation of antibiotic resistance”. *Molecular Microbiology* 115 (2021), 157–174.

[160] O. Berntsson et al. “Sequential conformational transitions and α-helical supercoiling regulate a sensor histidine kinase”. *Nature Communications* 8 (2017), 1–8.

[161] S. Wingbermühle and L. V. Schäfer. “Capturing the Flexibility of a Protein–Ligand Complex: Binding Free Energies from Different Enhanced Sampling Techniques”. *Journal of Chemical Theory and Computation* 16 (2020), 4615–4630.

[162] F. Marsico et al. “Multiscale approach to the activation and phosphotransfer mechanism of CpxA histidine kinase reveals a tight coupling between conformational and chemical steps”. *Biochemical and Biophysical Research Communications* 498 (2018), 305–312.

[163] S. Nakayama and H. Watanabe. “Involvement of cpxA, a sensor of a two-component regulatory system, in the pH-dependent regulation of expression of Shigella sonnei virF gene”. *Journal of Bacteriology* 177 (1995), 5062–5069.

A. Appendices of Chapter 3

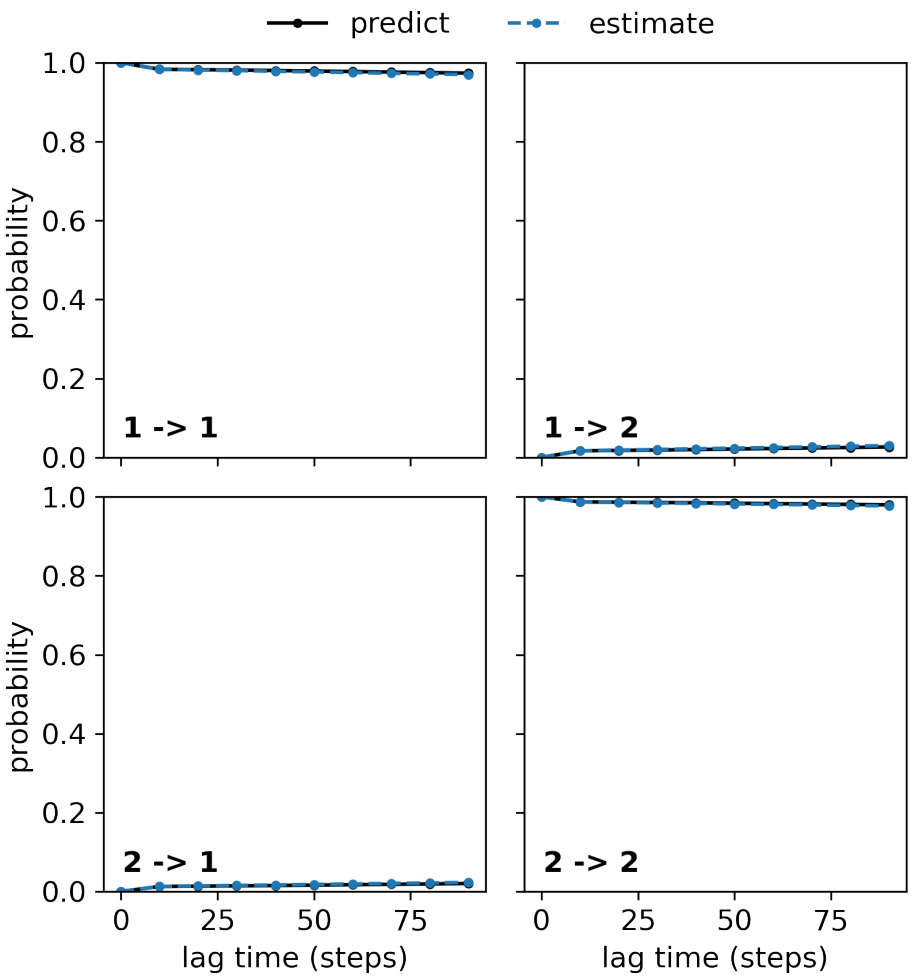


Figure A.1.: Chapman-Kolmogorov (CK) test for the dual-basin SBM simulation of WalK histidine kinase. The weighting parameters λ_α and λ_β were set to 0.85.

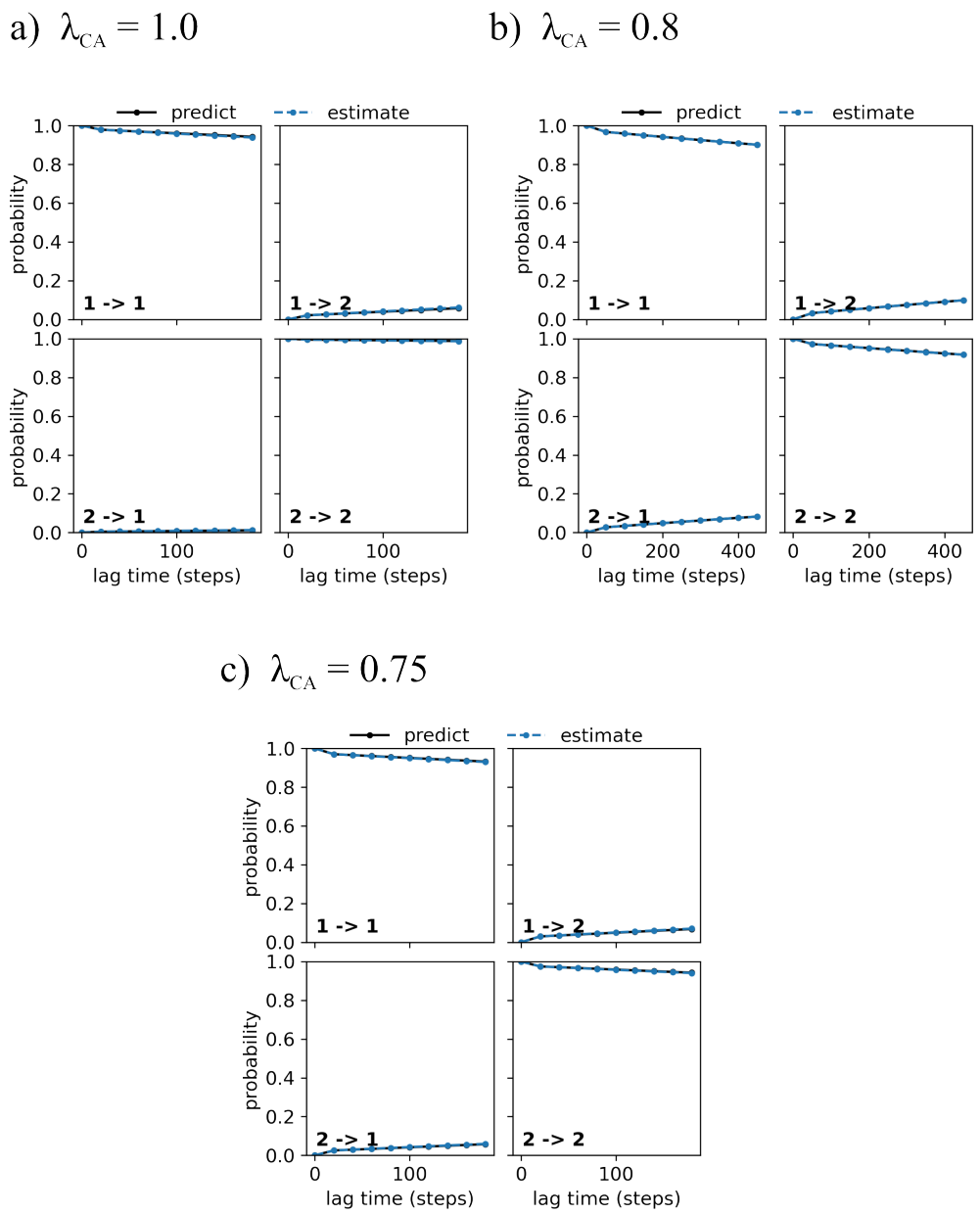


Figure A.2.: Chapman-Kolmogorov (CK) test for the dual-basin SBM simulations of WalK histidine kinase. The weighting parameters λ_{α}, λ_{DHp} and λ_{CA-DHp} were set to 0.85 in all simulations and λ_{CA} was set to (a) 1.0, (b) 0.8 and (c) 0.75 in each simulation.

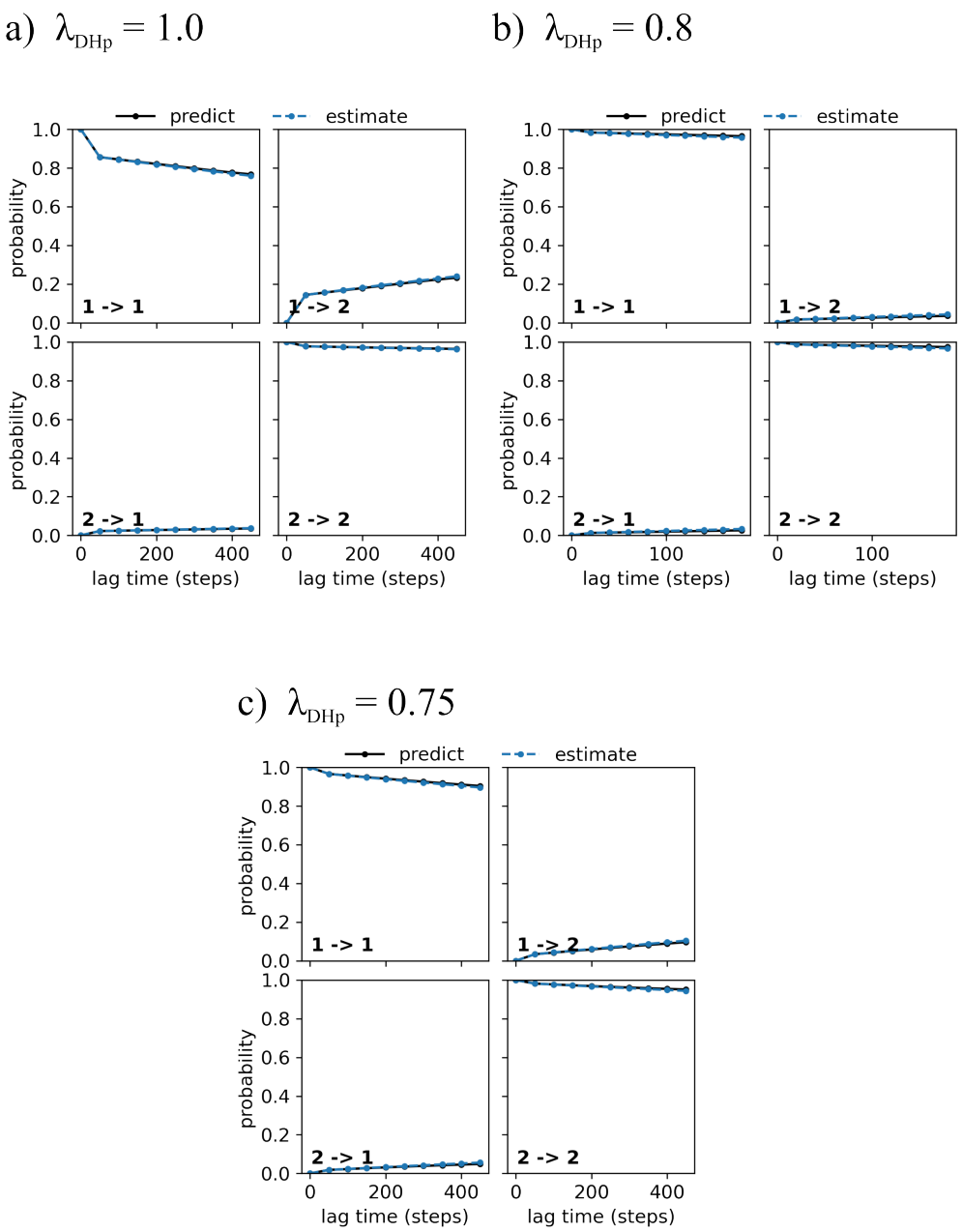


Figure A.3.: Chapman-Kolmogorov (CK) test for the dual-basin SBM simulations of WalK histidine kinase. The weighting parameters λ_{α}, λ_{CA} and λ_{CA-DHp} were set to 0.85 in all simulations and λ_{DHp} was set to (a) 1.0, (b) 0.8 and (c) 0.75 in each simulation.

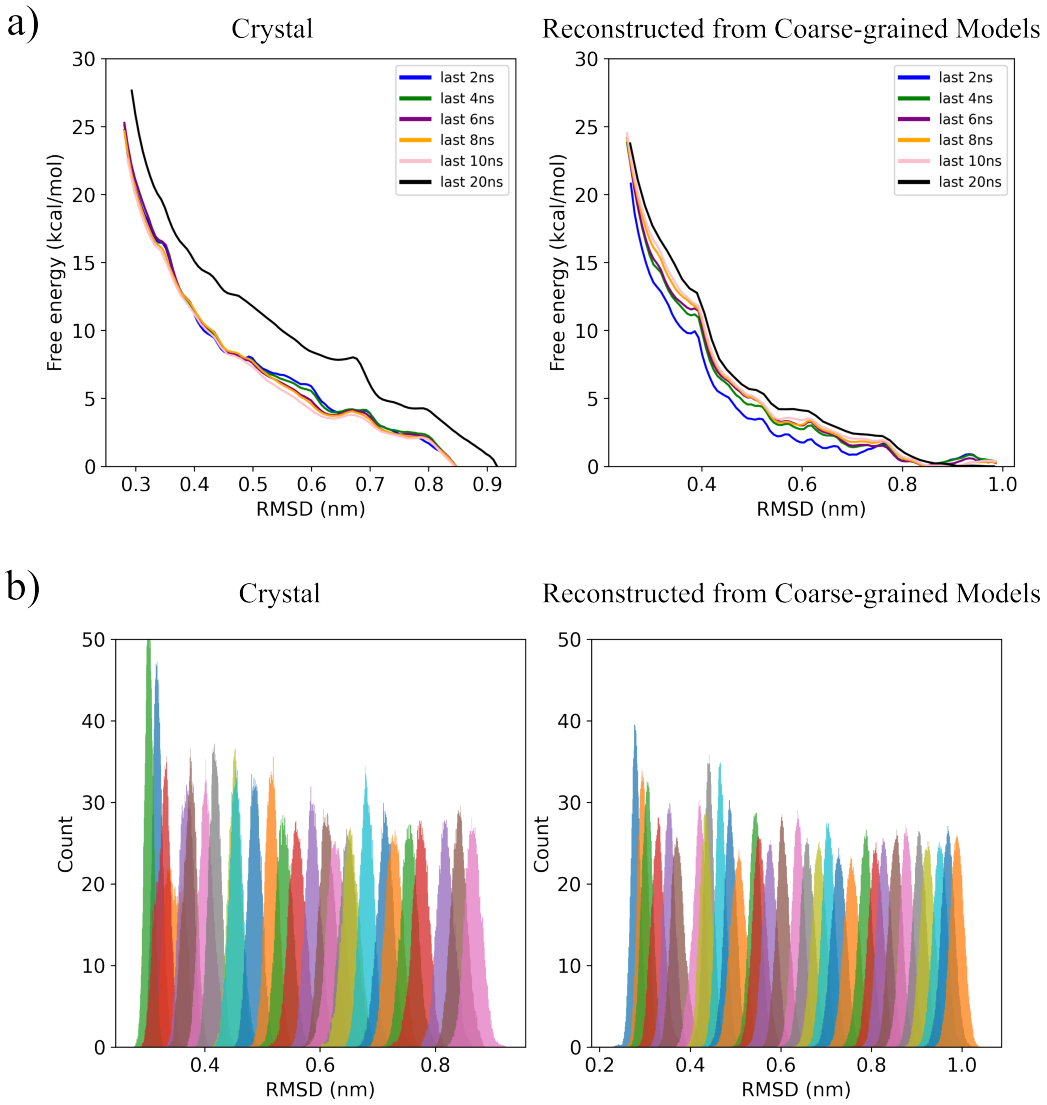


Figure A.4.: The convergence of the umbrella sampling simulations of WalK histidine kinase starting from the inactive crystal structure (left, PDB ID: 4U7N) or an all-atom model reconstructed from a coarse-grained model (right). a) The free energy profiles with respect to RMSD calculated using simulation data obtained from different time intervals. b) The normal distributions of the RMSD values. Each color histogram depicts an umbrella window.

B. Appendices of Chapter 4

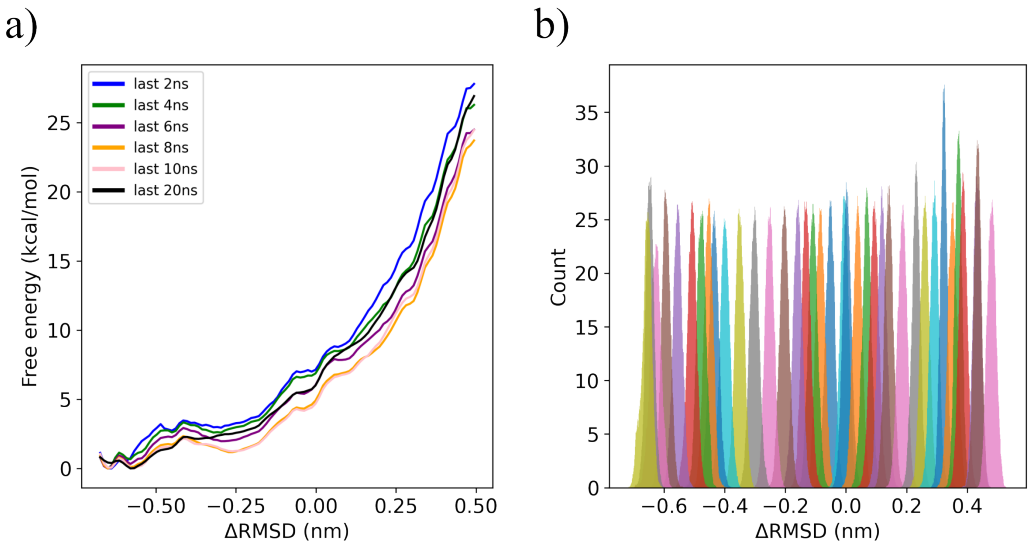


Figure B.1.: The convergence of the umbrella sampling simulations of WalK histidine kinase. a) The free energy profile with respect to $\Delta RMSD$ calculated using simulation data obtained from different time intervals. A total of 37 umbrella windows were simulated for 30 ns each. b) The normal distributions of the $\Delta RMSD$ values. Each color histogram depicts an umbrella window.

C. Appendices of Chapter 5

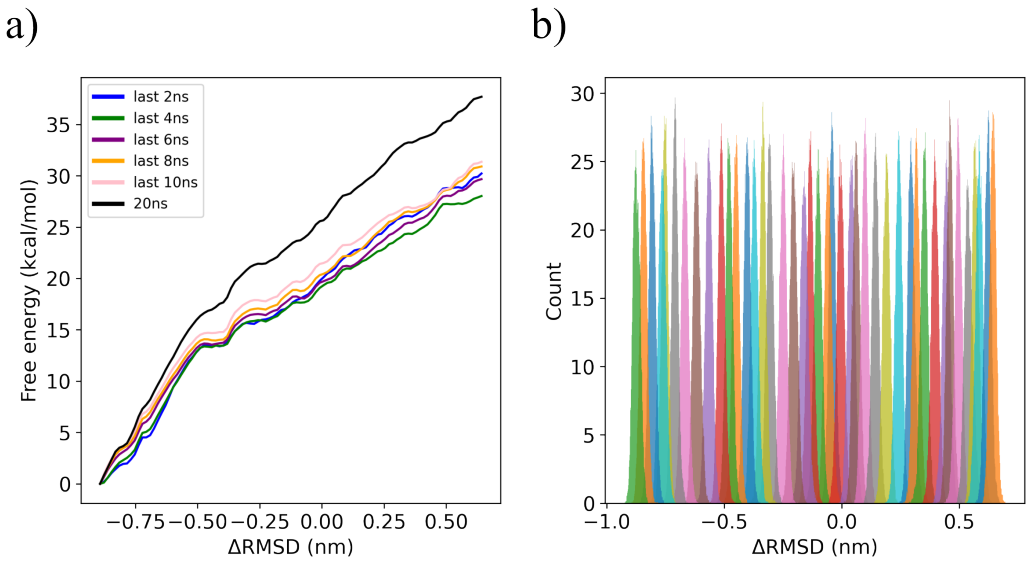


Figure C.1.: The convergence of the umbrella sampling simulations of CpxA histidine kinase. a) The free energy profile with respect to $\Delta RMSD$ calculated using simulation data obtained from different time intervals. A total of 42 umbrella windows were simulated for 20 ns each. b) The normal distributions of the $\Delta RMSD$ values. Each color histogram depicts an umbrella window.