Computational Investigation Of The Mechanism Of Molecular Recognition And Signaling Processes In Biomolecules: A Thermodynamic View

Thesis submitted for the degree of

Doctor of Philosophy (Science)

in Chemistry (Physical)

By

Krishnendu Sinha

Department of Chemistry (Physical)
University of Calcutta

2024

This thesis is dedicated to all my teachers, whose guidance, wisdom, and encouragement have shaped my academic journey. Each of them have played a pivotal role in my development and success, and their dedication to teaching and mentoring has inspired me throughout my studies. This work is a testament to your influence and support, and I am deeply grateful for everything you have imparted.

Acknowledgements

First and foremost, I would like to extend my sincere gratitude to my research guide, Dr. Suman Chakrabarty. As one of the first students in his group, I had the unique privilege of gaining direct experience from him. His dedication and insightful ideas have greatly enriched my Ph.D. journey. I would not have reached this point in my scientific career without his unwavering support, patience, and trust in me. Suman has instilled in me a profound ability to learn, teach, and think critically. I aspire to uphold the high standard of scientific excellence that he represents. Thank you, Sir.

I am most grateful to our collaborator Prof. Gautam Basu (Bose Institute), Dr. Hyunbum Jang (CCR-NCI, Bethesda), Prof. Ruth Nussinov (CCR-NCI, Bethesda), Dr. Rajib Mitra (SNBNCBS) and Dr. Salim Shah (CEO Sarfez Pharmaceuticals, Inc.) for lending me their expertise and intuition to my scientific and technical problems.

I would like to express my sincere gratitude to the members of my Doctoral advisory committee: Prof. Jaydeb Chakrabarty and Prof. Ranjit Biswas, for their insightful comments and encouragement which incented me to widen my research from various perspectives.

I gratefully acknowledge the S. N. Bose National Centre for Basic Sciences (SNBNCBS) for providing me financial support towards research fellowship and travel support for attending conferences.

I would like to extend my sincere thanks to all the members of SClab, both past and present, for their constant support and camaraderie. I am grateful to Dr. Mantu Santra, Dr. Amit Kumawat, Dr. Ipsita Basu, Dr. Saheb Dutta, Dr. Samir Ahamed, Shounak Mukherjee, Kankana Bhattacharya, Abhinandan Das, Dibyendu Maity, Sutanu Mukherjee, Sreyan Bhowmick, Sayari Bhattacharya, Shaheerah Shahid, Shivam Pandit, Muskan Sharma, and Adway Majumdar. Your collective contributions have made this journey an unforgettable experience. A special mention goes to Dr. Amit Kumawat. During his time at SNBNCBS, he was not only my mentor but also a

significant influence on my academic and personal growth. His guidance, patience, and wisdom have been invaluable, and I have learned a great deal from him. Thank you, Amit, for being a remarkable mentor. I would also like to express my heartfelt gratitude to Dr. Ipsita Basu, who is like an elder sister to me. Her unwavering support, encouragement, and ability to simplify complex tasks with ease have been a tremendous help throughout my Ph.D. journey. Her cheerful demeanor and positive attitude made even the toughest days bearable. The three of us, along with Dr. Amit Kumawat, shared many enjoyable moments, from lunch outings to spontaneous tea breaks, making the otherwise monotonous Ph.D. life enjoyable and memorable. A special note of thanks to Dibyendu Maity, my junior and partner in crime. Dibyendu's innovative ideas and willingness to assist in various aspects of my research have been indispensable. Dibyendu has always been there to lend a helping hand and offer creative solutions to problems. I consider myself fortunate to have had such a dedicated and resourceful junior.

A special thank you to the 2nd floor bay members: Premashis Kumar, my badminton partner, rival in chess and table tennis, and a source of inspiration with his never-give-up mentality and dedication; Tonmoy Chakrabarty, a great human being and my arch rival in badminton (matches between Premashis Kumar and me vs. Tonmoy Chakrabarty and Sayan Ghosh are unforgettable); and Krishnendu Patra, whose humor always brought laughter to us all. Me, Premashis Da, Tonmoy Da, Krishnendu Patra (Kaka), Dibyendu, Sayari, and Sutanu often found ourselves ordering food late at night, turning those moments into impromptu parties. Working late into the night with them, sharing stories, and creating countless memories are experiences that I will always cherish. I would also like to thank Sudipto Mitra, Prasun Boyal, Shobhan Deb Mandol, Anish Das, Jayarshi Bhattacharya, Sayan Ghosh, and Anuthosh. I am grateful for the memorable interactions and experiences shared with each of you.

I would also like to thank Ardhendu for his invaluable assistance with all official matters, always providing help whenever needed. Additionally, I am grateful to Ria for our engaging discussions, both scientific and non-scientific, which have been both captivating and enjoyable.

I would like to express my heartfelt gratitude to Dipanjan, my college friend, whose hard work and dedication taught me that perseverance is the key to success. He is the person who advised me to join Dr. Suman Chakrabarty's lab (SClab), and it turned out to be the greatest advice. His motivation and support have been invaluable, and I am forever grateful for his guidance. I also want to acknowledge Subhajit Singha my roommate during MSc, with whom I embarked on the Ph.D. journey. Subhajit is a great person with an exact opposite point of view towards life compared to mine, but those differences only strengthened our friendship, making it more profound and resilient. His unique perspectives have enriched my own understanding, and our bond has been a source of immense support and joy throughout this journey.

A special thanks to Rik Niranjan Mukherjee. Our discussions on science, philosophy, and mythology were always enriching and stimulating. These memories will always hold a special place in my heart, and I cherish the intellectual and personal growth that came from our interactions.

I would like to extend my heartfelt gratitude to my family, whose steadfast support and boundless encouragement have been the bedrock of my journey. To my incredible wife, Arpita Mukherjee, your constant understanding and inspiring presence have been my greatest strength. To my devoted parents, Mita Sinha and Tarun Kumar Sinha, your sacrifices, love, and unwavering belief in me have paved the way for my achievements. I am profoundly grateful to my elder sister, Ipsita Sinha, and her supportive husband, Koushik Pratihar, for their ceaseless encouragement and faith in my abilities. Your support has been a guiding light through every challenge.

I am also deeply appreciative of my extended family: my mother-in-law, Malabika Mukherjee, my brother-in-law, Avik Mukherjee, and his wife, Rumita Mukherjee.

A sincere thanks to my childhood friend, Sourav Biswas, for always having my back and being a constant source of support. Finally, a special thanks to Wow Momo for keeping the outlet open at midnight and to Utpal Da for the delicious Maggi. Both have played an important role in keeping me going throughout my Ph.D. tenure.

Abstract

Signal transduction, an essential cellular process, involves a series of organized biochemical and biophysical reactions triggered by various stimuli, leading to the effective propagation of signals through intramolecular dynamics and inter-molecular interactions. This thesis investigates the role of allosteric modulation and molecular interactions in several representative examples of signal transduction processes, employing molecular dynamics (MD) simulations to unravel the intricate details of these biological phenomena.

We have studied the role of metastable conformations in regulating biological functions, specifically focusing on the phosphorylation mechanism of the Escherichia coli toxin HipA protein. Through extensive MD simulations, the study reveals a multi-minima "in-out" pathway that explains the solvent exposure of the autophosphorylation site Ser150. This insight into the transient exposure of buried residues during phosphorylation provides a broader understanding of protein regulation.

We have explored the molecular mechanism of the "phosphorylation code," investigating how phosphorylation affects the binding affinity of Rho-specific guanine nucleotide dissociation inhibitors (RhoGDIs) to Rho family GTPases. The findings highlight the significance of electrostatic interactions in signal propagation and allosteric responses, offering a mechanistic model that underscores the role of specific electrostatic interactions in the selective release of GTPase partners.

Our research has demonstrated the impact of protonation states on phosphorylation-induced dissociation, utilizing constant pH MD simulations to demonstrate how phosphorylation alters the protonation of histidine residues. This change in protonation state influences the dynamics of the Rac1-GDI complex, facilitating its dissociation. The study emphasizes the importance of the N-terminal flexibility and changes in hydrogen bonding patterns in allosteric signal propagation and complex dissociation.

We also present an innovative approach to inhibiting protein-protein interactions (PPI) by leveraging the bidirectional nature of allostery. Using the PCSK9-LDLR

interaction, responsible for hypercholesterolemia, as a model system, the research identifies a distal loop whose conformational shifts correlate with binding affinity. This bidirectional allosteric relationship suggests potential targets for allosteric inhibitors, providing a new strategy for drug development.

We have investigated the mechanism of substrate binding to acetylcholinesterase (AChE), a key enzyme in Alzheimer's disease. The study proposes a new set of residues that better describe the opening of the AChE gorge, offering insights into how substrates and ligands affect protein conformations. This understanding is crucial for designing effective AChE inhibitors, which are essential for therapeutic applications.

Overall, our research advances the understanding of the regulatory mechanisms governing protein function, with significant implications for the rational design of drugs targeting allosteric sites and protein-protein interactions. Through a combination of computational methodologies, this work provides valuable insights into the dynamic conformational changes and regulatory mechanisms that underpin various biological processes.

Publications

Papers to be included in this thesis

1. "Phosphorylation-Competent Metastable State of Escherichia coli Toxin HipA", *Bhawna Pandey, ***Krishnendu Sinha**, Aditya Dev Himal K. Ganguly Smarajit Polley Suman Chakrabarty and Gautam Basu, Biochemistry, 62(5), 989-999 (2023), https://pubs.acs.org/doi/10.1021/acs.biochem.2c00614.

2. "Molecular mechanism of regulation of RhoA GTPase by phosphorylation of RhoGDI", **Krishnendu Sinha**, Amit Kumawat, Hyunbum Jang, Ruth Nussinov, Suman Chakrabarty Biophysical Journal , 123(1), 57-67, (2023), https://doi.org/10.1016/j.bpj.2023.11.018

3. "Leveraging Bidirectional Nature of Allostery To Inhibit Protein–Protein Interactions (PPIs): A Case Study of PCSK9–LDLR Interaction", **Krishnendu Sinha**, Ipsita Basu, Zacharia Shah, Salim Shah, Suman Chakrabarty, Journal of Chemical Information and Modeling, 64(9), 3923-3932, (2024), https://pubs.acs.org/doi/10.1021/acs.jcim.4c00294

4. "Molecular Insights into the Regulation of Rac1 by Phosphorylation of GDI.", **Krishnendu Sinha**,Amit Kumawat, Hyunbum Jang, Ruth Nussinov, Suman Chakrabarty (Manuscript under preparation)

5. "Investigation of binding mechanism of Acetylcholine to Acetylcholinesterase.", **Krishnendu Sinha**, Suman Chakrabarty (Manuscript under preparation)

* These authors have contributed equally.

Contents

List of Tables

List of Figures

1

Introduction

"It is strange that only extraordinary men make the discoveries, which later appear so easy and simple"

Georg C. Lichtenberg

1.1 Signaling Processes in Biological Systems

Signal transduction is an intricate series of biochemical processes that take place within cells in reaction to inputs from either the external environment or inside the cell itself[1]. The primary messengers, often referred to as stimuli, may be either chemical (such as metal ions, ligands, hormones, neurotransmitters, chemokines, paracrine and autocrine substances) or physical (such as pressure, light, heat)[2–6]. The stimuli trigger signaling pathways that initiate a series of closely regulated reactions, which in turn affect important physiological activities such as gene expression, enzyme activity, and cell division.

Proteins are the fundamental components that control the cellular signaling. Receptor proteins, often transmembrane proteins, facilitate the cellular recognition of signals by binding to signaling molecules outside the cell and transmitting the signal inside. These receptors exhibit specificity towards certain chemicals, such as dopamine, insulin, or nerve growth factors. The three primary categories receptors may be classified into are G-protein-coupled receptors, ion channel receptors, and enzyme-linked receptors. These classifications pertain to the processes via which they convert external signals into internal ones: protein activity, ion channel opening, or enzyme activation. Certain receptors are situated inside the cell or nucleus and have the ability to bind to substances such as gasses and steroid hormones that have the capability to traverse the plasma membrane.

When receptor proteins receive a signal, they undergo conformational changes that

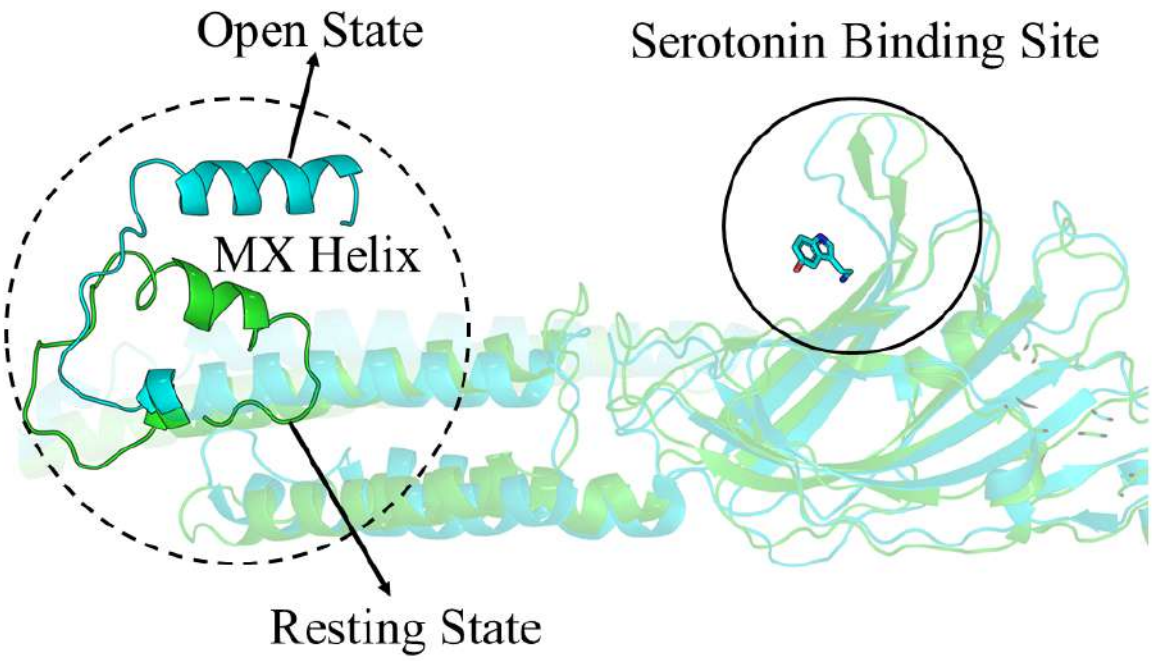

Figure 1.1: Structural Comparison of the MX Helix in 5-HT3 Receptor States. The figure illustrates the different positions of the MX helix in the resting state (green, PDB ID: 6BE1) and the open state (cyan, PDB ID: 6DG8) of the 5-HT3 receptor. The serotonin binding site is highlighted, showing the conformational changes triggered by serotonin binding.

trigger a sequence of biochemical processes within the cell. For example, 5-HT3 recep-

tors are members of the family of pentameric ligand-gated ion channels. Each subunit has an extracellular, transmembrane, and intracellular domain[7]. The MX helix in 5-HT3 receptors occupies distinct positions in the resting (green, PDB ID: 6BE1[8]) and open (cyan, PDB ID: 6DG8[9]) states. The conformational state of the MX helix is influenced by the presence or absence of the substrate, serotonin. This binding event induces a conformational change in the MX helix, which in turn initiates downstream signaling processes. The structural comparison is shown in Figure 1.1.

The ligand binding to receptor proteins give rise to intracellular signaling pathways or signal transduction cascades. These pathways enhance the signal, generating numerous intracellular signals for each receptor that is bound. Receptor activation may result in the production of second messengers such as cAMP, cGMP, Ca^{2+} ions, IP3. These second messengers and protein directed post translational modification like phosphorylation or ubiquitination regulate intracellular signaling[10].

1.2 Free Energy Landscapes: Understanding Protein Signaling

The concept of the free energy landscape has profoundly transformed our understanding of biomolecular functions. Rather than viewing proteins as static entities, this concept emphasizes that proteins exist as dynamic ensembles of conformations. This perspective is particularly useful for understanding protein signaling processes, which are often considered in terms of free energy surfaces.

A free energy landscape is a multidimensional surface that maps the free energy of a protein as a function of its conformational coordinates. The deepest minima on this surface correspond to the protein's most stable state, known as the "native conformation." Local minima, on the other hand, represent metastable states that play critical roles in protein function.

The dynamic nature of proteins allows them to respond to environmental changes and interact with various partners, thus facilitating complex signaling processes. For

instance, GTP-binding proteins (G-proteins) act as molecular switches, toggling between an active GTP-bound state and an inactive GDP-bound state. Similarly, post-translational modifications, such as phosphorylation, can stabilize specific protein conformations, thereby modulating the protein's activity.

Determining an accurate free energy surface for a protein involves significant chal-

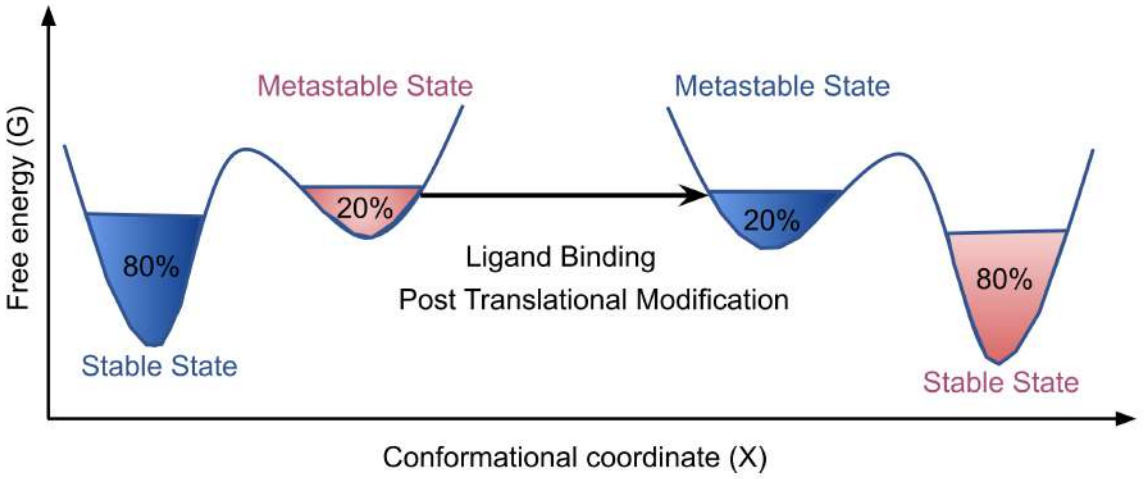

Figure 1.2: This diagram illustrates the free energy surface of a protein and how ligand binding or post-translational modifications can alter this landscape. On the left, the protein in its free form predominantly occupies a stable state (80%) with a smaller fraction in a metastable state (20%). Upon ligand binding or post-translational modifications, shown in the right part of the diagram, the distribution shifts, stabilizing a new conformational state. This modification results in the previous metastable state becoming the new stable state (80%), while the former stable state now becomes metastable (20%).

lenges, primarily in identifying a proper conformational coordinate and adequately sampling the conformational space. The choice of conformational coordinate (X) is crucial, as it must capture the relevant structural changes associated with the protein's function. Inappropriate coordinates can lead to an incomplete or misleading representation of the free energy landscape. Additionally, proper sampling of the conformational space is essential to obtain a reliable free energy surface. Inadequate sampling results in incorrect probability distributions, P(X), leading to erroneous calculations of the free energy, G(X) = -RT ln P(X). Ensuring comprehensive sampling

is difficult due to the vastness of the conformational space and the presence of high-energy barriers between states. Methods for identifying appropriate conformational coordinates and achieving proper sampling of the conformational space are detailed in the methodology section.
These perturbations can alter the free energy landscape in two key ways: by preferentially stabilizing a particular conformation or by altering the barrier heights between conformational states. Stabilizing a specific conformation can shift the equilibrium towards that state(Figure1.2), while changes in barrier heights influence the rates at which the protein transitions between different states. Together, these effects impact the protein's function and the regulation of signaling pathways. Protein signaling involves two key aspects:

- Intra-protein interactions (especially allostery)
- Inter-molecular interactions (protein-protein interaction and molecular recognition

In the subsequent sections we have discussed these in detail.

1.3 Intra-Protein Communication: Allostery

Allosteric regulation of proteins refers to the modulation of protein activity by a perturbation or stimulus at one site, known as the allosteric site, which then alters the function or characteristics of another site located further away, called the orthosteric site[11]. Allosteric perturbation might arise from several factors, including ligand binding, post-transcriptional modifications, and changes in pH and temperature[12,13]. The field of allostery is becoming more and more popular. It is unsurprising as allostery plays a crucial role in regulating cellular functions and serves as a mechanism by which the environment influences function[14].

Different Models of Allostery

The historical evolution of this allostery began in 1904 when Christian Bohr made a discovery known as the "Bohr effect." He discovered that the affinity of hemoglobin for binding oxygen is influenced by the presence of carbon dioxide and also by the pH of the enviroment[15]. This discovery emphasized the allosteric property of hemoglobin, in which the binding of one molecule affects the binding affinity of another molecule at a separate location. Bohr's finding was crucial, establishing the basis for comprehending how proteins may alter their shape and function in reaction to various environmental conditions. The Hill equation[16] provides a mathematical framework to describe the cooperative binding of ligands to a multi-subunit protein, such as hemoglobin. The Hill equation is given by:

$$\theta = \frac{[\mathrm{L}]^{\mathrm{n}}}{\mathrm{K_d} + [\mathrm{L}]^{\mathrm{n}}} \tag{1.1}$$

where θ is the fraction of occupied binding sites, [L] is the concentration of the ligand, $\mathrm{K_d}$ is the dissociation constant, and n is the Hill coefficient, indicating the degree of cooperativity.

The Hill coefficient, n, typically greater than 1 for hemoglobin, reflects positive cooperativity, meaning that the binding of one oxygen molecule increases the affinity for subsequent oxygen molecules.

In the presence of higher concentrations of $\mathrm{H^+}$ ions and $\mathrm{CO_2}$, the affinity of hemoglobin for oxygen decreases, which can be modeled by an increase in the effective $\mathrm{K_d}$ value in the Hill equation. This shift explains how hemoglobin more readily releases oxygen under conditions where tissues are more acidic and have higher levels of carbon dioxide, thus facilitating oxygen delivery to metabolically active tissues. After this several models and equations have been developed to describe allosteric behavior, such as the Adair equation[17], and refinements by Irving Klotz[18] and Linus Pauling[19]. These models provide mathematical frameworks to understand cooperative binding and the stepwise binding process, reflecting the intricacies of allosteric regulation. Despite being recognized long before the term "allosteric" was coined, it wasn't until 1961 that Jacques Monod and Francois Jacob introduced the term "allosteric inhibition" [20].

The MWC model, also known as the conformational selection model, was proposed

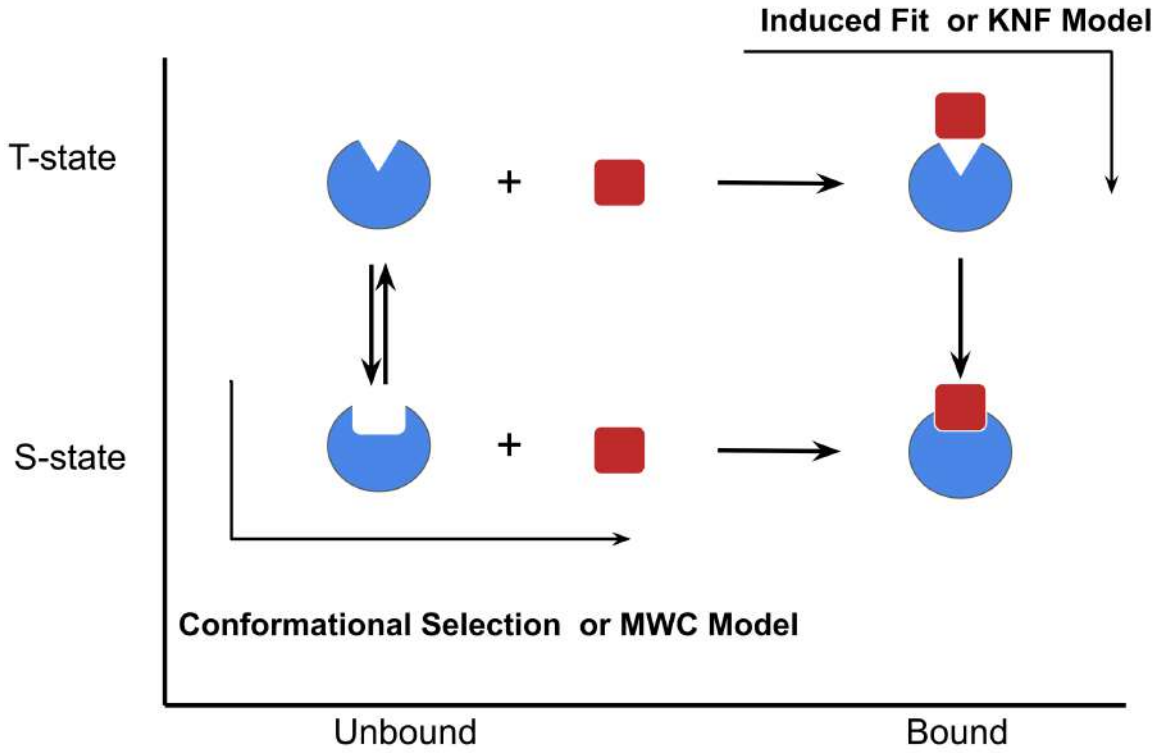

Figure 1.3: Schematic Representation of Conformational Selection (MWC Model) and Induced Fit (KNF Model) Mechanisms.

by Jacques Monod, Jeffries Wyman, and Jean-Pierre Changeux in 1965[21]. According to the MWC or conformation selection model, an allosteric protein exists in equilibrium between a triangular (T) conformation and a square (S) conformation and can shift between these two states upon ligand binding. The binding of a ligand to the protein shifts the equilibrium towards the high-affinity S state, making the transition more likely. The MWC model has been used to describe allostery in proteins such as GroEL and CheY.

In contrast, the KNF model, also known as the induced fit model, was proposed by Daniel Koshland and colleagues in 1966[22]. According to this model ligand binding, the protein undergoes a conformational change that reshapes the active site to better fit the ligand. This conformational adjustment increases the binding affinity between

the protein and the ligand. The binding event induces structural rearrangements in the protein, optimizing interactions such as hydrogen bonds, hydrophobic contacts, and van der Waals forces. The MWC and KNF models are not mutually exclusive; the MWC model can be viewed as the "all-or-none" limiting case of the sequential model. Both models are instrumental in understanding allosteric regulation in proteins and enzymes[23]. The Schematic representation of MWC and KNF model is shown in Figure 1.3.

For over next twenty years conformational change has been regarded as a distinctive feature in the idea of allostery. Then in 1984 Cooper et al.[24] proposed the term "dynamic allostery" challenging the notion that conformational change was the exclusive hallmark of allostery. These scientists have conducted a remarkable study where they establish a connection between thermal fluctuations and dynamic behavior using statistical mechanics. They highlighted the significance of conformational entropy in allostery. Evidently, alterations in the dynamics associated with the conformational entropy that provide valuable insights into the free energy of protein-ligand binding[25–28]. This concept demonstrates that allostery is a process driven by thermodynamics, influenced by both changes in enthalpy and entropy.

In 1999, the Nussinov group introduced the "conformation selection and population shift" model[29–31], which broadened the understanding of allostery by including the idea of multiple states instead of only two-state transitions. This model highlights the collective aspect of allostery, in which proteins exist in a dynamic balance of several structural states, and the binding of a ligand alters the distribution of these states. Simultaneously, the Ranganathan group emphasized the energetic interconnection among residues, revealing evolutionarily conserved residues that constitute allosteric networks[32,33]. These observations provide a more detailed perspective on the transmission of allosteric signals inside proteins. In 2004, the Nussinov group proposed that all proteins have allosteric characteristics, expanding the range of allostery in biological study[25]. In 2015, the Nussinov group reexamined the hypothesis of "allostery without a conformational change" and stressed that the lack of visible structural changes does not disprove the existence of conformational shifts. They at-

tributed this to limitations in the technologies used for detection[34].

The recent research conducted by Chakrabarty's group has significantly enhanced the understanding of dynamic allostery, namely in PDZ domain proteins, which are often seen in signaling complexes[35,36]. In the past, allostery in PDZ domain was believed to be caused by variations in entropy without any notable alterations in structure[37]. Nevertheless, their investigation demonstrates that the internal redistribution and rewiring of electrostatic contacts, particularly paired hydrogen-bonded interactions, play a vital role in the allosteric response[35]. The researchers showed that extrinsic perturbations, such as protonation, may greatly influence the conformational dynamics of distal locations, while without changing the overall structure. They also showed that the allosteric network for the PDZ2 domain protein consists of a group of residues that react to various external perturbations by reorganizing hydrogen-bond networks and redistributing electrostatic interaction energies[36].

Methods for Detection of Allostery

Allostery has been extensively examined over a span of fifty years via a combination of experimental and computational approaches.

Experimental Methods

X-ray crystallography is often used to get precise structural data both before and after perturbation, hence uncovering conformational differences[38–42]. Nevertheless, X-ray crystallography is limited in its ability to capture the dynamic character of allostery since it only provides static snapshots and may be influenced by biases from crystal packing. Time-resolved crystallography conducted using synchrotron sources significantly increases our comprehension by uncovering structural changes that occur over time[43]. On the other hand, nuclear magnetic resonance (NMR) spectroscopy provides a more dynamic perspective by recording temporary, less often occurring conformations. Nuclear magnetic resonance (NMR) techniques may effectively monitor changes in protein structure by using additional markers[44–50]. Furthermore, fluores-

cence resonance energy transfer (FRET)[51] and hydrogen-deuterium exchange mass spectrometry (HDXMS) are well-recognized techniques used to investigate allostery. For example, Chandramohan et al. used HDXMS to identify deuterium substitution for hydrogen atoms on the protein, which allows for the observation of conformational changes[52]. Double-mutant cycle analysis (DMC) is an experimental technique used to determine the energetic interactions between residues, hence improving our comprehension of allosteric networks[53].

Computational Method

In addition to these experimental procedures, computational methods provide effective tools for investigating allostery. Residues that participate in functional and correlated movements have been effectively found by topological studies, such as graph theory[54], statistical coupling analysis[32], and perturbation algorithms[55]. Methods such as the CONTACT method[56] and residue frustration analysis[57,58], which includes the "protein frustratometer" online service[59], have demonstrated significant concurrence with experimental results. This underscores the significance of steric interactions between different amino acids in the transmission of allosteric signals.

Molecular dynamics (MD) simulations produce comprehensive sets of molecular conformations, which may be used to analyze allosteric networks using techniques such as normal mode analysis[60–62], correlation matrices[63–65], community-network analysis[66], mutual information[67], and dynamical network analysis[68–70].Study by Gerhard Stock's group used non-equilibrium molecular dynamics simulations to model the allosteric communication pathway[71]. It has been demonstrated that the nonbonded interaction energy can be used to find the allosteric pathway[35,36,72,73]. Coarse-grained simulations have been demonstrated to be a highly effective tool for studying the allosteric behavior of larger biological systems[74]. These MD-based techniques successfully reveal the dynamic nature of allostery, pinpointing critical residues and pathways in proteins, and facilitating the study of allosteric mechanisms across various biological systems.

The Role of Allosteric Mutations in Disease and the Success of Targeted Allosteric Drugs

Allosteric mutations can cause a range of disorders by interfering with the regular functioning of proteins and signaling networks. The majority of mutations that cause diseases in proteins are located in regions of the protein other than the binding sites[75]. These mutations can disrupt signaling through allosteric pathways, resulting in disease either by changing the balance between active and inactive states[76] or by modifying the shape and dynamics of the binding site, or by eliminating or introducing sites for allosteric posttranslational modifications[77]. Furthermore, these mutations have the ability to modify patterns of posttranslational modifications[78].The L834R (or L858R) mutation in EGFR, which occurs in 41% of lung cancer patients, stabilizes the active αC-helix-in conformation through heterodimerization, even in the absence of ligand-promoted dimerization[79]. Conversely, mutations like T790M in EGFR, T315I in BCR-ABL, T334I in c-ABL, T341I in Src, T670I in KIT, and T674I in PDGFRA stabilize the hydrophobic regulatory spine, thereby promoting an active kinase conformation[78]. These examples illustrate the diverse mechanisms through which allosteric mutations activate kinases.

Allosteric drugs, which bind away from the active site of a target protein, offer significant advantages, including high specificity and the ability to modulate rather than completely inhibit protein activity[80]. These drugs can work alongside endogenous ligands, allowing for fine-tuned regulation and reduced side effects. Examples include benzodiazepines targeting the GABA receptor of GPCR[81] and cinacalcet modulating the Ca^{2+}-sensing receptor[82]. Despite their benefits, discovering and developing allosteric drugs is challenging due to unknown binding sites and unpredictable modulatory effects. Additionally, allosteric drugs face issues like drug-resistant mutations and variable patient responses due to genetic and metabolic differences[83]. Toxicity and dosage requirements further complicate their development. Nonetheless, the rapid expansion of allosteric modulators highlights their potential, particularly in overcoming resistance to traditional orthosteric drugs, as demonstrated by advancements in

treatments for conditions like cancer and HIV-1[80].
Although there are difficulties, researchers are actively studying allosteric inhibitors to address the problem of resistance to competitive inhibitors. However, it is important to note that allosteric inhibitors are not always more effective, as shown by the case of rapamycin and mTOR regulation[84].

1.4 Inter-Molecular Communication: Molecular Recogniton

Molecular recognition is a fundamental principle in the fields of chemistry and biology, including the precise interactions between two or more molecules via non-covalent bonding. The subject of study has seen substantial development since the first investigations conducted by scholars such as Lehn and Cram, who examined the interaction between crown ethers and cations in organic solvents[85–87]. Understanding molecular recognition in biological systems such as protein-ligand, protein-protein, protein-nuclic acid interactions which are essential for comprehending complex biological phenomena. Molecular recognition depends on many non-covalent interactions, including as hydrogen bonding, dipole-dipole interactions, induced dipole effects, cation-π interactions, lone pair-π interactions, and van der Waals forces. The hydrophobic effects are also important in driving the recognition processes between protein-ligand and protein-protein. Although it is significant, the hydrophobic effect continues to be one of the least understood phenomena in the field of molecular research. Each of these interactions (forces) contributes to the overall free energy of the interaction, making the prediction and mechanistic understanding of molecular recognition a complex but essential challenge in chemistry and biology. In this work we are mainly interested in protein-protein interactions.

1.4.1 Protein-Protein Interaction

Proteins play a crucial role in carrying out many activities such as signaling and enzyme activity. Due to their involvement in all these cellular activities, there have been substantial endeavors to predict their functions based on sequences and structures[88–90]. One effective approach for predicting protein function is to identify their binding partners, since protein-protein interactions play a crucial role in most biological functions. Comprehending these interactions aids in the delineation of cellular pathways and their intricate interrelationships[91–93]. The dynamic regulation of proteins may be deduced from their associations, since it is not possible for two protein partners to bind concurrently at the same location. Identifying these interactions is essential for functional genomics and drug development. An understanding of the pathways by which interactions occur, including their structure, duration, and how they change over time, might provide useful information for anticipating the potential adverse effects of drugs. Forecasting protein associations is a difficult task because of the wide range of interactions. Under particular circumstances, any two proteins can interact with variable intensities, mostly due to the hydrophobic effect[94,95], as well as hydrogen bonds[96] and electrostatic interactions[97,98]. The Gibbs free energy of complex formation, as determined by the association (K_a) and dissociation (K_d) constants, quantifies the stability of the interaction, with free energy changes (ΔG_a) varying from -6 to -19 kcal/mol. For predicting protein-protein interactions it is important to keep in mind the chemical aspects protein associations such as shape and charge complementarity[99,100]. Proteins exhibit flexibility and exist as collections of interchanging conformations, with both their backbones and side-chains in constant motion. Native proteins often exist in several energy minima that are separated by low barriers, indicating the presence of different conformational states and allosteric influences[25]. However, it is important to note that seeing proteins as rigid structures might be deceptive, particularly when it comes to hub proteins or those involved in regulation, since various conformations may selectively interact with different partners. A considerable proportion of cellular proteins exhibit a state of “natively dis-

order," indicating their instability in solution[101,102] while yet retaining some residual structure[103–105]. These disordered regions often make up distinct parts or domains of the protein and have functions in regulating the cell cycle, transmitting signals, and forming large molecular complexes. Although they are disordered individually, their natural conformation becomes stable when they bind[106]. Disordered proteins maintain their overall structure when they attach to various partners, but there may be variations in their local conformation, which makes it difficult to anticipate protein interactions.

Different Types of Protein-Protein Interactions (PPIs)

Protein-protein interactions (PPIs) may be classified according to their binding affinity[107], which varies from brief interactions with weak binding affinity to very strong, persistent interactions. Transient contacts, such as electron transfer complexes and mono-ubiquitin-binding interactions, exhibit brief durations and possess affinities within the high micromolar to millimolar range[108,109]. Conversely, very close contacts, such as RNAse-inhibitor[110] and protease-inhibitor complexes[111], have half-lives that span many days and affinities that are measured in sub-picomolar units. Although they possess considerable strength, these interactions are classified as transitory because the proteins operate in both unbound and complexed states. Obligatory interactions, which make up the bulk of protein-protein interactions (PPIs), include oligomeric proteins that form permanent complexes, serving as the only stable structure inside the cell. Within E. coli, almost 75% of proteins exist as oligomers[112], with more than 50% forming either homo or hetero oligomers[113]. These complexes exhibit greater stability against denaturation and have a reduced surface area exposed to the solvent. Symmetric homo-oligomers are common, emphasizing the importance of symmetry in evolution, since a single mutation may affect several subunits[114]. There are notable differences in the composition of oligomeric protein interfaces between obligatory and transitory interactions. Contacts between side chains of proteins are the primary kind of contacts for protein-protein interactions, but these sidechain-

sidechain contacts are less important for intra-protein interaction[115]. These disparities enable precise forecasting of the interaction category to which a binding site belongs[116]. Recently, there have been published comprehensive reviews on multiple PPI databases, and various tools.[107,117,118]

Regulation of Protein Association and Dissociation Rates

Protein binding affinity is determined by intricate interactions that are controlled by individual residues, and it has important consequences for both the rates of association and dissociation. At first, proteins come together to form an "encounter complex", which is mostly held together by long-distance electrostatic interactions. The presence of charged residues located on the edge or immediately next to the interface greatly increases the pace at which binding occurs. This is mainly due to the electrostatic complementarity, which plays a major role in promoting quicker association[119,120]. The association rate is influenced by structural rearrangement and short-range interactions[121]. For dissociation breaking of these short range interacts are essential[122]. The primary driving force behind faster association is electrostatic interactions, whereas slower association is characterized by substantial structural rearrangements[122]. The design and prediction of these faster interactions are facilitated by electrostatic complementarity, where charged residues located on or near the interface have a crucial impact[120,123,124]. Nevertheless, it is more challenging to anticipate the residues that dictate the dissociation rate, as dissociation entails the disruption of connections occurring at close proximity. Mutations, particularly hot-spot mutations, may significantly impact dissociation rates, leading to a decrease in binding affinity of up to 4 orders of magnitude[122]. Advancements in computational design hold promise for creating protein complexes with slower dissociation rates and improved affinity, despite the challenges[125,126].

Features of Protein-Protein Interactions (PPIs)

PPI interfaces have different properties in comparison to the entire protein surface and core. Research suggests that PPI interfaces have distinct amino acid compositions, hydrophobicity, polarity, and charge. It is important to note that not all PPIs (protein-protein interactions) exhibit the same characteristics. For instance, the interactions involving RNAse inhibitors are distinct from those involving electron transfer complexes or antibody-antigen contacts, particularly in terms of their interface features. Some of the important features are:

- Permanent protein-protein interactions (PPIs) tend to cover greater surface areas and possess higher hydrophobicity compared to transient PPIs[127].
- Non-interacting surfaces are less hydrophobic than both transient and permanent PPIs[128].
- Interface residues often exhibit higher levels of evolutionary conservation, with this phenomenon being more prominent in necessary contacts compared to transitory ones[129].
- The majority of the binding energy is focused on certain residues known as "hot-spots" [130].
- Tryptophan, arginine, and tyrosine are more frequently found in PPI hot-spots compared to other amino acids[131,132].
- When hot-spot residues are mutated to alanine, there is a substantial decrease in binding energy[133].

Despite the diversity observed in protein complexes, these fundamental principles are crucial for understanding the composition of protein-protein interactions (PPIs).

Methods for PPI Detection

PPI detection utilizes a range of methodologies, which may be classified into in vitro, in vivo, and in silico approaches.

In Vitro Methods

In vitro experiments allow for precise control over experimental settings, using techniques like Tandem Affinity Purification-Mass Spectroscopy (TAP-MS)[134,134] and affinity chromatography[135]. Co-Immunoprecipitation(co-IP) validates interactions in their original states using intact cellular extracts[136]. An exciting new tool for discovering PPIs and other targeted biochemical reactions is the protein array[137–139], which allows for the simultaneous detection of several accessible components in a single experiment. In proof-of-principle tests, Zhu et al. printed 5,800 pure yeast proteins onto coated glass slides and found new proteins that interact with calmodulin and phospholipids[140]. Protein-fragment complementation assays (PCAs) provide a simple and direct method for detecting protein-protein interactions in living cells, multicellular organisms, or in vitro, applicable to proteins of any molecular weight and expressed at endogenous levels[141]. Phage display involves incorporating proteins into phage particles to find potential partners[142]. X-ray crystallography observes protein structures at the atomic level[143], while Nuclear Magnetic Resonance (NMR) spectroscopy discovers PPIs by monitoring magnetically active nuclei in a powerful magnetic field[144].

In Vivo Methods

In contrast, in vivo methods are performed within living organisms, offering a more accurate representation of natural cellular environments. The Yeast Two-Hybrid (Y2H) system is a widely used in vivo method that screens for interactions by using two protein domains to activate a reporter gene[145]. Synthetic lethality identifies functional interactions by creating gene mutations that are viable alone but lethal when combined, revealing essential interaction networks[146,147].

In Silico Methods

Computational approaches, known as in silico methods, utilize biological data to predict protein-protein interactions (PPIs). These methods are extremely useful for managing vast amounts of data and providing ideas that may be experimentally validated. Sequence-based methods utilize orthologous sequences or conserved domains to forecast interactions[148], whereas structure-based methods deduce interactions based on common structural characteristics[149]. Functional links may be inferred when genes are positioned in close vicinity on a chromosome, which is a commonly employed strategy for analyzing bacterial genomes[150]. Gene fusion predicts relationships by combining single-domain proteins into multidomain proteins across several animals[151,152]. The In Silico Two-Hybrid (I2H) technique relies on the assumption that interacting proteins co-evolve[153]. Phylogenetic tree analysis use evolutionary histories to forecast interactions by examining co-evolution[154]. Gene expression analysis can predict interactions by detecting shared expression profiles, indicating that proteins from genes that are co-expressed are likely to interact[155].

Advancements and Challenges in Targeting Protein-Protein Interactions for Drug Discovery

The human proteome comprises around 20,000 proteins and a multitude of protein-protein interactions (PPIs) that are essential for biological functions. Dysregulation of these interactions can result in the development of different diseases[156,157]. Current approaches for the discovery of PPI modulators involve various advanced methodologies tailored to address the challenges posed by the unique interfaces of protein-protein interactions (PPIs). Traditional medicinal chemistry techniques often fall short because PPI interfaces tend to be flat and extensive, lacking the typical binding pockets found in conventional drug targets[158–160]. To overcome these hurdles, several strategies have been developed to identify potential modulators effectively. High-throughput screening (HTS) is widely utilized to discover compounds targeting PPI hot-spots[160], although it requires a diverse compound library due to the specific nature of PPI interfaces. HTS has been successful in identifying inhibitors for interactions like MDM2/p53[161–163]. Fragment-based drug discovery (FBDD) is another critical approach, leveraging the identification of small molecular fragments that can be optimized and linked to form effective modulators[164], with techniques like surface plasmon resonance and X-ray crystallography aiding in validation[165,166]. FBDD has shown success in discovering modulators for PPIs such as XIAP/caspase-9[167] and Bcl-2/Bax[168]. Current structure-based design strategies for PPI modulators include targeting hot-spots through bioisosterism and de novo design[169], and using peptidomimetic[170] design to simulate key peptides' secondary structures, particularly the α-helix. Successful modulators like VHL/HIF1α[171], c-Myc/Max[172], Bcl-2/Bax[173], and MDM2/p53[174] has been found. Virtual screening employs computational methods to screen compound libraries, focusing on both structure-based and ligand-based approaches to identify hits. This method has proven effective in developing modulators for interactions such as Ubc13/Uevl[175]. The discovery of PPI modulators also explores the mechanisms of action, distinguishing between orthosteric inhibition, where modulators bind directly at the PPI interface, and allosteric inhibition, where

binding occurs at a non-interaction site, inducing conformational changes that affect the interaction[176,177]. While most modulators are inhibitors, PPI stabilization, where modulators enhance the interaction between proteins, is emerging as a promising area[178]. The three main types of PPI modulators include small molecule modulators, antibodies, and peptides, each with its advantages and challenges[179]. Small molecules are suitable for tight, narrow interfaces but face pharmacokinetic issues due to size and solubility[169], while antibodies and peptides offer higher specificity and affinity but peptide is susceptible to hydrolysis by various hydrolases in the body. The continued advancement in structural biology and high-throughput screening techniques holds promise for the future development of effective PPI modulators.

1.4.2 Protein-Ligand Interaction and Its Implications in Drug Discovery

Protein-ligand interactions are fundamental to understanding protein functions and optimizing drug design and development. Ligands are molecules that bind to proteins with high specificity and affinity, forming complexes through noncovalent bonds[180]. This binding mechanism, crucial for molecular recognition, drives many cellular processes such as DNA replication and cellular signaling. Key to this interaction is the specificity and affinity between the protein and ligand, akin to a lock-and-key fit[181], ensuring that only perfectly fitting molecules interact[182]. The shape, size, and surface properties of both the protein and ligand must be compatible to enable effective binding. These interactions are essential for a protein to perform its function, such as enzyme catalysis or signaling. In drug discovery, this principle is harnessed to design drugs that can selectively bind to target proteins with high affinity, minimizing side effects and maximizing therapeutic efficacy[181]. The most common types of interactions include hydrogen bonds, ionic interactions, and hydrophobic interactions[183]. Hydrogen bonds involve electrostatic interactions between a hydrogen-bond donor and acceptor, providing strength through close

proximity of the atoms involved. Ionic interactions occur between oppositely charged groups, while hydrophobic interactions involve non-polar groups of the ligand and protein, crucial for the stability of the protein-ligand complex. The binding process involves initial hydrogen bonding with water molecules, followed by the displacement of these molecules as the ligand binds directly to the protein. This thermodynamic process, governed by Gibbs free energy changes, determines the stability and binding affinity of the complex. Understanding these interactions and the energetics behind them enables the design of drugs with optimal binding properties, advancing the field of drug discovery and the development of new therapeutic agents.

1.5 Systems Used for Studying Biomolecular Signaling

In our work, we have tried to understand the role of allosteric modulation and molecular interaction in signaling processes using MD simulations of different systems:

1. HipA Protein: A Bacterial Toxin
2. RhoGTPases: Siganling Proteins
3. PCSK9: A Potential Drug Target for Cardiovacular Disease
4. Acetylcholinesterase: A Known Drug Target for Alzheimer's Disease

1.6 Overview of the Thesis

The thesis is divided into seven chapters described as follows

Chapter 1: Introduction

In this chapter, we introduce two major regulating factors of signal transduction: Allostery and Molecular Recognition, with a particular emphasis on protein-protein

interactions. This sets the stage for understanding the fundamental mechanisms underlying the studies presented in the following chapters.

Chapter 2: Methods

Here, we describe the computational and experimental methods used throughout the study. This includes detailed explanations of the molecular dynamics simulations, thermodynamic analyses, and other techniques employed to investigate the biological phenomena discussed in subsequent chapters.

Chapter 3: Role of Metastable Conformations in Regulating Biological Functions

Bacterial persistence involves a subset of bacteria surviving antibiotic treatment without acquiring genetic resistance, primarily through dormancy mechanisms often driven by toxin-antitoxin (TA) systems[184]. A key player in this phenomenon is the HipBA type-II TA module, with the kinase HipA being crucial for the formation of persister cells.[185,186] HipA's activity is tightly regulated by autophosphorylation at the Ser150 residue, which is pivotal for its function. Autophosphorylation induces a conformational change in HipA, inhibiting its interaction with glutamyl-tRNA synthetase, leading to cell stasis and promoting persistence. This phosphorylation acts as a biochemical switch, maintaining HipA in an inactive state crucial for its regulatory role.

Several experimental studies have provided insights into the structural changes associated with Ser150 autophosphorylation[187]. These studies reveal that the Ser150 residue, initially buried in the inactive state of HipA, becomes exposed upon phosphorylation[187,188]. This exposure allows HipA to interact with its substrates and dynamically regulate its function, underscoring the critical role of Ser150 in HipA's regulatory mechanism.

This chapter addresses the intriguing question of how Ser150 is phosphorylated despite being buried inside the protein core. Through molecular dynamics simulations,

a multi-minima "in–out" pathway is uncovered, illustrating the gradual exposure of Ser150 with a free energy difference of approximately 2–2.5 kcal/mol. These findings provide valuable insights into phosphorylation events involving the transient exposure of buried residues, applicable to protein systems beyond HipA.

Chapter 4: Understanding the Molecular Mechanism of the "Phosphorylation Code"

Rho GTPases, part of the Ras superfamily, regulate actin cytoskeleton dynamics and various signal transduction pathways in eukaryotic cells[189]. Their dysregulation is linked to cancer, neurological disorders, and bacterial infections[190]. These GTPases cycle between active GTP-bound and inactive GDP-bound states[191], controlled by GTPase-activating proteins (GAPs), guanine nucleotide exchange factors (GEFs), and guanine nucleotide dissociation inhibitors (GDIs)[192]. GAPs facilitate GTP hydrolysis, converting Rho proteins to their inactive state, while GEFs promote nucleotide exchange, activating the GTPases. GDIs bind to GDP-bound Rho GTPases, preventing their activation and regulating their localization. Phosphorylation of RhoGDI is crucial for the selective release of RhoGTPases by altering their interactions with RhoGDIs. The "phosphorylation code" refers to the specific phosphorylation patterns on RhoGDIα, which determine its binding affinity to various RhoGTPases (Rac1, RhoA and Cdc42), thereby selectively modulating their release and activity[193]. This mechanism ensures precise control over RhoGTPase activity, allowing for accurate spatial and temporal regulation of cellular signaling pathways. In this chapter, we aim to understand the molecular mechanism behind the "phosphorylation code." The chapter is divided into two sections.

Section 1 reveals the mechanism of selective release of RhoA from the RhoA-RhoGDI complex. It highlights the role of electrostatic interactions in signal propagation from the mutation site to distant sites. The proposed mechanistic model involves allosteric responses, emphasizing the importance of specific electrostatic interactions in the manifestation of the phosphorylation code.

Section 2 elucidates the mechanism of selective release of Rac1 from the Rac1-RhoGDI complex. It demonstrates the importance of histidine protonation states in the selective release of specific GTPases. Correctly assigning the protonation state of titrable residues is crucial for understanding these systems.

Chapter 5: Leveraging the Bidirectional Nature of Allostery to Inhibit Protein-Protein Interactions (PPIs)

Proprotein Convertase Subtilisin/Kexin type 9 (PCSK9) is prominently expressed in the liver and small intestine, indicating its broad systemic implications[194]. Notably, mutations in the PCSK9 gene have been linked to autosomal dominant hypercholesterolemia, a condition characterized by elevated low-density lipoprotein cholesterol (LDL-c) levels and an increased risk of cardiovascular diseases[195]. PCSK9's critical role in regulating LDL-c levels makes it a promising target for cardiovascular disease treatment. By binding to LDL receptors (LDLR), PCSK9 facilitates their degradation, thereby reducing the number of receptors available to clear LDL-c from circulation[196,197]. Elevated LDL-c levels are a well-established risk factor for atherosclerosis, leading to coronary artery disease and other cardiovascular conditions. Traditional treatments like statins reduce cholesterol production but do not address LDLR regulation. PCSK9 inhibitors, however, increase LDLR expression on hepatocytes, enhancing LDL-c clearance and significantly lowering plasma LDL-c levels.[198–200]

PCSK9 is a 692-amino-acid protein consisting of three domains: the prodomain, catalytic domain, and C-terminal Cys/His-rich domain[201]. The prodomain (residues 31–152) is crucial for PCSK9 folding and maturation, requiring autocatalytic cleavage for the protein to be secreted. The catalytic domain (residues 153–421) is essential for PCSK9's interaction with LDLR, facilitating its degradation in lysosomes, thereby reducing the number of receptors available to clear LDL-c from the bloodstream. The C-terminal histidine-rich domain (CHRD, residues 453–692) enhances PCSK9's stability and function.

This chapter focuses on developing an alternative strategy for inhibiting protein-

protein interactions, using the PCSK9-LDLR interaction responsible for hypercholesterolemia as a model system. Extensive molecular dynamics simulations of both apo and bound systems identify a distal loop that exists in different conformations in the apo and bound states. This conformational shift correlates with binding affinity, establishing a bidirectional allosteric relationship. Thermodynamic analysis suggests that loop conformations predominant in the apo state of PCSK9 have lower LDLR binding affinity, making them potential targets for allosteric inhibitors.

Chapter 6: Mechanism of Substrate Binding to Acetylcholinesterase Protein

Alzheimer's disease (AD) is a progressive neurological disorder marked by cognitive decline, memory loss, and behavioral changes, predominantly affecting individuals aged 65 and older. A key theory underlying AD is the cholinergic hypothesis[202], which attributes cognitive impairment to a deficiency of the neurotransmitter acetylcholine (ACh) in the brain. Acetylcholine is essential for learning and memory processes. The enzyme acetylcholinesterase (AChE) hydrolyzes ACh into choline and acetate at synaptic junctions, thereby terminating cholinergic neurotransmission[203]. By inhibiting AChE activity, the availability of ACh at synapses is increased, which improves cholinergic transmission and alleviates AD symptoms[204].
Here, we study the mechanism of substrate binding to the acetylcholinesterase (AChE) protein. The findings indicate that previously known gating residues do not fully describe the gorge opening motion of the AChE protein. We propose a set of residues that better describe the opening of the gorge, enhancing our understanding of how the presence of a substrate/ligand can affect protein conformations.

Chapter 7: Future Outlook

The final chapter concludes the thesis by highlighting the significant findings from the molecular investigation of protein-protein interactions and allosteric regulation processes. It discusses the implications of these findings in the field of drug design

and outlines the challenges faced during the study, particularly regarding the selection of reaction coordinates/collective variables (CVs) and sampling the conformational space. The chapter also proposes future directions, including the application of machine learning techniques to address these challenges.

Bibliography

[1] Bradshaw, R. A & Dennis, E. A, eds. (2010) in *Handbook of Cell Signaling (Second Edition)*. (Academic Press, San Diego), Second edition edition, pp. 2983–3047.

[2] Dalous, J, Burghardt, E, Müller-Taubenberger, A, Bruckert, F, Gerisch, G, & Bretschneider, T. (2008) Reversal of cell polarity and actin-myosin cytoskeleton reorganization under mechanical and chemical stimulation. *Biophysical Journal* **94**, 1063–1074.

[3] Artemenko, Y, Lampert, T. J, & Devreotes, P. N. (2014) Moving towards a paradigm: common mechanisms of chemotactic signaling in dictyostelium and mammalian leukocytes. *Cellular and Molecular Life Sciences* **71**, 3711–3747.

[4] Artemenko, Y, Axiotakis, L, Borleis, J, Iglesias, P. A, & Devreotes, P. N. (2016) Chemical and mechanical stimuli act on common signal transduction and cytoskeletal networks. *Proceedings of the National Academy of Sciences* **113**.

[5] Stroka, K. M & Konstantopoulos, K. (2014) Physical biology in cancer. 4. physical cues guide tumor cell adhesion and migration. *American Journal of Physiology-Cell Physiology* **306**, C98–C109.

[6] Cortese, B, Palamà, I. E, D'Amone, S, & Gigli, G. (2014) Influence of electrotaxis on cell behaviour. *Integr. Biol.* **6**, 817–830.

[7] Mocatta, J, Mesoy, S. M, Dougherty, D. A, & Lummis, S. C. R. (2022) 5-ht3 receptor mx helix contributes to receptor function. *ACS Chemical Neuroscience* **13**, 2338–2345.

[8] Basak, S, Gicheru, Y, Rao, S, Sansom, M. S. P, & Chakrapani, S. (2018) Cryo-em reveals two distinct serotonin-bound conformations of full-length 5-ht3a receptor. *Nature* **563**, 270–274.

[9] Hassaine, G, Deluz, C, Grasso, L, Wyss, R, Tol, M. B, Hovius, R, Graff, A, Stahlberg, H, Tomizaki, T, Desmyter, A, Moreau, C, Li, X.-D, Poitevin, F, Vogel, H, & Nury, H. (2014) X-ray structure of the mouse serotonin 5-ht3 receptor. *Nature* **512**, 276–281.

[10] Lee, M. J & Yaffe, M. B. (2016) Protein regulation in signal transduction. *Cold Spring Harbor Perspectives in Biology* **8**, a005918.

[11] Liu, J & Nussinov, R. (2016) Allostery: An overview of its history, concepts, methods, and applications. *PLOS Computational Biology* **12**, e1004966.

[12] Özkan, E, Yu, H, & Deisenhofer, J. (2005) Mechanistic insight into the allosteric activation of a ubiquitin-conjugating enzyme by ring-type ubiquitin ligases. *Proceedings of the National Academy of Sciences* **102**, 18890–18895.

[13] Tsai, C.-J, del Sol, A, & Nussinov, R. (2009) Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Molecular BioSystems* **5**, 207.

[14] Goodey, N. M & Benkovic, S. J. (2008) Allosteric regulation and catalysis emerge via a common route. *Nature Chemical Biology* **4**, 474–482.

[15] Bohr, C, Hasselbalch, K, & Krogh, A. (1904) Ueber einen in biologischer beziehung wichtigen einfluss, den die kohlensäurespannung des blutes auf dessen sauerstoffbindung übt 1. *Skandinavisches Archiv Für Physiologie* **16**, 402–412.

[16] Hill, A. V, Hill, A. V, & Paganini-Hill, A. (year?) The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *The Journal of Physiology* **40**, 4–7.

[17] Adair, G, Bock, A, & Field, H. (1925) The hemoglobin system. *Journal of Biological Chemistry* **63**, 529–545.

[18] Klotz, I. M. (1946) The application of the law of mass action to binding by proteins; interactions with calcium. *Archives of biochemistry* **9**, 109–17.

[19] Pauling, L. (1935) The oxygen equilibrium of hemoglobin and its structural interpretation. *Proceedings of the National Academy of Sciences* **21**, 186–191.

[20] Monod, J & Jacob, F. (1961) General conclusions: Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology* **26**, 389–401.

[21] Monod, J, Wyman, J, & Changeux, J.-P. (1965) On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology* **12**, 88–118.

[22] Koshland, D. E, Némethy, G, & Filmer, D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits*. *Biochemistry* **5**, 365–385.

[23] Nelson, D. L & Cox, M. M. (2004) *Lehninger Principles of Biochemistry, Fourth Edition* ed. Freeman. Fourth edition edition.

[24] Cooper, A & Dryden, D. T. F. (1984) Allostery without conformational change: A plausible model. *European Biophysics Journal* **11**, 103–109.

BIBLIOGRAPHY

[25] Gunasekaran, K, Ma, B, & Nussinov, R. (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics* **57**, 433–443.

[26] Popovych, N, Sun, S, Ebright, R. H, & Kalodimos, C. G. (2006) Dynamically driven protein allostery. *Nature Structural & Molecular Biology* **13**, 831–838.

[27] Bu, Z & Callaway, D. J. (2011) *Proteins MOVE ! Protein dynamics and long-range allostery in cell signaling.* (Elsevier), p. 163–221.

[28] Tsai, C.-J, del Sol, A, & Nussinov, R. (2008) Allostery: Absence of a change in shape does not imply that allostery is not at play. *Journal of Molecular Biology* **378**, 1–11.

[29] Kumar, S, Ma, B, Tsai, C.-J, Wolfson, H, & Nussinov, R. (1999) Folding funnels and conformational transitions via hinge-bending motions. *Cell Biochemistry and Biophysics* **31**, 141–164.

[30] Tsai, C, Kumar, S, Ma, B, & Nussinov, R. (1999) Folding funnels, binding funnels, and protein function. *Protein Science* **8**, 1181–1190.

[31] Ma, B, Kumar, S, Tsai, C.-J, & Nussinov, R. (1999) Folding funnels and binding mechanisms. *Protein Engineering, Design and Selection* **12**, 713–720.

[32] Lockless, S. W & Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299.

[33] Süel, G. M, Lockless, S. W, Wall, M. A, & Ranganathan, R. (2002) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology* **10**, 59–69.

[34] Nussinov, R & Tsai, C.-J. (2015) Allostery without a conformational change? revisiting the paradigm. *Current Opinion in Structural Biology* **30**, 17–24.

[35] Kumawat, A & Chakrabarty, S. (2017) Hidden electrostatic basis of dynamic allostery in a pdz domain. *Proceedings of the National Academy of Sciences* **114**.

[36] Kumawat, A & Chakrabarty, S. (2020) Protonation-induced dynamic allostery in pdz domain: Evidence of perturbation-independent universal response network. *The Journal of Physical Chemistry Letters* **11**, 9026–9031.

[37] Petit, C. M, Zhang, J, Sapienza, P. J, Fuentes, E. J, & Lee, A. L. (2009) Hidden dynamic allostery in a pdz domain. *Proceedings of the National Academy of Sciences* **106**, 18249–18254.

[38] Bauman, J. D, Patel, D, Dharia, C, Fromer, M. W, Ahmed, S, Frenkel, Y, Vijayan, R. S. K, Eck, J. T, Ho, W. C, Das, K, Shatkin, A. J, & Arnold, E. (2013) Detecting allosteric sites of hiv-1 reverse transcriptase by x-ray crystallographic fragment screening. *Journal of Medicinal Chemistry* **56**, 2738–2746.

[39] Krojer, T, Fraser, J. S, & von Delft, F. (2020) Discovery of allosteric binding sites by crystallographic fragment screening. *Current Opinion in Structural Biology* **65**, 209–216.

[40] Liu, X, Ahn, S, Kahsai, A. W, Meng, K.-C, Latorraca, N. R, Pani, B, Venkatakrishnan, A. J, Masoudi, A, Weis, W. I, Dror, R. O, Chen, X, Lefkowitz, R. J, & Kobilka, B. K. (2017) Mechanism of intracellular allosteric β2ar antagonist revealed by x-ray crystal structure. *Nature* **548**, 480–484.

[41] Fraser, J. S, van den Bedem, H, Samelson, A. J, Lang, P. T, Holton, J. M, Echols, N, & Alber, T. (2011) Accessing protein conformational ensembles using room-temperature x-ray crystallography. *Proceedings of the National Academy of Sciences* **108**, 16247–16252.

[42] Günther, S, Reinke, P. Y. A, Fernández-García, Y, Lieske, J, Lane, T. J, Ginn, H. M, Koua, F. H. M, Ehrt, C, Ewert, W, Oberthuer, D, Yefanov, O, Meier, S, Lorenzen, K, Krichel, B, Kopicki, J.-D, Gelisio, L, Brehm, W, Dunkel, I, Seychell, B, Gieseler, H, Norton-Baker, B, Escudero-Pérez, B, Domaracky, M, Saouane, S, Tolstikova, A, White, T. A, Hänle, A, Groessler, M, Fleckenstein, H, Trost, F, Galchenkova, M, Gevorkov, Y, Li, C, Awel, S, Peck, A, Barthelmess, M, Schlünzen, F, Lourdu Xavier, P, Werner, N, Andaleeb, H, Ullah, N, Falke, S, Srinivasan, V, França, B. A, Schwinzer, M, Brognaro, H, Rogers, C, Melo, D, Zaitseva-Doyle, J. J, Knoska, J, Peña-Murillo, G. E, Mashhour, A. R, Hennicke, V, Fischer, P, Hakanpää, J, Meyer, J, Gribbon, P, Ellinger, B, Kuzikov, M, Wolf, M, Beccari, A. R, Bourenkov, G, von Stetten, D, Pompidor, G, Bento, I, Panneerselvam, S, Karpics, I, Schneider, T. R, Garcia-Alai, M. M, Niebling, S, Günther, C, Schmidt, C, Schubert, R, Han, H, Boger, J, Monteiro, D. C. F, Zhang, L, Sun, X, Pletzer-Zelgert, J, Wollenhaupt, J, Feiler, C. G, Weiss, M. S, Schulz, E.-C, Mehrabi, P, Karničar, K, Usenik, A, Loboda, J, Tidow, H, Chari, A, Hilgenfeld, R, Uetrecht, C, Cox, R, Zaliani, A, Beck, T, Rarey, M, Günther, S, Turk, D, Hinrichs, W, Chapman, H. N, Pearson, A. R, Betzel, C, & Meents, A. (2021) X-ray screening identifies active site and allosteric inhibitors of sars-cov-2 main protease. *Science* **372**, 642–646.

[43] Šrajer, V & Schmidt, M. (2017) Watching proteins function with time-resolved x-ray crystallography. *Journal of Physics D: Applied Physics* **50**, 373001.

[44] Boulton, S, Akimoto, M, Selvaratnam, R, Bashiri, A, & Melacini, G. (2014) A tool set to map allosteric networks through the nmr chemical shift covariance analysis. *Scientific Reports* **4**.

[45] Piazza, F. (2014) Nonlinear excitations match correlated motions unveiled by nmr in proteins: a new perspective on allosteric cross-talk. *Physical Biology* **11**, 036003.

[46] Palmer, A. G. (2015) Enzyme dynamics from nmr spectroscopy. *Accounts of Chemical Research* **48**, 457–465.

[47] Grutsch, S, Brüschweiler, S, & Tollinger, M. (2016) Nmr methods to study dynamic allostery. *PLOS Computational Biology* **12**, e1004620.

[48] Lisi, G. P & Loria, J. P. (2016) Solution nmr spectroscopy for the study of enzyme allostery. *Chemical Reviews* **116**, 6323–6369.

[49] Kadavath, H & Riek, R. (2022) *Multistate Structures and Dynamics at Atomic Resolution Using Exact Nuclear Overhauser Enhancements (eNOEs).* (The Royal Society of Chemistry), p. 29–55.

[50] Skeens, E & Lisi, G. P. (2023) Analysis of coordinated nmr chemical shifts to map allosteric regulatory networks in proteins. *Methods* **209**, 40–47.

[51] Mazal, H & Haran, G. (2019) Single-molecule fret methods to study the dynamics of proteins at work. *Current Opinion in Biomedical Engineering* **12**, 8–17.

[52] Chandramohan, A, Krishnamurthy, S, Larsson, A, Nordlund, P, Jansson, A, & Anand, G. S. (2016) Predicting allosteric effects from orthosteric binding in hsp90-ligand interactions: Implications for fragment-based drug design. *PLOS Computational Biology* **12**, e1004840.

[53] Horovitz, A. (1996) Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Folding and Design* **1**, R121–R126.

[54] Kannan, N & Vishveshwara, S. (1999) Identification of side-chain clusters in protein structures by a graph spectral method 1 1edited by j. m. thornton. *Journal of Molecular Biology* **292**, 441–464.

[55] Daily, M. D, Upadhyaya, T. J, & Gray, J. J. (2007) Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins: Structure, Function, and Bioinformatics* **71**, 455–466.

[56] van den Bedem, H, Bhabha, G, Yang, K, Wright, P. E, & Fraser, J. S. (2013) Automated identification of functional dynamic contact networks from x-ray crystallography. *Nature Methods* **10**, 896–902.

[57] Ferreiro, D. U, Hegler, J. A, Komives, E. A, & Wolynes, P. G. (2007) Localizing frustration in native proteins and protein assemblies. *Proceedings of the National Academy of Sciences* **104**, 19819–19824.

[58] Ferreiro, D. U, Hegler, J. A, Komives, E. A, & Wolynes, P. G. (2011) On the role of frustration in the energy landscapes of allosteric proteins. *Proceedings of the National Academy of Sciences* **108**, 3499–3503.

[59] Jenik, M, Parra, R. G, Radusky, L. G, Turjanski, A, Wolynes, P. G, & Ferreiro, D. U. (2012) Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Research* **40**, W348–W351.

[60] Brooks, B & Karplus, M. (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences* **80**, 6571–6575.

[61] Go, N, Noguti, T, & Nishikawa, T. (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences* **80**, 3696–3700.

[62] Levitt, M, Sander, C, & Stern, P. S. (1983) The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *International Journal of Quantum Chemistry* **24**, 181–199.

[63] García, A. E. (1992) Large-amplitude nonlinear motions in proteins. *Physical Review Letters* **68**, 2696–2699.

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

[65] Lange, O. F & Grubmüller, H. (2007) Full correlation analysis of conformational protein dynamics. *Proteins: Structure, Function, and Bioinformatics* **70**, 1294–1312.

[66] Girvan, M & Newman, M. E. J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**, 7821–7826.

[67] McClendon, C. L, Friedland, G, Mobley, D. L, Amirkhani, H, & Jacobson, M. P. (2009) Quantifying correlations between allosteric sites in thermodynamic ensembles. *Journal of Chemical Theory and Computation* **5**, 2486–2502.

[68] Sethi, A, Eargle, J, Black, A. A, & Luthey-Schulten, Z. (2009) Dynamical networks in trna:protein complexes. *Proceedings of the National Academy of Sciences* **106**, 6620–6625.

[69] Eargle, J & Luthey-Schulten, Z. (2012) Networkview: 3d display and analysis of protein·rna interaction networks. *Bioinformatics* **28**, 3000–3001.

[70] VanWart, A. T, Eargle, J, Luthey-Schulten, Z, & Amaro, R. E. (2012) Exploring residue component contributions to dynamical network models of allostery. *Journal of Chemical Theory and Computation* **8**, 2949–2961.

[71] Ali, A. A. A. I, Gulzar, A, Wolf, S, & Stock, G. (2022) Nonequilibrium modeling of the elementary step in pdz3 allosteric communication. *The Journal of Physical Chemistry Letters* **13**, 9862–9868.

[72] Sinha, K, Kumawat, A, Jang, H, Nussinov, R, & Chakrabarty, S. (2024) Molecular mechanism of regulation of rhoa gtpase by phosphorylation of rhogdi. *Biophysical Journal* **123**, 57–67.

[73] Sinha, K, Basu, I, Shah, Z, Shah, S, & Chakrabarty, S. (2024) Leveraging bidirectional nature of allostery to inhibit protein–protein interactions (ppis): A case study of pcsk9–ldlr interaction. *Journal of Chemical Information and Modeling* **64**, 3923–3932.

[74] Kravats, A. N, Tonddast-Navaei, S, & Stan, G. (2016) Coarse-grained simulations of topology-dependent mechanisms of protein unfolding and translocation mediated by clpy atpase nanomachines. *PLOS Computational Biology* **12**, e1004675.

[75] Chi, C. N, Bach, A, Engström, A, Strømgaard, K, Lundström, P, Ferguson, N, & Jemth, P. (2011) Biophysical characterization of the complex between human papillomavirus e6 protein and synapse-associated protein 97. *Journal of Biological Chemistry* **286**, 3597–3606.

[76] Nussinov, R. (2012) Allosteric modulators can restore function in an amino acid neurotransmitter receptor by slightly altering intra-molecular communication pathways. *British Journal of Pharmacology* **165**, 2110–2112.

[77] Deng, X, Gao, F, & May, W. S. (2009) Protein phosphatase 2a inactivates bcl2's antiapoptotic function by dephosphorylation and up-regulation of bcl2-p53 binding. *Blood* **113**, 422–428.

[78] Nussinov, R & Tsai, C.-J. (2013) Allostery in disease and in drug discovery. *Cell* **153**, 293–305.

[79] Shan, Y, Eastwood, M, Zhang, X, Kim, E, Arkhipov, A, Dror, R, Jumper, J, Kuriyan, J, & Shaw, D. (2012) Oncogenic mutations counteract intrinsic disorder in the egfr kinase and promote receptor dimerization. *Cell* **149**, 860–870.

[80] Nussinov, R, Tsai, C.-J, & Csermely, P. (2011) Allo-network drugs: harnessing allostery in cellular networks. *Trends in Pharmacological Sciences* **32**, 686–693.

[81] Sigel, E. (2002) Mapping of the benzodiazepine recognition site on gaba-a receptors. *Current Topics in Medicinal Chemistry* **2**, 833–839.

[82] Jensen, A & Brauner-Osborne, H. (2007) Allosteric modulation of the calcium-sensing receptor. *Current Neuropharmacology* **5**, 180–186.

[83] Sadowsky, J. D, Burlingame, M. A, Wolan, D. W, McClendon, C. L, Jacobson, M. P, & Wells, J. A. (2011) Turning a protein kinase on or off from a single allosteric site via disulfide trapping. *Proceedings of the National Academy of Sciences* **108**, 6056–6061.

[84] Gupta, M, Hendrickson, A. E. W, Yun, S. S, Han, J. J, Schneider, P. A, Koh, B. D, Stenson, M. J, Wellik, L. E, Shing, J. C, Peterson, K. L, Flatten, K. S, Hess, A. D, Smith, B. D, Karp, J. E, Barr, S, Witzig, T. E, & Kaufmann, S. H. (2012) Dual mtorc1/mtorc2 inhibition diminishes akt activation and induces puma-dependent apoptosis in lymphoid malignancies. *Blood* **119**, 476–487.

[85] Cram, D. J. (1992) Molecular container compounds. *Nature* **356**, 29–36.

[86] Lehn, J. (1995) *Supramolecular Chemistry: Concepts and Perspectives.* (Wiley).

[87] Cram, D. J. (1988) The design of molecular hosts, guests, and their complexes. *Science* **240**, 760–767.

[88] Aytuna, A. S, Gursoy, A, & Keskin, O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* **21**, 2850–2855.

[89] Malmström, L, Riffle, M, Strauss, C. E. M, Chivian, D, Davis, T. N, Bonneau, R, & Baker, D. (2007) Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biology* **5**, e76.

[90] Sharan, R, Ulitsky, I, & Shamir, R. (2007) Network-based prediction of protein function. *Molecular Systems Biology* **3**.

[91] Ewing, R. M, Chu, P, Elisma, F, Li, H, Taylor, P, Climie, S, McBroom-Cerajewski, L, Robinson, M. D, O'Connor, L, Li, M, Taylor, R, Dharsee, M, Ho, Y, Heilbut, A, Moore, L, Zhang, S, Ornatsky, O, Bukhman, Y. V, Ethier, M, Sheng, Y, Vasilescu, J, Abu-Farha, M, Lambert, J, Duewel, H. S, Stewart, I. I, Kuehl, B, Hogue, K, Colwill, K, Gladwish, K, Muskat, B, Kinach, R, Adams, S, Moran, M. F, Morin, G. B, Topaloglou, T, & Figeys, D. (2007) Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular Systems Biology* **3**.

[92] Hart, G. T, Ramani, A. K, & Marcotte, E. M. (2006). *Genome Biology* **7**, 120.

[93] Kim, P. M, Lu, L. J, Xia, Y, & Gerstein, M. B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938–1941.

[94] Tsai, C, Lin, S. L, Wolfson, H. J, & Nussinov, R. (1997) Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Science* **6**, 53–64.

[95] Young, L, Jernigan, R, & Covell, D. (1994) A role for surface hydrophobicity in protein-protein recognition. *Protein Science* **3**, 717–729.

[96] Xu, D, Tsai, C. J, & Nussinov, R. (1997) Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering Design and Selection* **10**, 999–1012.

[97] Norel, R, Sheinerman, F, Petrey, D, & Honig, B. (2001) Electrostatic contributions to protein–protein interactions: Fast energetic filters for docking and their physical basis. *Protein Science* **10**, 2147–2161.

[98] Sheinerman, F. B, Norel, R, & Honig, B. (2000) Electrostatic aspects of protein–protein interactions. *Current Opinion in Structural Biology* **10**, 153–159.

[99] Li, Y, Zhang, X, & Cao, D. (2013) The role of shape complementarity in the protein-protein interactions. *Scientific Reports* **3**.

[100] Sulea, T & Purisima, E. O. (2003) Profiling charge complementarity and selectivity for binding at the protein surface. *Biophysical Journal* **84**, 2883–2896.

[101] Teilum, K, Olsen, J. G, & Kragelund, B. B. (2021) On the specificity of protein–protein interactions in the context of disorder. *Biochemical Journal* **478**, 2035–2050.

[102] Dyson, H. J & Wright, P. E. (2005) Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology* **6**, 197–208.

[103] Gunasekaran, K, Tsai, C.-J, Kumar, S, Zanuy, D, & Nussinov, R. (2003) Extended disordered proteins: targeting function with less scaffold. *Trends in Biochemical Sciences* **28**, 81–85.

[104] Möglich, A, Joder, K, & Kiefhaber, T. (2006) End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proceedings of the National Academy of Sciences* **103**, 12394–12399.

[105] Shortle, D & Ackerman, M. S. (2001) Persistence of native-like topology in a denatured protein in 8 m urea. *Science* **293**, 487–489.

[106] Morris, O. M, Torpey, J. H, & Isaacson, R. L. (2021) Intrinsically disordered proteins: modes of binding with emphasis on disordered domains. *Open Biology* **11**.

[107] Sowmya, G & Ranganathan, S. (2014) Protein-protein interactions and prediction: a comprehensive overview. *Protein and peptide letters* **21**, 779–789.

[108] Schilder, J & Ubbink, M. (2013) Formation of transient protein complexes. *Current opinion in structural biology* **23**, 911–918.

[109] Hurley, J. H, Lee, S, & Prag, G. (2006) Ubiquitin-binding domains. *Biochemical Journal* **399**, 361–372.

[110] Vreven, T, Moal, I. H, Vangone, A, Pierce, B. G, Kastritis, P. L, Torchala, M, Chaleil, R, Jiménez-García, B, Bates, P. A, Fernandez-Recio, J, et al. (2015) Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology* **427**, 3031–3041.

[111] Schreiber, G & Fersht, A. R. (1993) Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry* **32**, 5145–5150.

[112] Goodsell, D. S & Olson, A. J. (2000) Structural symmetry and protein function. *Annual review of biophysics and biomolecular structure* **29**, 105–153.

[113] Levy, E. D, Pereira-Leal, J. B, Chothia, C, & Teichmann, S. A. (2006) 3d complex: A structural classification of protein complexes. *PLoS Computational Biology* **2**, e155.

[114] Garcia-Seisdedos, H, Empereur-Mot, C, Elad, N, & Levy, E. D. (2017) Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244–247.

[115] Cohen, M, Reichmann, D, Neuvirth, H, & Schreiber, G. (2008) Similar chemistry, but different bond preferences in inter versus intra-protein interactions. *Proteins: Structure, Function, and Bioinformatics* **72**, 741–753.

[116] La, D, Kong, M, Hoffman, W, Choi, Y. I, & Kihara, D. (2013) Predicting permanent and transient protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics* **81**, 805–818.

[117] Keskin, O, Tuncbag, N, & Gursoy, A. (2016) Predicting protein–protein interactions from the molecular to the proteome level. *Chemical reviews* **116**, 4884–4909.

[118] Kastritis, P. L & Bonvin, A. M. (2013) On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface* **10**, 20120835.

[119] Phillip, Y, Harel, M, Khait, R, Qin, S, Zhou, H.-X, & Schreiber, G. (2012) Contrasting factors on the kinetic path to protein complex formation diminish the effects of crowding agents. *Biophysical journal* **103**, 1011–1019.

[120] Kiel, C, Selzer, T, Shaul, Y, Schreiber, G, & Herrmann, C. (2004) Electrostatically optimized ras-binding ral guanine dissociation stimulator mutants increase the rate of association by stabilizing the encounter complex. *Proceedings of the National Academy of Sciences* **101**, 9223–9228.

[121] Harel, M, Spaar, A, & Schreiber, G. (2009) Fruitful and futile encounters along the association reaction between proteins. *Biophysical journal* **96**, 4237–4248.

[122] Schreiber, G, Haran, G, & Zhou, H.-X. (2009) Fundamental aspects of protein- protein association kinetics. *Chemical reviews* **109**, 839–860.

[123] Selzer, T, Albeck, S, & Schreiber, G. (2000) Rational design of faster associating and tighter binding protein complexes. *Nature structural biology* **7**, 537–541.

[124] Alsallaq, R & Zhou, H.-X. (2007) Prediction of protein-protein association rates from a transition-state theory. *Structure* **15**, 215–224.

[125] Netzer, R, Listov, D, Lipsh, R, Dym, O, Albeck, S, Knop, O, Kleanthous, C, & Fleishman, S. J. (2018) Ultrahigh specificity in a network of computationally designed protein-interaction pairs. *Nature communications* **9**, 5286.

[126] Baran, D, Pszolla, M. G, Lapidoth, G. D, Norn, C, Dym, O, Unger, T, Albeck, S, Tyka, M. D, & Fleishman, S. J. (2017) Principles for computational design of binding antibodies. *Proceedings of the National Academy of Sciences* **114**, 10900–10905.

[127] Conte, L. L, Chothia, C, & Janin, J. (1999) The atomic structure of protein-protein recognition sites. *Journal of molecular biology* **285**, 2177–2198.

[128] Mintseris, J & Weng, Z. (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences* **102**, 10930–10935.

[129] Levy, E. D. (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology* **403**, 660–670.

[130] Clackson, T & Wells, J. A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383–386.

[131] Scott, D. E, Bayly, A. R, Abell, C, & Skidmore, J. (2016) Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery* **15**, 533–550.

[132] Moreira, I. S, Fernandes, P. A, & Ramos, M. J. (2007) Hot spots—a review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics* **68**, 803–812.

[133] Chakrabarti, P & Janin, J. (2002) Dissecting protein–protein recognition sites. *Proteins: Structure, Function, and Bioinformatics* **47**, 334–343.

[134] Rigaut, G, Shevchenko, A, Rutz, B, Wilm, M, Mann, M, & Séraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology* **17**, 1030–1032.

[135] Rao, V. S, Srinivas, K, Sujini, G. N, & Kumar, G. N. S. (2014) Protein-protein interaction detection: Methods and analysis. *International Journal of Proteomics* **2014**, 1–12.

[136] Lin, J.-S & Lai, E.-M. (2017) *Protein–Protein Interactions: Co-Immunoprecipitation.* (Springer New York), p. 211–219.

[137] Emili, A. Q & Cagney, G. (2000) Large-scale functional analysis using peptide or protein arrays. *Nature biotechnology* **18**, 393–397.

[138] Kersten, B, Wanker, E. E, Hoheisel, J. D, & Angenendt, P. (2005) Multiplex approaches in protein microarray technology. *Expert Review of Proteomics* **2**, 499–510.

[139] Ramachandran, N, Hainsworth, E, Bhullar, B, Eisenstein, S, Rosen, B, Lau, A. Y, Walter, J. C, & LaBaer, J. (2004) Self-assembling protein microarrays. *Science* **305**, 86–90.

[140] Zhu, H, Bilgin, M, Bangham, R, Hall, D, Casamayor, A, Bertone, P, Lan, N, Jansen, R, Bidlingmaier, S, Houfek, T, et al. (2001) Global analysis of protein activities using proteome chips. *science* **293**, 2101–2105.

[141] Michnick, S. W, Ear, P. H, Landry, C, Malleshaiah, M. K, & Messier, V. (2011) *Protein-Fragment Complementation Assays for Large-Scale Analysis, Functional Dissection and Dynamic Studies of Protein–Protein Interactions in Living Cells.* (Humana Press), p. 395–425.

[142] Smith, G. P. (1985) Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317.

[143] Tong, A. H. Y, Drees, B, Nardelli, G, Bader, G. D, Brannetti, B, Castagnoli, L, Evangelista, M, Ferracuti, S, Nelson, B, Paoluzi, S, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324.

[144] Tong, A. H. Y, Evangelista, M, Parsons, A. B, Xu, H, Bader, G. D, Pagé, N, Robinson, M, Raghibizadeh, S, Hogue, C. W, Bussey, H, et al. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368.

[145] Uetz, P, Giot, L, Cagney, G, Mansfield, T. A, Judson, R. S, Knight, J. R, Lockshon, D, Narayan, V, Srinivasan, M, Pochart, P, Qureshi-Emili, A, Li, Y, Godwin, B, Conover, D, Kalbfleisch, T, Vijayadamodar, G, Yang, M, Johnston, M, Fields, S, & Rothberg, J. M. (2000) A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature* **403**, 623–627.

[146] Rutherford, S. L. (2000) From genotype to phenotype: buffering mechanisms and the storage of genetic information. *BioEssays* **22**, 1095–1105.

[147] Hartman IV, J. L. (2001) Principles for the buffering of genetic variation. *Science* **291**, 1001–1004.

[148] Lee, S.-A, Chan, C.-h, Tsai, C.-H, Lai, J.-M, Wang, F.-S, Kao, C.-Y, & Huang, C.-Y. F. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics* **9**.

[149] Lu, L, Lu, H, & Skolnick, J. (2002) Multiprospector: An algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins: Structure, Function, and Bioinformatics* **49**, 350–364.

[150] Yamada, M, Kabir, M. S, & Tsunedomi, R. (2003) Divergent promoter organization may be a preferred structure for gene control in escherichia coli. *Microbial Physiology* **6**, 206–210.

[151] Enright, A. J, Iliopoulos, I, Kyrpides, N. C, & Ouzounis, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90.

[152] Marcotte, E. M, Pellegrini, M, Ng, H.-L, Rice, D. W, Yeates, T. O, & Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753.

[153] Pazos, F & Valencia, A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function, and Bioinformatics* **47**, 219–227.

[154] Craig, R. A & Liao, L. (2007) Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. *BMC Bioinformatics* **8**.

[155] Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Research* **29**, 3513–3519.

[156] White, A. W, Westwell, A. D, & Brahemi, G. (2008) Protein–protein interactions as targets for small-molecule therapeutics in cancer. *Expert Reviews in Molecular Medicine* **10**.

[157] Blazer, L. L & Neubig, R. R. (2008) Small molecule protein–protein interaction inhibitors as cns therapeutic agents: Current progress and future hurdles. *Neuropsychopharmacology* **34**, 126–141.

[158] Coyne, A. G, Scott, D. E, & Abell, C. (2010) Drugging challenging targets using fragment-based approaches. *Current Opinion in Chemical Biology* **14**, 299–307.

[159] Winter, A, Higueruelo, A. P, Marsh, M, Sigurdardottir, A, Pitt, W. R, & Blundell, T. L. (2012) Biophysical and computational fragment-based approaches to targeting protein–protein interactions: applications in structure-guided drug discovery. *Quarterly Reviews of Biophysics* **45**, 383–426.

[160] Arkin, M. R & Wells, J. A. (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery* **3**, 301–317.

[161] Vassilev, L. T, Vu, B. T, Graves, B, Carvajal, D, Podlaski, F, Filipovic, Z, Kong, N, Kammlott, U, Lukacs, C, Klein, C, Fotouhi, N, & Liu, E. A. (2004) In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science* **303**, 844–848.

[162] Grasberger, B. L, Lu, T, Schubert, C, Parks, D. J, Carver, T. E, Koblish, H. K, Cummings, M. D, LaFrance, L. V, Milkiewicz, K. L, Calvo, R. R, Maguire, D, Lattanze, J, Franks, C. F, Zhao, S, Ramachandren, K, Bylebyl, G. R, Zhang, M, Manthey, C. L, Petrella, E. C, Pantoliano, M. W, Deckman, I. C, Spurlino, J. C, Maroney, A. C, Tomczuk, B. E, Molloy, C. J, & Bone, R. F. (2005) Discovery and cocrystal structure of benzodiazepinedione hdm2 antagonists that activate p53 in cells. *Journal of Medicinal Chemistry* **48**, 909–912.

[163] Allen, J. G, Bourbeau, M. P, Wohlhieter, G. E, Bartberger, M. D, Michelsen, K, Hungate, R, Gadwood, R. C, Gaston, R. D, Evans, B, Mann, L. W, Matison, M. E, Schneider, S, Huang, X, Yu, D, Andrews, P. S, Reichelt, A, Long, A. M, Yakowec, P, Yang, E. Y, Lee, T. A, & Oliner, J. D. (2009) Discovery and optimization of chromenotriazolopyrimidines as potent inhibitors of the mouse double minute 2–tumor protein 53 protein–protein interaction. *Journal of Medicinal Chemistry* **52**, 7044–7053.

[164] Hajduk, P. J & Greer, J. (2007) A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery* **6**, 211–219.

[165] Silvestre, H. L, Blundell, T. L, Abell, C, & Ciulli, A. (2013) Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery. *Proceedings of the National Academy of Sciences* **110**, 12984–12989.

[166] Magee, T. V. (2015) Progress in discovery of small-molecule modulators of protein–protein interactions via fragment screening. *Bioorganic & Medicinal Chemistry Letters* **25**, 2461–2468.

[167] Wu, B, Zhang, Z, Noberini, R, Barile, E, Giulianotti, M, Pinilla, C, Houghten, R. A, Pasquale, E. B, & Pellecchia, M. (2013) Hts by nmr of combinatorial libraries: a fragment-based approach to ligand discovery. *Chemistry & biology* **20**, 19–33.

[168] Lugovskoy, A. A, Degterev, A. I, Fahmy, A. F, Zhou, P, Gross, J. D, Yuan, J, & Wagner, G. (2002) A novel approach for characterizing protein ligand complexes: molecular basis for specificity of small-molecule bcl-2 inhibitors. *Journal of the American Chemical Society* **124**, 1234–1240.

[169] Sheng, C, Dong, G, Miao, Z, Zhang, W, & Wang, W. (2015) State-of-the-art strategies for targeting protein–protein interactions by small-molecule inhibitors. *Chemical Society Reviews* **44**, 8238–8259.

[170] Mason, J. M. (2010) Design and development of peptides and peptide mimetics as antagonists for therapeutic intervention. *Future Medicinal Chemistry* **2**, 1813–1822.

[171] Buckley, D. L, Gustafson, J. L, Van Molle, I, Roth, A. G, Tae, H. S, Gareiss, P. C, Jorgensen, W. L, Ciulli, A, & Crews, C. M. (2012) Small molecule inhibitors of the interaction between the e3 ligase vhl and hif1α. *Angewandte Chemie (International Ed. in English)* **51**, 11463.

[172] Yap, J. L, Wang, H, Hu, A, Chauhan, J, Jung, K.-Y, Gharavi, R. B, Prochownik, E. V, & Fletcher, S. (2013) Pharmacophore identification of c-myc inhibitor 10074-g5. *Bioorganic & Medicinal Chemistry Letters* **23**, 370–374.

[173] Yin, H, Lee, G.-i, Sedey, K. A, Kutzki, O, Park, H. S, Orner, B. P, Ernst, J. T, Wang, H.-G, Sebti, S. M, & Hamilton, A. D. (2005) Terphenyl-based bak bh3 α-helical proteomimetics as low-molecular-weight antagonists of bcl-xl. *Journal of the American Chemical Society* **127**, 10191–10196.

[174] Chen, L, Yin, H, Farooqi, B, Sebti, S, Hamilton, A. D, & Chen, J. (2005) p53 α-helix mimetics antagonize p53/mdm2 interaction and activate p53. *Molecular Cancer Therapeutics* **4**, 1019–1025.

[175] Scheper, J, Guerra-Rebollo, M, Sanclimens, G, Moure, A, Masip, I, González-Ruiz, D, Rubio, N, Crosas, B, Meca-Cortés, Ó, Loukili, N, et al. (2010) Protein-protein interaction antagonists as novel inhibitors of non-canonical polyubiquitylation. *PloS one* **5**, e11403.

[176] Lu, S, Shen, Q, & Zhang, J. (2019) Allosteric methods and their applications: Facilitating the discovery of allosteric drugs and the investigation of allosteric mechanisms. *Accounts of Chemical Research* **52**, 492–500.

[177] Changeux, J.-P. (2013) The concept of allosteric modulation: an overview. *Drug Discovery Today: Technologies* **10**, e223–e228.

[178] Petta, I, Lievens, S, Libert, C, Tavernier, J, & De Bosscher, K. (2016) Modulation of protein–protein interactions for the development of novel therapeutics. *Molecular Therapy* **24**, 707–718.

[179] Lu, H, Zhou, Q, He, J, Jiang, Z, Peng, C, Tong, R, & Shi, J. (2020) Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy* **5**.

[180] Du, X, Li, Y, Xia, Y.-L, Ai, S.-M, Liang, J, Sang, P, Ji, X.-L, & Liu, S.-Q. (2016) Insights into protein–ligand interactions: Mechanisms, models, and methods. *International Journal of Molecular Sciences* **17**, 144.

[181] (2013) Drug design: Methodology, concepts, and mode-of-action.

[182] Demchenko, A. P. (2001) Recognition between flexible protein molecules: induced and assisted folding. *Journal of Molecular Recognition* **14**, 42–61.

[183] Bongrand, P. (1999) Ligand-receptor interactions. *Reports on Progress in Physics* **62**, 921–968.

[184] Eisenreich, W, Rudel, T, Heesemann, J, & Goebel, W. (2021) Persistence of intracellular bacterial pathogens—with a focus on the metabolic perspective. *Frontiers in Cellular and Infection Microbiology* **10**.

[185] Moyed, H. S & Bertrand, K. P. (1983) hipa, a newly recognized gene of escherichia coli k-12 that affects frequency of persistence after inhibition of murein synthesis. *Journal of Bacteriology* **155**, 768–775.

[186] Jayaraman, R. (2008) Bacterial persistence: some new insights into an old phenomenon. *Journal of Biosciences* **33**, 795–805.

[187] Wen, Y, Sobott, F, & Devreese, B. (2016) Atp and autophosphorylation driven conformational changes of hipa kinase revealed by ion mobility and crosslinking mass spectrometry. *Analytical and Bioanalytical Chemistry* **408**, 5925–5933.

[188] Schumacher, M. A, Piro, K. M, Xu, W, Hansen, S, Lewis, K, & Brennan, R. G. (2009) Molecular mechanisms of hipa-mediated multidrug tolerance and its neutralization by hipb. *Science* **323**, 396–401.

[189] Jaffe, A. B & Hall, A. (2005) Rho gtpases: Biochemistry and biology. *Annual Review of Cell and Developmental Biology* **21**, 247–269.

[190] Ellenbroek, S. I. J & Collard, J. G. (2007) Rho gtpases: functions and association with cancer. *Clinical & Experimental Metastasis* **24**, 657–672.

[191] Dvorsky, R & Ahmadian, M. R. (2004) Always look on the bright site of rho: structural implications for a conserved intermolecular interface. *EMBO reports* **5**, 1130–1136.

[192] Cherfils, J & Zeghouf, M. (2013) Regulation of small gtpases by gefs, gaps, and gdis. *Physiological Reviews* **93**, 269–309.

[193] Garcia-Mata, R, Boulter, E, & Burridge, K. (2011) The "invisible hand": regulation of rho gtpases by rhogdis. *Nature Reviews Molecular Cell Biology* **12**, 493–504.

[194] Seidah, N. G, Benjannet, S, Wickham, L, Marcinkiewicz, J, Jasmin, S. B, Stifani, S, Basak, A, Prat, A, & Chrétien, M. (2003) The secretory proprotein convertase neural apoptosis-regulated convertase 1 (narc-1): Liver regeneration and neuronal differentiation. *Proceedings of the National Academy of Sciences* **100**, 928–933.

[195] Bao, X, Liang, Y, Chang, H, Cai, T, Feng, B, Gordon, K, Zhu, Y, Shi, H, He, Y, & Xie, L. (2024) Targeting proprotein convertase subtilisin/kexin type 9 (pcsk9): from bench to bedside. *Signal Transduction and Targeted Therapy* **9**.

[196] Maxwell, K. N, Fisher, E. A, & Breslow, J. L. (2005) Overexpression of pcsk9 accelerates the degradation of the ldlr in a post-endoplasmic reticulum compartment. *Proceedings of the National Academy of Sciences* **102**, 2069–2074.

[197] Maxwell, K. N & Breslow, J. L. (2004) Adenoviral-mediated expression of pcsk9 in mice results in a low-density lipoprotein receptor knockout phenotype. *Proceedings of the National Academy of Sciences* **101**, 7100–7105.

[198] Stein, E. A, Mellis, S, Yancopoulos, G. D, Stahl, N, Logan, D, Smith, W. B, Lisbon, E, Gutierrez, M, Webb, C, Wu, R, Du, Y, Kranz, T, Gasparino, E, & Swergold, G. D. (2012) Effect of a monoclonal antibody to pcsk9 on ldl cholesterol. *New England Journal of Medicine* **366**, 1108–1118.

[199] Sullivan, D, Olsson, A. G, Scott, R, Kim, J. B, Xue, A, Gebski, V, Wasserman, S. M, & Stein, E. A. (2012) Effect of a monoclonal antibody to pcsk9 on low-density lipoprotein cholesterol levels in statin-intolerant patients: The gauss randomized trial. *JAMA* **308**, 2497.

[200] Chai, M, He, Y, Zhao, W, Han, X, Zhao, G, Ma, X, Qiao, P, Shi, D, Liu, Y, Han, W, An, P, Li, H, Yan, S, Ma, Q, Deng, H, Qian, L, & Zhou, Y. (2023) Efficacy and safety of tafolecimab in chinese patients with heterozygous familial hypercholesterolemia: a randomized, double-blind, placebo-controlled phase 3 trial (credit-2). *BMC Medicine* **21**.

[201] Piper, D. E, Jackson, S, Liu, Q, Romanow, W. G, Shetterly, S, Thibault, S. T, Shan, B, & Walker, N. P. (2007) The crystal structure of pcsk9: A regulator of plasma ldl-cholesterol. *Structure* **15**, 545–552.

[202] BARTUS, R. T, DEAN, R. L, PONTECORVO, M. J, & FLICKER, C. (1985) The cholinergic hypothesis: A historical overview, current perspective, and future directions. *Annals of the New York Academy of Sciences* **444**, 332–358.

[203] Rosenberry, T. L. (1975) Acetylcholinesterase. *Advances in enzymology and related areas of molecular biology* **43**, 103–218.

[204] Colovic, M. B, Krstic, D. Z, Lazarevic-Pasti, T. D, Bondzic, A. M, & Vasic, V. M. (2013) Acetylcholinesterase inhibitors: pharmacology and toxicology. *Current neuropharmacology* **11**, 315–335.

2

Methodology

"A man who dares to waste one hour of time has not discovered the value of life"

Charles Darwin

2.1 MD Simulation of Biological Systems and its Historical Development

Biological macromolecules exhibit inherent dynamism, where internal motions and subsequent conformational changes profoundly influence their functional behavior. Achieving a comprehensive understanding of biological processes necessitates a thorough grasp of the structure, dynamics, energetics, and function of these macromolecules. Conventional experimental techniques like Raman spectroscopy, NMR, and X-ray crystallography have been extensively used to unravel the fundamental physiological principles. However, these techniques, while valuable, often provide only a

limited grasp of molecular interactions. To bridge the lack of understanding, it is necessary to provide a detailed description at the atomistic level, utilizing ideas derived from theoretical chemistry and statistical mechanics. Molecular Dynamics (MD) simulations are powerful computational tools used to study the physical movements of atoms and molecules. From its inception, the approach has seen significant advances, enabling thorough research of the dynamic behavior of proteins, nucleic acids, and other complex systems. Historical milestones in MD simulations include the development of the first algorithms for numerical integration of Newton's equations of motion and the creation of various force fields that accurately represent interatomic interactions. This method allows for the visualization of processes like folding of proteins, protein-drug interactions, and other functional dynamics that are frequently unattainable by experimental methods. The primary methodology employed in this thesis is MD simulation, which is utilized to thermodynamic and kinetic properties of different protein systems.

Earliest computer simulation study, dating back to 1953[2], focused on computing the thermodynamic features of liquids modeled as "hard sphere". This groundbreaking research established the fundamental principles that is now known as "Metropolis Monte Carlo simulation." Subsequently, the first molecular dynamics investigation, conducted by Alder and Wainright, examined dynamic properties within a system of hard spheres using Newton's equations of motion[3]. A significant advancement occurred in 1964 when Rahman pioneered the molecular dynamics simulation of liquid Argon using a soft sphere (Lennard-Jones) potential[4], marking a pivotal step toward modern simulations. Building upon these foundational studies, rapid progress ensued in modeling larger systems, including proteins. First MD simulation of a protein was conducted in a vacuum in 1977 by McCammon and Karplus[5]. This Field has experienced significant milestones and setbacks (Figure 2.1). In the 1980s, supercomputers facilitated remarkable progress[6,7]. The 1990s brought further disappointments, particularly with unfulfilled biomedical expectations[1], such as the slow translation of human genome information into medical solutions[8]. Limitations in force fields and conformational sampling further impeded practical applications.

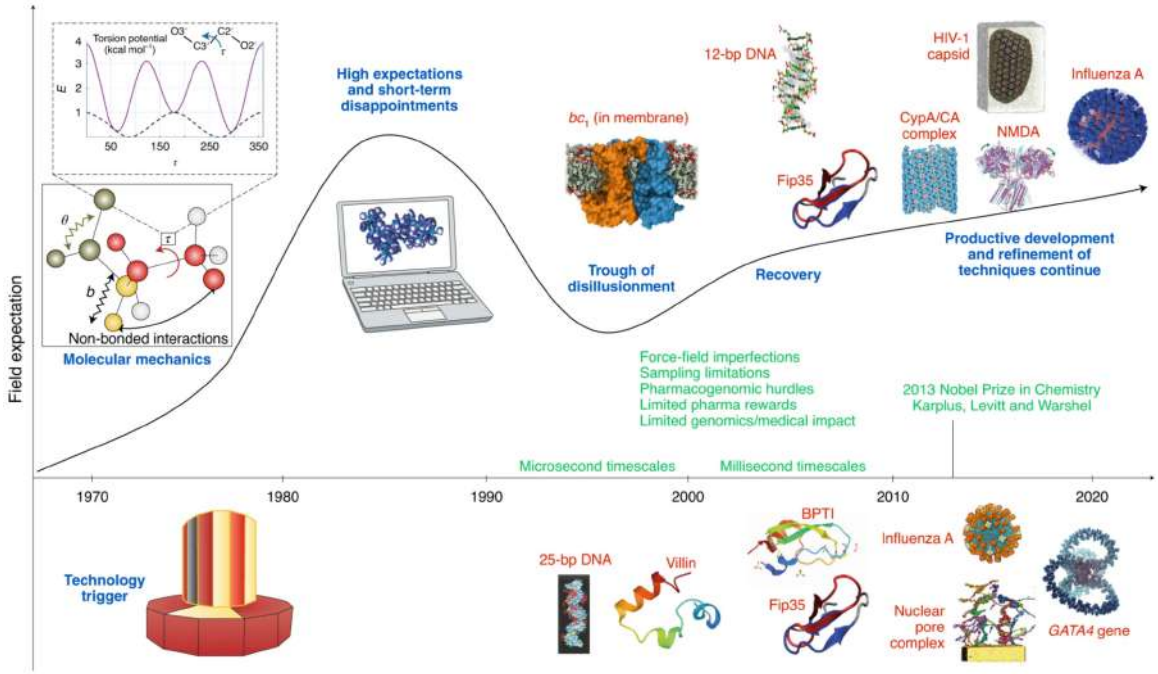

Figure 2.1: Historical Development of MD Simulation. Reprinted from "Biomolecular modeling thrives in the age of technology" by Tamar Schlick & Stephanie Portillo-Ledesma, 2021, Nature Computational Science, 1(5), 321-331. Copyright ©2021, Springer Nature America, Inc.[1]

However, this challenging period was succeeded by innovative approaches in both software and hardware, which addressed many of these limitations. Over the past two decades, the field has celebrated significant triumphs, including millisecond all-atom simulations of protein folding[9], insights into the mechanisms of large biomolecular networks such as virus simulations[10], and advancements in drug discovery, notably in the search for COVID-19 treatments[11]. The foundational work of force-field pioneers like Allinger, Lifson, Scheraga, and Kollman paved the way for these achievements. The 2013 Nobel Prize in Chemistry, given to Martin Karplus, Michael Levitt, and Arieh Warshel, highlighted the significance of computational techniques in the field of biology. Today, experimentation and modeling are integral and complementary components of a dynamic and successful scientific discipline.

2.2 Basic Theory of MD Simulation

Using Newton's equations of motion and classical mechanics, the positions and velocities of the atoms in an N-particle system are determined by molecular dynamics simulation:

$$F_i(t) = m_i \ddot{r}_i(t) \tag{2.1}$$

In this case, $r_i(t)$ is the atom's position at time t, m_i is its mass, and F_i is the force on i-th atom at that moment. The potential energy function is used to compute the force on atom i:

$$F_i(t) = -\frac{\partial V(r_i)}{\partial r_i} \tag{2.2}$$

The system's potential energy, denoted by $V(r_i)$, is determined by the atomic coordinates of the N particles. Each atom's position, velocity, and acceleration may be found using this force calculation.

2.2.1 Integration Methods

Multiple algorithms can be employed to integrate Newton's equations of motion in molecular dynamics simulations. A robust molecular dynamics program necessitates an effective integration algorithm that meets the following criteria:

1. Demands a minimal amount of cycles to compute forces every time step, resulting in computational efficiency and reduced memory use.
2. Must preserve both energy and momentum
3. Must exhibit time reversibility.
4. Should maintain the volume of the phase.
5. Demands more precision.
6. It is important for the system to remain stable and not generate any visual distortions while using greater time intervals (Δt).

Several integration methods are available, including the Verlet algorithm, Velocity-Verlet algorithm, and Leap-Frog algorithm. For this study, we have utilized the Leap-Frog algorithm due to its specific advantages.

Leap-Frog Integrator

The leap-frog method is developed from the Verlet algorithm, resulting in similar trajectories. It is the integrator that is used most frequently. This algorithm calculates the velocities at time steps that are halfway between two integers, and then uses these velocities to determine the new places. The method may be expressed as follows:

$$\begin{aligned} v\left(t+\frac{\Delta t}{2}\right) &= v\left(t-\frac{\Delta t}{2}\right)+a(t)\Delta t \\ r(t+\Delta t) &= r(t)+v\left(t+\frac{\Delta t}{2}\right)\Delta t \end{aligned} \tag{2.3}$$

This computation uses the position at time ($r(t)$) and the velocity at the previous half time step ($v(t-\Delta t/2)$) to determine the velocity at the next half time step ($v(t+\Delta t/2)$). Based on the provided information, it is feasible to ascertain the positions at the subsequent integer time step ($r(t+\Delta t)$). Thus, the velocity exceeds the position by half of the time step, and subsequently, the position overcomes the velocity to identify the location at the full-time step.

Notable features: :

1. Velocities are precisely computed, improving velocity evaluation.

2. More computationally intensive compared to the Verlet algorithm.

2.2.2 Force Fields in Molecular Dynamics

A force field serves as a mathematical model that illustrates how the energy of a system is influenced by the spatial arrangement of its components. It includes an analytical representation of the potential energy between atoms, denoted as $U(r_1, r_2, \ldots, r_N)$,

along with a set of parameters used in this representation. These parameters are typically derived from *ab initio* or semi-empirical quantum mechanical calculations, or by fitting them to experimental data obtained through techniques such as X-ray, and neutron diffraction, NMR, infrared, Raman spectroscopy.

In molecular dynamics simulations, the force field is vital as it allows for the modeling of systems with thousands of atoms over microsecond timescales, which would be computationally infeasible using direct *ab initio* methods. Essentially, a force field describes the interactions among atoms and determines the system's energy based on their positions. The accuracy of these simulations is strongly dependent on the chosen parameters and potential energy functions. A typical force field includes components that account for both bonded (covalent) and non-bonded (non-covalent) interactions. The potential energy V within a force field can be generally expressed by a specific equation.

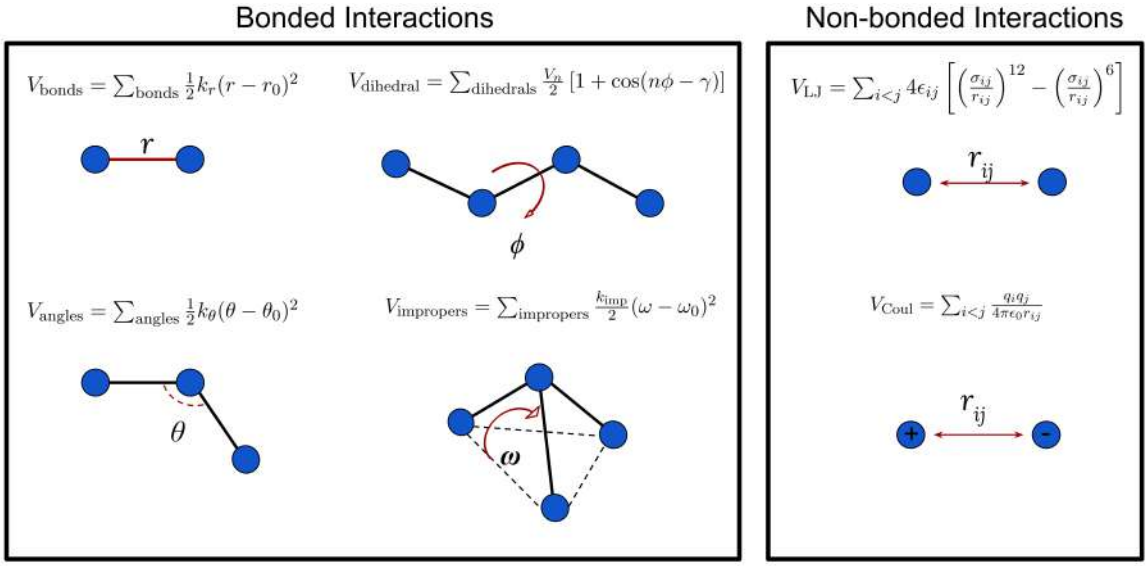

Figure 2.2: Schematic Representation of Bonded and Non-Bonded Interactions

A typical force field includes terms for both bonded (covalent) and non-bonded (non-covalent) interactions. A general expression for the potential energy V in a force field can be expressed as:

$$V = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{dihedral}} + V_{\text{impropers}} + V_{\text{LJ}} + V_{\text{Coul}} \tag{2.4}$$

Here, V_{bonds} describes the energy for the bond stretching, normally represented by a harmonic potential:

$$V_{\text{bonds}} = \sum_{\text{bonds}} \frac{1}{2} k_r (r - r_0)^2 \tag{2.5}$$

In this equation, k_r is the bond force constant and r_0 is the equilibrium bond length. This harmonic approximation assumes that bonds cannot break, limiting the simulation of chemical reactions.

Similarly, the angle bending energy V_{angles} is often modeled by another harmonic potential:

$$V_{\text{angles}} = \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2 \tag{2.6}$$

where k_θ is the angle force constant and θ_0 is the equilibrium bond angle.

For molecules with more than four atoms, the dihedral or torsional energy V_{dihedral} is included to account for rotations around bonds:

$$V_{\text{dihedral}} = \sum_{\text{dihedrals}} \frac{V_n}{2} \left[1 + \cos(n\phi - \gamma)\right] \tag{2.7}$$

In this term, V_n is the barrier height, ϕ represents the dihedral angle, γ represents the phase factor, and n represents the multiplicity. The combination of these terms allows for the construction of a dihedral potential with minima having different depths.

Additionally, improper torsion potentials $V_{\text{impropers}}$ are included to ensure the planarity of certain groups, such as sp2 hybridized carbons in carbonyl groups or aromatic rings:

$$V_{\text{impropers}} = \sum_{\text{impropers}} \frac{k_{\text{imp}}}{2} (\omega - \omega_0)^2 \tag{2.8}$$

where k_{imp} is the force constant for the improper torsion, and ω is the improper angle deviation from planarity.

Non-bonded interactions include Van der Waals forces and electrostatic interactions. The Van der Waals energy V_{LJ} is typically modeled using the 12-6 Lennard-Jones potential:

$$V_{\text{LJ}} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.9}$$

Here, ϵ_{ij} represents the depth of the potential well, σ_{ij} is the distance at which the inter-particle potential is zero, and r_{ij} is the distance between particles i and j.

The Coulombic energy V_{Coul} for electrostatic interactions is given by:

$$V_{\text{Coul}} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.10}$$

where q_i and q_j are the partial charges on atoms i and j, ϵ_0 is the permittivity of free space, and r_{ij} is the distance between the charged particles.

The specific form and parameters of the potential function can vary among different force fields, depending on the systems and properties used for their characterization. In this study, various force fields such as AMBER99SB-ILDN[12], AMBER14SB[13], and CHARMM36m[14] have been employed for protein simulations, each selected based on its suitability for different aspects of the research.

2.2.3 Technical Details

Periodic Boundary Conditions (PBC) and Minimum Image Convention

MD simulations are computational techniques used to investigate the thermodynamic properties of large-scale systems by representing a limited number of particles, usually ranging from hundreds to thousands. One significant challenge in these simulations is the presence of surface effects, where particles at the boundaries of the simulation box experience different forces compared to those in the bulk. This discrepancy arises because a large fraction of particles lie on the surface, unlike in a real bulk system

where the number of surface particles is negligible.
To address this, periodic boundary conditions (PBCs) are implemented. In PBCs, the simulation box containing N particles is replicated infinitely in all directions, creating a lattice of identical boxes. Each particle in the central simulation box interacts not only with other particles within the same box but also with particles in the adjacent periodic images of the box. This setup effectively removes the surface and edge effects by making the simulation box "wall-less," as any particle leaving the box on one side re-enters from the opposite side. Thus, PBCs ensure a consistent number of particles and density throughout the simulation, closely mimicking an infinite bulk system.
The use of periodic boundary conditions (PBCs) increses the number of interacting particles since each particle within the box can interact with its periodic counterparts. The minimum image convention (MIC) addresses this by ensuring that each particle interacts only with the closest image of any other particle, thereby streamlining the calculations.

Short Range and Long Range Interactions

To further reduce computational demands, interaction truncation is applied. Short-range interactions, which decay rapidly with distance, are truncated beyond a cutoff distance r_c. The potential function $U(r)$ is set to zero for $r > r_c$, which simplifies the calculations:

$$U_{\text{trunc}}(r) = \begin{cases} U(r) & \text{for } r \leqslant r_c \\ 0 & \text{for } r > r_c \end{cases} \tag{2.11}$$

This approach ensures that only interactions within the cutoff distance are considered. To avoid discontinuities in the potential at r_c, a shifted potential is often used:

$$U_{\text{trunc}}(r) = \begin{cases} U(r) - U(r_c) & \text{for } r \leqslant r_c \\ 0 & \text{for } r > r_c \end{cases} \tag{2.12}$$

The cutoff distance r_c must be less than half the box length to prevent a particle from interacting with multiple images.

Long-range interactions, such as electrostatic forces, pose a challenge in molecular dynamics simulations because each particle interacts with all other particles in the system, including their periodic images. These interactions decay slowly and cannot be truncated like short-range interactions, as doing so would result in significant errors.

Various methods have been developed to handle long-range interactions efficiently. One common technique is the Ewald summation[15], which decomposes the interaction potential into short-range and long-range components. The total electrostatic potential energy V is calculated as a sum of contributions from real space, reciprocal space, self-interaction corrections, and dipolar corrections:

$$V = V_{\text{real}} + V_{\text{reciprocal}} + V_{\text{self}} + V_{\text{dipolar}} \tag{2.13}$$

The real space contribution V_{real} is computed by direct summation over the short-range part of the potential, while the reciprocal space contribution $V_{\text{reciprocal}}$ is evaluated using Fourier transforms. The self-interaction term V_{self} corrects for the interaction of a particle with its own periodic images, and the dipolar term V_{dipolar} accounts for the interaction with an assumed uniform background charge.

Another efficient method is the Particle Mesh Ewald (PME) summation[16], which improves the computational cost of the reciprocal space calculations. PME uses a mesh-based approach to interpolate charges onto a grid and employs Fast Fourier Transform (FFT) to solve the Poisson equation. This method scales as $\mathcal{O}(N \log N)$, making it suitable for large systems.

In summary, long-range interactions require specialized techniques like Ewald summation and Particle Mesh Ewald to ensure accurate and efficient calculations in molecular dynamics simulations.

Neighbour and Cell Lists

In molecular dynamics simulations, truncating interactions helps reduce computational load, but it still requires calculating distances between each atom and all others to identify those within a cutoff radius. This results in a time complexity of N^2, where

N is the number of atoms. To accelerate computations, techniques like neighbor and cell lists are employed.
The neighbor list method involves setting a cutoff radius r_{list} larger than the interaction cutoff r_{c} and maintaining a list of atoms within this radius. Interactions are then computed only for atoms in this list, reducing unnecessary calculations and enhancing overall efficiency. The list is updated periodically when atoms move beyond $r_{\text{list}} - r_{\text{c}}$. Alternatively, the cell list method is used for large atom counts and simulation boxes much larger than the cutoff radius. Here, the box is divided into cells, each slightly larger than r_{c}. Initially, a list of atoms in each cell and neighboring cells is generated, remaining unchanged unless the simulation domain alters. At each time step, interactions for an atom are computed with those within its cell and neighboring cells, ensuring efficiency with a scaling of N.

2.2.4 Thermodynamics Ensembles

Nevertheless, the majority of practical situations entail engagements with the environment and variations in energy, volume, or particle count. Molecular dynamics simulations tackle this issue by employing two thermodynamic ensembles: the NVT ensemble and the NPT ensemble. In the NVT ensemble, the temperature, particle number, and volume are kept constant. In the NPT ensemble, the temperature, particle number, and pressure are maintained at a constant level, while allowing for fluctuations in volume. In order to create such ensembles, the system is connected to a thermostat or barostat, which are external baths that maintain a constant temperature and pressure, respectively. The NVT ensemble was utilized to achieve temperature equilibration, whereas the NPT ensemble was applied for all production runs in this study.

Temperature Regulation Methods in Simulations (Thermostat)

The system's temperature is associated with its kinetic energy as per the equipartition theorem. There are several methods to maintain the temperature of the system:

1. **Velocity rescaling**: involves modifying particle velocities to align with the target temperature. The scaling factor, denoted by λ, is calculated as the ratio of the intended temperature (T_B) to the instantaneous temperature before scaling ($T(t)$), as shown by the equation

$$\lambda = \sqrt{\frac{T_B}{T(t)}} \tag{2.14}$$

2. **Berendsen thermostat**: The Berendsen thermostat[17]functions as an improved version of velocity rescaling, utilizing a weak connection to an external heat bath that is maintained at a certain target temperature (T_B). This technique modifies Newton's equations of motion by incorporating a pseudo friction coefficient ($\gamma(t)$). The temperature scaling factor (λ) is calculated using the equation:

$$\lambda = \sqrt{\frac{\Delta t\,(T_B)}{1 + \tau\,(\langle T\rangle - 1)}} \tag{2.15}$$

In this context, Δt represents the duration of each time step, whereas τ is the coupling parameter that characterizes the interaction between the heat bath and the system. In this thesis, we employed a modified Berendsen thermostat[18], which included an extra stochastic term to guarantee the proper formation of a Canonical ensemble. The eq2.16 represents the rate of change of kinetic energy (dK) as a function of time (t) and other variables. It may be expressed as follows:

$$\frac{dK}{dt} = \frac{(K_0 - K)}{\tau_T} + \frac{2KK_0dW}{N_f\tau_T} \tag{2.16}$$

The variables in the equation are defined as follows: K represents kinetic energy, N_f represents the number of degrees of freedom, dW represents a Wiener process, and τ_T represents the temperature coupling time constant.

Other methods include the Anderson thermostat[19], Nose-Hoover thermostat[20,21], among others, which provide alternative approaches to temperature regulation in simulations. In this thesis, we have used Berendsen thermostat for temperature regulation.

Pressure Regulation Methods in Simulations (Barostat)

Constant pressure simulations, similar to thermostats, involve linking the system to a pressure bath maintained at a required pressure. This is accomplished by adjusting the volume of the box. In NPT ensemble, the cell's shape and size are modified to maintain a steady pressure. The rate of pressure change over time is determined by the following equation:

$$\frac{dP}{dt} = \frac{1}{\tau_p}(P_0 - P) \tag{2.17}$$

$$\lambda = 1 - \kappa\frac{\delta t}{\tau_p}(P_0 - P) \tag{2.18}$$

In this equation, κ represents the isothermal compressibility of the system.

$$\kappa = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \tag{2.19}$$

Therefore, at each iteration, the coordinates and dimensions of the box are modified by a factor of λ in order to maintain the pressure at P_0. The study utilized the Parrinello-Rahman Pressure Coupling technique[22]. In contant pressure simulations, the system can be connected to a pressure bath, similar to thermostats, in order to regulate the required pressure by adjusting the volume of the simulation box. This ensures consistent pressure inside the NPT ensemble. The rate of pressure change over time is dictated by the isothermal compressibility of the system.

2.2.5 Steps of Running an MD Simulation

System Preparation

Initiating a simulation involves establishing the starting configuration of the system. This can be achieved by utilizing pre-established structural models or by randomly dispersing molecules within a simulation box that accurately mimics the volume of the system. Proteins and nucleic acids, which are biological entities, can have their atomic coordinates determined by studying three-dimensional structures obtained

using techniques such as X-ray diffraction or NMR. Subsequently, a suitable force field is chosen to allocate non-bonded and bonded characteristics to the atoms inside the system, facilitating the calculation of potential energy at a later stage. To accurately mimic biological conditions, the system (such as proteins or nucleic acids) is submerged in water and ions are put into the simulation box.

Energy Minimization and Equilibration

Afterwards, the system is subjected to energy minimiation using techniques such as steepest descent or conjugate gradient in order to address any atom overlaps and steric stresses caused by atoms being positioned too near together. The atoms' initial velocities are allocated randomly from a Maxwellian distribution. Then, the resultant velocities are adjusted to scale the average kinetic energy to the required value. To address the potential deviation of the initial state from thermodynamic equilibrium, a two-step equilibration process is implemented. This procedure involves linking the system to a thermostat and a barostat, which respectively regulate temperature (NVT) and pressure (NPT). The purpose of this equilibration process is to bring the system to a state of equilibrium. At times, some portions of the solute in the system are held in place to help the solvent component reach equilibrium under specific conditions.

Production Run

After achieving equilibrium, the system undergoes evolution within the conformational space during the production run. The time intervals used to calculate location and velocities should be large enough to accommodate the fastest event in the simulation, which takes around 2 femtoseconds. Additionally, the production run should be of sufficient duration to enable thorough sampling of the phase space.

2.2.6 Analysis of MD Trajectory

Data on the position and velocity of atoms is collected during the simulation run. This data can then be used to compute various physical properties of the system. A reaction coordinate, also known as an order parameter, is a parameter that represents the progress of a transition between different states in a molecular system. It is a way to describe the changes occurring during a conformational transition in a simplified manner, often reducing a multi-dimensional problem to a lower dimension.

Traditional or Intution Based Order Parameters

Mass Weighted Root Mean Square Deviation (RMSD): Mass-weighted RMSD is a metric that quantifies the average spatial separation between atoms in aligned proteins, taking into account the masses of the atoms. It is characterized or described as:

$$\mathrm{RMSD}(t) = \sqrt{\frac{\sum_{i=1}^{N} m_i[\mathbf{r}_i^{\mathrm{ref}} - \mathbf{r}_i^{\mathrm{sim}}(t)]^2}{\sum_{i=1}^{N} m_i}} \tag{2.20}$$

where N is the number of atoms, m_i is the mass of the i-th atom, $\mathbf{r}_i^{\mathrm{ref}}$ is the position of the i-th atom in the reference structure, and $\mathbf{r}_i^{\mathrm{sim}}$ is the position of the i-th atom in the simulated structure.

Mass weighted RMSD is used to quantify the similarity between protein structures, accounting for the contribution of heavier atoms more significantly. It helps in comparing the conformational changes of a protein over time or between different conditions (e.g., bound vs. unbound states).

Root Mean Square Fluctuation (RMSF): RMSF quantifies the oscillation of individual atoms relative to their mean location within a given time period. It is defined as:

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle)^2} \tag{2.21}$$

where T is the total number of time frames, $\mathbf{r}_i(t)$ is the position of the i-th atom at time t, and $\langle \mathbf{r}_i \rangle$ is the average position of the i-th atom.

Usage:
RMSF provides insights into the flexibility and dynamic behavior of individual residues in a protein. It is particularly useful for identifying regions with high or low mobility, which can be important for understanding protein function and stability.

Radius of Gyration (Rg): The Radius of Gyration is a metric that quantifies the degree of compactness in a protein structure. It is characterized or described as:

$$\mathrm{R}_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{\mathrm{cm}})^2} \tag{2.22}$$

The variables in the equation are defined as follows: N represents the total number of atoms, $\mathbf{r}_i$ represents the position of the i-th atom, and $\mathbf{r}_{\mathrm{cm}}$ represents the center of mass of the protein.
Rg is used to assess the overall size and compactness of the protein structure. It helps in understanding the folding and unfolding processes and comparing the structural compactness between different states of the protein.

Hydrogen Bond Occupancy Analysis: A hydrogen bond is defined based on specific geometric criteria:

- The donor and the acceptor atom distance must be ($r \leqslant 0.35\,\mathrm{nm}$).
- A-D-H angle has a maximum value of 30 degrees A-D-H ($\alpha \leqslant 30°$).

Hydrogen bond occupancy is the fraction of simulation time during which a particular hydrogen bond exists.
Hydrogen bond occupancy analysis helps in understanding the stability and interactions within a protein or between proteins and other molecules. It provides insights into key interactions that stabilize the protein structure or facilitate binding.

Coordination Number (CN): It is a measure of the number of number of contacts i.e interactions between two groups of atoms, A and B. It is calculated using the following formula:

$$CN = \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \frac{\left(1 - \left(\frac{r_{ij}}{d}\right)^{12}\right)}{\left(1 - \left(\frac{r_{ij}}{d}\right)^{6}\right)} \tag{2.23}$$

The variables i and j are used to indicate the atoms that have been chosen, whereas the variables N_a and N_b represent the number of atoms that belong to groups a and b, respectively. It is the distance between the i-th and j-th atoms that is represented by the symbol r_{ij}. In accordance with the predetermined criteria for hydrogen bond lengths, the cutoff distance, which is designated by the letter d and is fixed at 3.0Å, is a perfect match. In light of this, the CN provides a computation of the number of hydrogen bonds that are formed between the atoms that are specified as being in the indicated groups.

Dimensionality Reduction Techniques

PCA: Principal Component Analysis (PCA) is a widely used statistical technique that aims to reduce the dimensionality of a dataset while preserving as much variance as possible. It achieves this by transforming the original data into a new coordinate system where the greatest variances by any projection of the data come to lie on the first coordinates (called principal components), the second greatest variances on the second coordinates, and so on. PCA is particularly useful in simplifying complex datasets, identifying patterns, and highlighting differences and similarities within the data.

Mathematically, PCA involves the following steps:

Given a dataset X consisting of T samples, each with N features:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_T \end{pmatrix} \in \mathbb{R}^{N \times T} \tag{2.24}$$

where x_t represents the feature vector at sample t.

The data is mean-centered by subtracting the mean of each feature across all samples:

$$\tilde{X} = X - \mu, \quad \text{where} \quad \mu = \frac{1}{T} \sum_{t=1}^{T} x_t \tag{2.25}$$

Compute the covariance matrix C:

$$C = \frac{1}{T}\tilde{X}\tilde{X}^T$$

Solve the eigenvalue problem for the covariance matrix:

$$C\vec{v}_i = \lambda_i \vec{v}_i \tag{2.26}$$

where λ_i are the eigenvalues and $\vec{v}_i$ are the corresponding eigenvectors.

Sort the eigenvalues λ_i in descending order and reorder the eigenvectors $\vec{v}_i$ accordingly. Select the top d eigenvectors corresponding to the largest d eigenvalues to form the transformation matrix V:

$$V = \begin{pmatrix} \vec{v}_1 & \vec{v}_2 & \dots & \vec{v}_d \end{pmatrix} \tag{2.27}$$

Project the mean-centered data onto the selected eigenvectors to obtain the principal components:

$$Y = V^T \tilde{X} \tag{2.28}$$

where Y represents the principal components, capturing the most significant variances in the system.

Key Points of PCA:

- PCA is a technique for reducing the dimensionality of a dataset while retaining the most important variance.
- It identifies patterns in the data by transforming the original variables into a new set of orthogonal variables (principal components).
- PCA helps in visualizing and understanding complex datasets by highlighting the most significant directions of variance.
- It is computationally efficient and widely used in data analysis and preprocessing, particularly in large datasets.
- PCA is applicable in various fields, including molecular dynamics, image processing, and finance, where dimensionality reduction and pattern recognition are essential.

TICA Time-lagged Independent Component Analysis (tICA)[23–25] is a statistical technique designed to identify and analyze the slow collective motions in time-series data. It is particularly useful in the context of molecular dynamics simulations, where understanding these slow dynamics can reveal important insights into the conformational changes and functional mechanisms of biomolecules. By focusing on the slowest motions, tICA helps in reducing the dimensionality of complex datasets, making it easier to identify and interpret significant dynamical processes.

Mathematically, tICA involves the following steps:

Given time-series data X consisting of T frames, each with N features:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_T \end{pmatrix} \in \mathbb{R}^{N \times T} \tag{2.29}$$

where x_t represents the feature vector at time t.

The data is mean-centered by subtracting the mean of each feature across all time frames:

$$\tilde{X} = X - \mu, \quad \text{where} \quad \mu = \frac{1}{T} \sum_{t=1}^{T} x_t \tag{2.30}$$

Compute the instantaneous covariance matrix C_0 and the time-lagged covariance matrix C_τ:

$$C_0 = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \tilde{x}_t \tilde{x}_t^T \tag{2.31}$$

$$C_\tau = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \tilde{x}_t \tilde{x}_{t+\tau}^T \tag{2.32}$$

where τ is the chosen time lag. Solve the generalized eigenvalue problem for the covariance matrices:

$$C_\tau \vec{v}_i = \lambda_i C_0 \vec{v}_i \tag{2.33}$$

where λ_i are the eigenvalues and $\vec{v}_i$ are the corresponding eigenvectors.

Sort the eigenvalues λ_i in descending order and reorder the eigenvectors $\vec{v}_i$ accordingly. Select the top d eigenvectors corresponding to the largest d eigenvalues to form the transformation matrix V:

$$V = \begin{pmatrix} \vec{v}_1 & \vec{v}_2 & \dots & \vec{v}_d \end{pmatrix} \tag{2.34}$$

Project the mean-centered data onto the selected eigenvectors to obtain the tICA components:

$$Y = V^T \tilde{X} \tag{2.35}$$

where Y represents the tICA components, capturing the slow collective motions in the system.

Key Points of tICA:

- tICA reduces the dimensionality of complex datasets, focusing on the most significant slow dynamics.
- The method helps in identifying and characterizing the slow transitions and conformational changes in molecular systems.
- By isolating the slowest collective motions, tICA makes it easier to interpret the underlying processes within the dataset.
- tICA is particularly useful in molecular dynamics simulations and other time-series data analyses where slow dynamics play a crucial role.
- The method is computationally efficient, making it suitable for analyzing large datasets typical in molecular dynamics studies.

2.2.7 Residue and Pairwise Change in Electrostatic Interaction:

The change in electrostatic interaction energy ($\Delta\langle E_i \rangle$) is a measure of how the electrostatic interactions between residues are altered due to a structural change, such as mutation of ligand binding. This is particularly useful for detecting subtle structural changes and long-range allosteric effects in proteins.

$$\Delta\langle E_i \rangle = \langle E_i^B \rangle - \langle E_i^A \rangle \tag{2.36}$$

where $\langle E_i^B \rangle$ and $\langle E_i^A \rangle$ represent the average nonbonded interaction energy of the i-th residue with its environment (protein, water, and ions) for unperturbed system (wild

type or apo type) and perturbed system (mutated type or ligand bound type) respectively. It is pivotal in understanding the propagation of the energetic perturbation. Similarly, for residue pairs:

$$\Delta\langle E_{ij}\rangle = \langle E_{ij}^{B}\rangle - \langle E_{ij}^{A}\rangle \tag{2.37}$$

where $\langle E_{ij}^{B}\rangle$ and $\langle E_{ij}^{A}\rangle$ represent the average pairwise electrostatic interaction energy of the i-th residue with j-th residue for unperturbed system (wild type or apo type) and perturbed system (mutated type or ligand bound type) respectively.
This analysis helps to understand how local perturbations, such as mutation or ligand binding, affect the overall electrostatic environment of the protein and propagate through the structure. By comparing the electrostatic interaction energies in different states, one can identify which residues are most affected by the perturbation and how the signal transmits to distant parts of the protein, potentially influencing protein-protein interactions (PPI).

2.2.8 Binding Energy Calculation Using MM/PBSA Method:

The binding free energy of a complex can be estimated using the following equation:

$$\Delta G_{\text{bind}} = \langle G_{\text{Com}}\rangle - \langle G_{\text{Rec}}\rangle - \langle G_{\text{Lig}}\rangle \tag{2.38}$$

where each of the term is defined as :

$$\langle G_{\text{x}}\rangle = \langle E_{\text{MM}}\rangle - \langle G_{\text{Sol}}\rangle - \langle TS\rangle \tag{2.39}$$

The binding energy can also be expressed as:

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S \tag{2.40}$$

where ΔH represents the enthalpy of binding and -TΔS accounts for the conformational entropy change upon ligand binding. When the entropic term is neglected, the

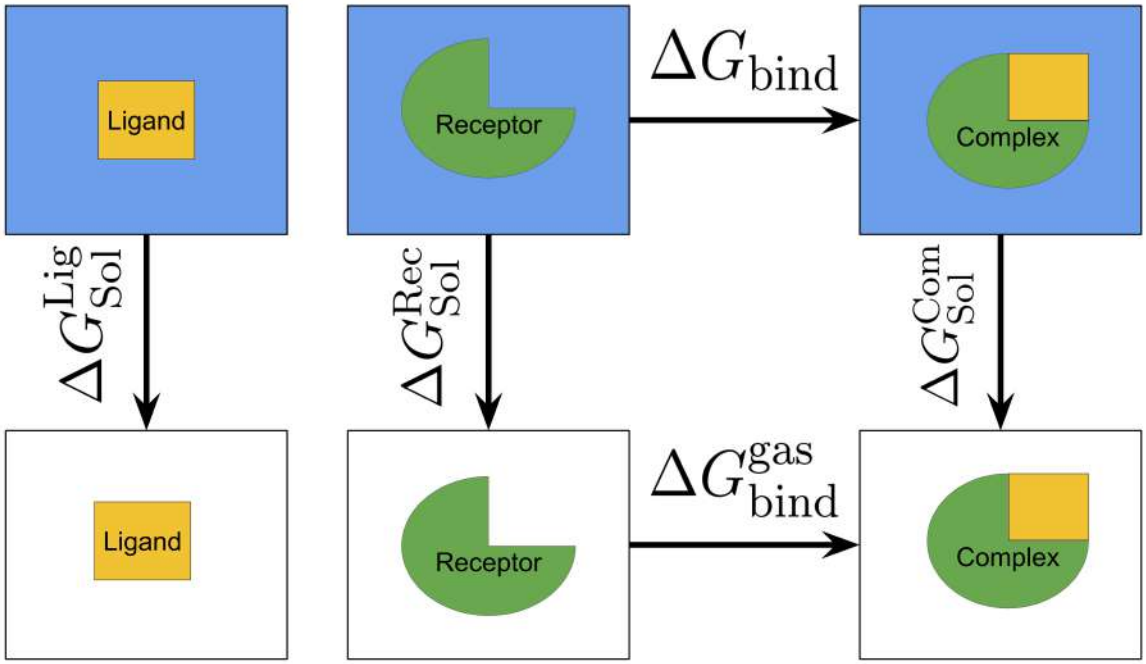

Figure 2.3: Thermodynamic Cycle Used to Calculate Binding Free Energy Using MM-PBSA

resulting value is the effective free energy, which is typically sufficient for comparing relative binding free energies of related ligands. The enthalpy change (ΔH) is decomposed into:

$$\Delta H = \Delta E_{\text{MM}} + \Delta G_{\text{Sol}} \tag{2.41}$$

where,

$$\Delta E_{\text{MM}} = \Delta E_{\text{Ele}} + \Delta E_{\text{vdw}} \tag{2.42}$$

The solvation free energy (ΔG_{Sol}) is given by:

$$\Delta G_{\text{Sol}} = \Delta G_{\text{Polar}} + \Delta G_{\text{Non-Polar}} = \Delta G_{\text{PB}} + \Delta G_{\text{Non-Polar}} \tag{2.43}$$

For the non-polar component:

$$\Delta G_{\text{Non-Polar}} = NP_{\text{TENSION}} \times \Delta SASA + NP_{\text{OFFSET}} \tag{2.44}$$

or,

$$\begin{aligned}\Delta G_{\text{Non-Polar}} &= \Delta G_{\text{disp}} + \Delta G_{\text{Cavity}} \\ &= \Delta G_{\text{disp}} + CAVITY_{\text{TENSION}} \times \Delta SASA + CAVITY_{\text{OFFSET}}\end{aligned} \tag{2.45}$$

In the equations provided,ΔE_{MM} corresponds to the alterations in molecular mechanical energy that occur in the gas phase. It consists of different components, the electrostatic interaction (ΔG_{Ele}) ,and the van der Waals interaction (ΔG_{vdW}). The solvation energy (ΔG_{Sol}) is determined differently based on the method used. While the Poisson-Boltzmann(PB) model estimate only the polar component. The non-polar component is often directly proportional to the solvent accessible surface area (SASA), using a proportionality constant determined from experimental data on tiny non-polar compounds. Alternatively, contemporary techniques divide non-polar solvation free energy into two components: cavity and dispersion. The cavity term is determined using SASA, while the dispersion term is calculated using surface-integration methods. We have not yet computed the entropic contribution to the binding energy in our works.

2.3 Free Energy Calculation

The underlying free energy landscape plays a vital role in driving biophysical and biochemical processes, like protein-protein interaction, allostery, molecular recognition and folding of protein. In order to comprehend these mechanisms at the molecular level, it is necessary to elucidate these processes in the context of thermodynamics. Statistical mechanics is essential in this context as it offers a probability distribution that links the microstates of a system to macroscopic thermodynamic variables such as free energy, temperature, and density. Free energy is a fundamental thermodynamic property that is used to describe the tendency of a chemical process to occur spontaneously and to determine the equilibrium state of a system. The free energy of a system is commonly represented by either the Helmholtz or Gibbs free energy. The Helmholtz free energy in a canonical ensemble may be determined using statistical

mechanics, and it is represented by the partition function Q.

$$A = -k_B T \ln Q(N, V, T) \tag{2.46}$$

Similarly, in an isothermal-isobaric ensemble, the Gibbs free energy can be described as:

$$G = -k_B T \ln Z(N, P, T) \tag{2.47}$$

The partition function Q(N,V,T) for an N-particle system in a canonical ensemble is defined as:

$$Q(N, V, T) = \frac{1}{h^N N!} \int \int e^{-\frac{H(p,r)}{k_B T}} \, d^3p \, d^3r \tag{2.48}$$

Here, r and p are 3N-dimensional vectors representing the coordinates and momenta of all particles in the system. Processes like enzyme-catalyzed transformations or protein folding occur in high-dimensional phase spaces with numerous degrees of freedom, leading to complex energy landscapes. In practical biochemical systems, the presence of solvent molecules, ions, and other substrates leads to the development of low-energy states on the free energy surface, which are filled with different conformations. Consequently, free energy calculations are frequently conducted using a series of coordinates that describe the advancement of a chemical process, referred to as collective variables(CV) or reaction coordinates(RC). These functions are commonly determined by the Cartesian coordinates of atoms in the system. The selection of reaction coordinates is contingent upon the particular reaction under investigation. In a bond-dissociation process, the bond length between the atoms serves as a measure to define the advancement of the reaction. Protein folding and unfolding may be analyzed by employing several measures, such as dihedral angles (ϕ, ψ), radius of gyration (R_g) root-mean-square deviation (RMSD), and the number of native contacts (N_C).

2.3.1 Potential of Mean Force

Although theoretically, the precise free energy of a system may be computed using the partition function, in reality, it becomes difficult for systems with significant free

energy barriers. Once an appropriate reaction coordinate is determined, a free energy profile may be created by averaging across the remaining degrees of freedom. The expression provided yields the potential of mean force (PMF):

$$A(X) = -k_B T \ln \left(\int \delta(f(r^N) - X) \exp\left(-\frac{U(r^N)}{k_B T} \right) dr^N \right) \tag{2.49}$$

The symbol $\delta(f(r^N)–X)$ represents the Dirac delta function applied to the reaction coordinate X, which relates to the location coordinates of all particles (r_N). This equation calculates the diminished amount of energy that is available during a reaction, while considering all other factors and variables. The probability distribution of the response coordinate may also be determined by utilizing the average distribution function.

$$A(X) = -k_B T \ln P(X) + c \tag{2.50}$$

P(X) represents the probability distribution for the reaction coordinate X. This method has been effectively utilized to compute the changes in free energy in various processes like ligand binding, protein folding.[26–28]

2.3.2 Sampling of Rare Event

Biomolecular processes, such as protein folding/unfolding, signaling, and ligand binding/unbinding, typically occur over timescales ranging from milliseconds to hours. However, these events are beyond the scope of unbiased all-atom molecular dynamics (MD) simulations, which currently reach microsecond timescales. To capture these rare events, various enhanced sampling techniques have been developed, including Metadynamics[29],Umbrella Sampling[30], and Steered Molecular Dynamics[31]. These methods apply a bias to the system, which encourages movement along a chosen reaction coordinate x, such as RMSD, Native Contacts, or protein-ligand distance.

Umbrella Sampling

Umbrella sampling is a method used to address the issue of sampling by adding a biased potential. This potential enhances the chances of exploring stable states that

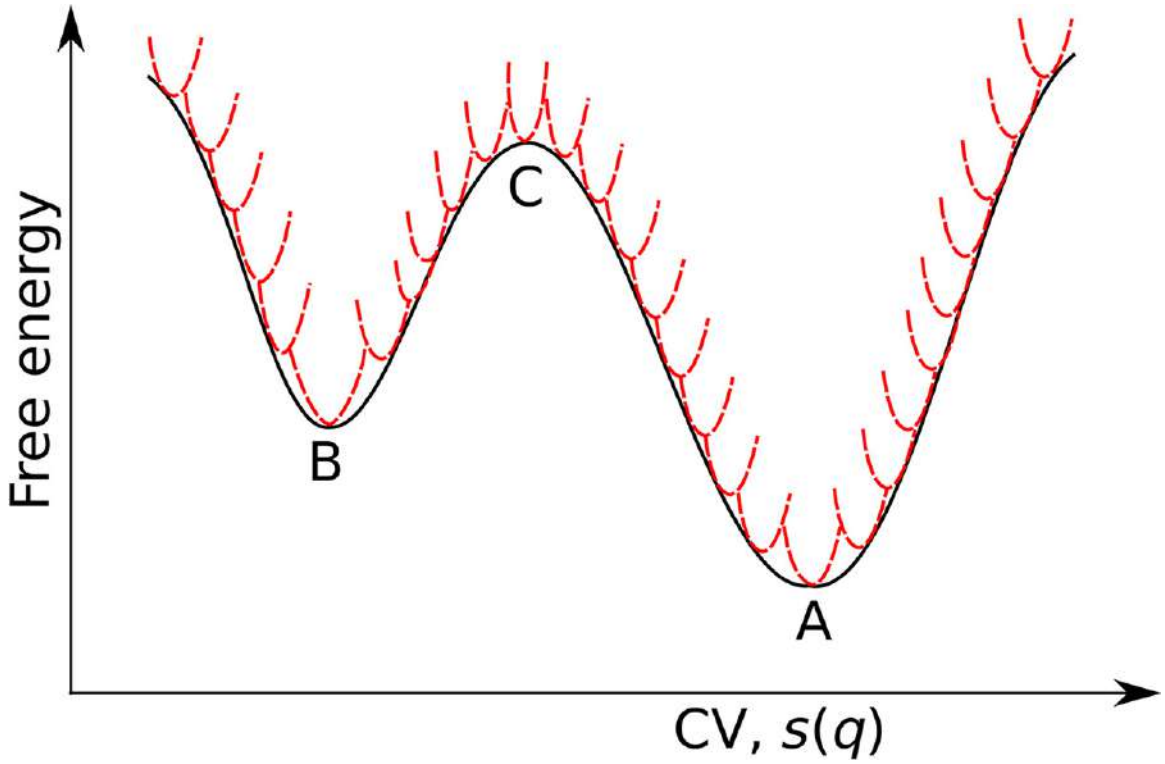

Figure 2.4: Diagram depicting the umbrella sampling technique. The red dashed lines correspond to the harmonic bias potentials that are introduced into the system Hamiltonian at various places (windows) along the collective variable (CV) space. Reprinted from "Chapter Four - Enhanced sampling and free energy calculations for protein simulations" by Qinghua Liao, 2020,Progress in Molecular Biology and Translational Science[32]

are separated by large free energy barriers[30,33]. The approach computes the Potential of Mean Force (PMF) along a selected response coordinate by including a bias, usually in the form of a harmonic potential, represented as:

$$W(X) = \frac{1}{2}k(X - X_0)^2 \tag{2.51}$$

In this context, k represents the strength of the harmonic constraint applied to the collective variable (CV), whereas X_0 denotes the equilibrium state inside each window. The harmonic constraint is utilized to ensure that the system remains in close proximity to the equilibrium state. The whole potential thus becomes:

$$U(r) = U_0(r) + W(X) \tag{2.52}$$

In order to select the equilibrium value of the bias potential, the CV is designed in such a way that it allows for the selection of a set of intermediate points between the respective end states. Sampling inside each window or umbrella will result in a probability distribution that is skewed towards each individual point. Subsequently, the mean value of quantity F for an impartial trajectory is obtained using ensemble averaging.

$$\langle F \rangle = \frac{\langle F e^{\beta U_0} \rangle_W}{\langle e^{\beta U_0} \rangle_W} \tag{2.53}$$

where $\langle \cdot \rangle_W$ indicates sampling under the biased potential. Umbrella sampling involves simulations at multiple windows, generating several PMFs along the reaction coordinate.

To combine these simulations and generate a complete PMF, the Weighted Histogram Analysis Method (WHAM) is commonly used[34,35]. The accuracy of umbrella sampling relies on sufficient overlap between windows and the proper choice of biased potential. This technique has been widely used to study large conformational changes in proteins, such as misfolding, ligand binding events, and ion transport through channels.[36–39]

Metadynamics

Metadynamics is a sophisticated sampling technique used for rebuilding the free energy surface as a function of chosen reaction coordinate (RC)[29,40]. In contrust with umbrella sampling, which applies a static bias, metadynamics dynamically accelerates the sampling of infrequent events by introducing a time-dependent repulsive bias that propels the system out of local minima[41].

At any given time t, the bias potential is the cumulative sum of repulsive Gaussians placed along the trajectory:

$$V(X,t) = \sum_{t_i<t} w \exp\left(-\frac{(X - X(t_i))^2}{2\sigma^2}\right) \tag{2.54}$$

Here, $X(t)$ represents the value of the reaction coordinate at time t, σ is the width of the Gaussian, and w is its height. This bias potential prevents the system from remaining in low free energy states for too long. To ensure the system has time to relax

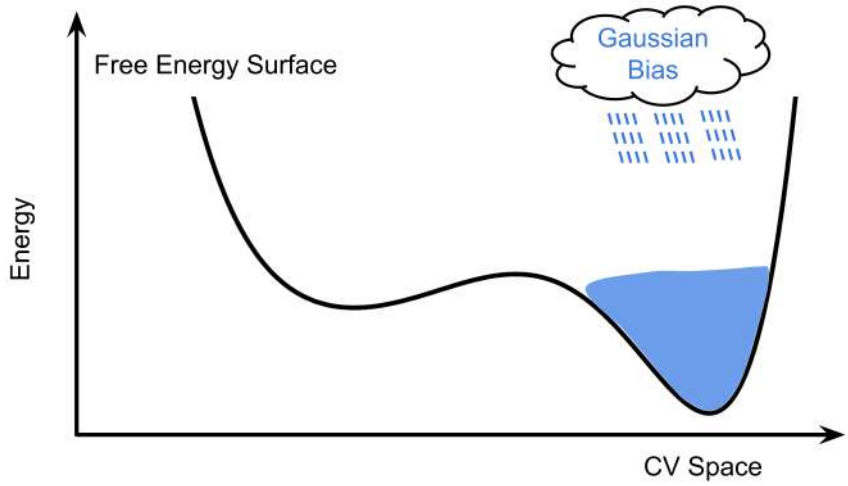

Figure 2.5: An illustrative depiction of metadynamics in the form of a cartoon. Metadynamics simulations include the deposition of gaussian form biases along specified coordinates, often known as collective variables (CVs).

to the nearest local minimum before another Gaussian is added, the interval between successive hill additions (deposition frequency) must be chosen carefully. Proper adjustment of the Gaussian parameters—width, height, and deposition frequency—is crucial for an accurate and efficient reconstruction of the free energy surface. Incorrect parameter settings can either lead to significant errors (if hills are too large or added too frequently) or increased computational cost (if hills are too small or added too infrequently).

The principle of metadynamics is that by progressively adding enough Gaussian hills over a long period, the bias potential approximates the unbiased energy landscape:

$$F(X) = -V(X, t \to \infty) + c \tag{2.55}$$

where c is a constant.

To overcome the convergence limitations of early metadynamics, well-tempered metadynamics was developed. In this method, the bias potential is defined as:

$$V(X,t) = \frac{1}{\Delta T} \ln \left(1 + \frac{\Delta T}{k_B T} N(X,t)\right) \tag{2.56}$$

where $N(X,t)$ is the histogram of the collective variable X, and ΔT is a parame-

ter controlling the extent of exploration away from the lowest energy states. This approach allows for a gradual reduction in the Gaussian deposition rate over time, thereby preventing overfilling of the free energy landscape.

The free energy surface (FES) can be estimated as:

$$F(X) = -\frac{T + \Delta T}{\Delta T} V(X, t \to \infty) + c \tag{2.57}$$

Metadynamics has proven effective in exploring reaction pathways, identifying metastable states, and reconstructing free energy landscapes without prior knowledge[42–45].

Steered Molecular Dynamics (SMD)

Steered Molecular Dynamics (SMD) was first introduced by Klaus Schulten and colleagues in 1997[46]. This approach takes inspiration from atomic force microscopy (AFM) investigations, where a mechanical probe is employed to collect force-extension data from different structures, such as biomolecules. Although AFM yielded significant insights into the structure-function correlations of several systems, it could not completely illuminate the atomic-level processes and interactions[47]. Steered molecular dynamics (SMD) simulations have demonstrated their reliability in offering comprehensive understanding of the connections between the structure and function of large molecular complexes, including protein-ligand interactions. This makes them a valuable tool that complements experimental data. In simulations using the SMD method, a specific atom is linked to a fake atom by a virtual spring. The dummy atom is displaced at a consistent speed, and the force exerted between the SMD atom and the dummy atom is quantified using the subsequent equations:

$$F = -\nabla U \tag{2.58}$$

$$U = \frac{1}{2} k \left[vt - (\mathbf{r} - \mathbf{r}_0) \cdot \mathbf{n} \right]^2 \tag{2.59}$$

The variables in the equation are as follows: U represents the potential energy, k represents the spring constant, v represents the pulling velocity, t represents time,$\mathbf{r}$

represents the current location of the SMD atom, $\mathbf{r}_0$ represents the initial position of the SMD atom, and $\mathbf{n}$ represents the direction of pulling. The work done along a trajectory can be approximated by:

$$W_{0\to t} = -kv \int_0^t (r - (r_0 + vt))\, dt \tag{2.60}$$

As per the second law of thermodynamics, the average work performed on the system is either more than or equal to the difference in free energy between the starting and end states. This can be stated as:

$$\Delta G \leqslant \overline{W} \tag{2.61}$$

The overbar is the average value obtained from a collection of measurements of WW. The principle of equality remains valid when the procedure is conducted at an indefinitely slow pace. In 1997, Christopher Jarzynski developed an equation that establishes a connection between non-equilibrium processes occurring between two states and the difference in equilibrium free energy between those states.[48]. This relationship is expressed as:

$$\langle \exp(-\beta W) \rangle = \exp(-\beta \Delta G) \tag{2.62}$$

or equivalently,

$$\Delta G = -\beta^{-1} \ln \langle \exp(-\beta W) \rangle \tag{2.63}$$

where $(\beta = 1/(k_B T))$.

2.4 Markov State Models (MSM)

Zwanzig's 1983 paper examines the random walk of a dynamical system through a sequence of states, focusing on short memory approximations.[49] He observes that if these states are selected "sensibly" and the dynamics they represent are "sufficiently complex," the system will lack memory of how it arrived at its present state. Under these conditions, Zwanzig notes, there is a "remarkably simple way to estimate

the transition rates between cells."[49] The approach for predicting transition rates has developed into the discipline of kinetic analysis, which is frequently referred to as Markov state models (MSMs). The MSM framework, when integrated with advancements in molecular dynamics (MD) simulations, has evolved into a complex and statistically rigorous collection of tools for studying the temporal development of a system. Crucially, the process of choosing states no longer depends on personal experience and intuition, due to the methodological progress made in the subject. As a result, the larger scientific community now has access to the whole capabilities of MSM approaches, which are a well established collection of procedures. A Markov State Model (MSM) provides a framework based on the master equation, enabling a comprehensive description of a system's dynamics. The master equation approach has wide-ranging applications across various scientific disciplines, as detailed in reference. In the context of MSMs, this method is employed to model dynamical systems typically assumed to be in thermodynamic equilibrium. The MSM is represented by an $n \times n$ square matrix, known as the transition probability matrix, where the system's entire configuration space is divided into n discrete states. By identifying these states, we can monitor the system's dynamical progression (e.g., in molecular dynamics simulations) by recording the state it occupies at intervals of τ, known as the lag time. For the lag time τ to be considered Markovian, the system must exhibit "memorylessness," meaning the probability of transitioning from state x to state y is independent of the system's history before entering state x.

The transition probabilities offer kinetic understanding and delineate potential routes between any two states.The need for the system to be in thermodynamic equilibrium implies that all transitions occur with equal probability in both the forward and backward directions, resulting in symmetry about an equilibrium distribution. In addition, the system must exhibit ergodicity, meaning that any state may be reached from any other state given enough time, and aperiodicity, meaning that all initial configurations of the system will ultimately converge to the same equilibrium distribution. These conditions enable additional study of the system by use of the eigendecomposition of the transition matrix. The process of decomposition results in a collection of

eigenvectors and their corresponding eigenvalues. The eigenvectors provide an approximation of the eigenfunctions of the transfer operator, which is a continuous integral operator that is approximated by the transition probability matrix. By analyzing the characteristics of the transition matrix, we may deduce certain important factors related to the eigenvalues and eigenvectors:

1. Each eigenvalue's corresponding eigenvector has n elements, each linked to one of the n states, illustrating the states' contributions to the process identified by that eigenvalue.

2. The eigenvalue with the greatest magnitude is 1, and its related eigenfunction indicates the state of equilibrium.

3. All eigenvalues, except for one, have absolute values that are smaller than 1. This suggests that the processes either decay to equilibrium (for positive eigenvalues) or fluctuate (for negative eigenvalues). However, the negative eigenvalues are usually not relevant for analytical purposes.

4. The lag time τ can be used to transform positive eigenvalues into physically significant timeframes.

The ultimate goal is to select n states that effectively capture the system's dynamics and choose a lag time τ that is sufficiently long to ensure Markovian properties while still being short enough to resolve the system's dynamics. When this is achieved, the MSM provides extensive and valuable information about the system, derived solely from its transition matrix.

2.4.1 Workflow for Building MSM

Here we present the typical work flow for building MSMs for protein systems Figure 2.6.

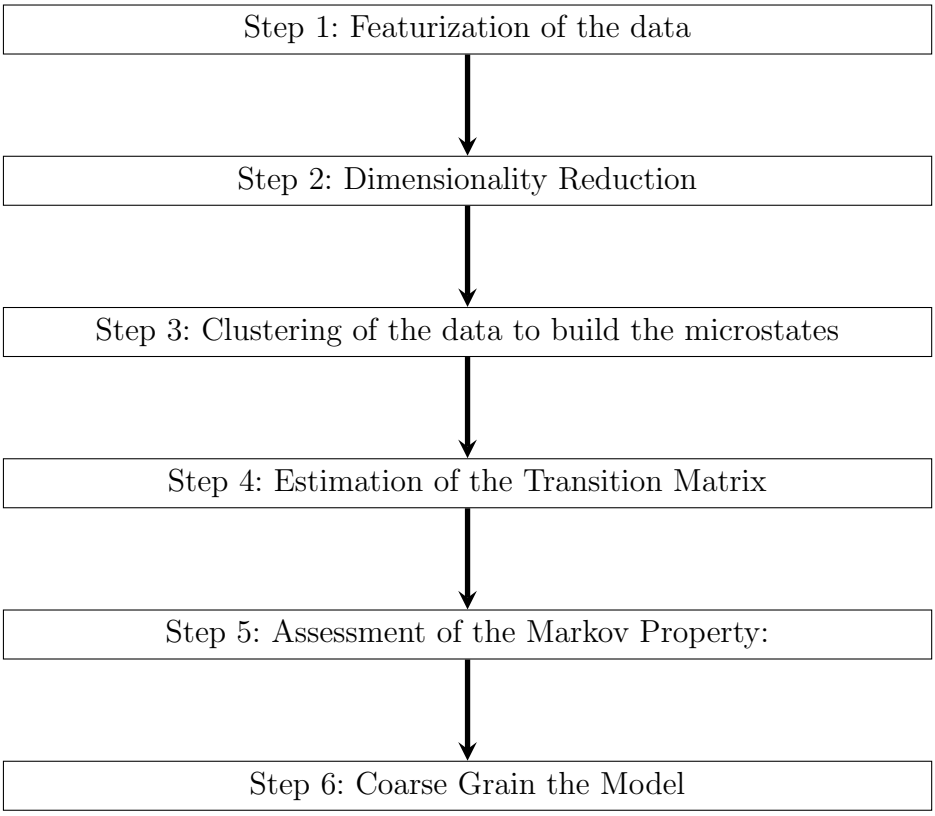

Figure 2.6: Workflow for Building Markov State Models (MSMs)

Feature Selection

Feature selection is crucial in Markov state modeling to accurately represent the essential dynamics of biomolecular processes. By choosing appropriate input coordinates or properties, such as backbone dihedral angles and tertiary connections, we can effectively capture the structural and dynamic traits of the system. These features help differentiate between structural states, elucidate energy distributions, and track slow motions vital for understanding complex biological mechanisms like protein folding. For instance, if we focus on the movement of a specific loop, we might select important Cα-Cα distances or dihedral angles as features.

Automated methods for identifying significant features include RF-tICA and MoSAIC. MoSAIC constructs a correlation matrix measuring the correlation between each pair of Cα-Cα distances. By analyzing this matrix, we can exclude movements

with weak or no correlation, refining the feature set[50]. RF-tICA uses Random Forest machine learning to identify critical residue pairs sensitive to ligand binding, aiding in the characterization of protein-ligand interactions. This supervised classification approach enhances our understanding and analysis of key features related to these interactions[51].

Dimensionality Reduction

Once features are selected, the next step is to reduce the dimensionality of the data to capture the essential dynamics while discarding noise and irrelevant information. Several methods are commonly used for this purpose, including Principal Component Analysis (PCA), Time-lagged Independent Component Analysis (TICA), dihedral PCA (dPCA), and Kernel Time-lagged Independent Component Analysis (Kernel TiCA). The basic theory of TICA and PCA is mentioned in the section 4.8

Clustering for Building Microstates:

Clustering is a crucial step in constructing a Markov State Model (MSM), as it involves grouping similar conformations or states of a molecular system into discrete microstates. Several clustering methods can be employed, each with its strengths and suitable applications:

- **k-means Clustering:** This method partitions the data into k clusters by minimizing the variance within each cluster. It is computationally efficient and straightforward but requires the number of clusters to be specified beforehand.
- **Hierarchical Clustering:** This approach builds a hierarchy of clusters by either agglomerative (bottom-up) or divisive (top-down) strategies. It is flexible and does not require a predefined number of clusters but can be computationally intensive for large datasets.
- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN):** This method identifies clusters based on the density of data points,

allowing it to find arbitrarily shaped clusters and handle noise effectively. It does not require specifying the number of clusters but depends on the choice of density parameters.

Estimation of the Transition Matrix

In theory, estimating a transition matrix involves straightforward counting. First, data is assigned to clusters, labeled from 0 to $n-1$. Each trajectory is then seen as a series of microstate assignments, allowing the construction of a transition count matrix $\mathbf{C}$, where C_{ij} represents the number of transitions from state i to state j. With infinite data, the transition probability matrix $\mathbf{T}$ can be estimated using:

$$T_{ij}(\tau) = \frac{C_{ij}}{\sum_k C_{ik}} \tag{2.64}$$

where τ is the lag time. However, finite data complicates this process due to sampling limitations and microstate definition imperfections.

Counting Transitions

Various approaches exist for counting transitions. One must choose a lag time that satisfies the Markov assumption, often requiring estimation at multiple lag times. Ideally, with abundant data, transitions at intervals of τ can be observed, counting transitions as $\sigma(0) \to \sigma(\tau)$, $\sigma(\tau) \to \sigma(2\tau)$, and so forth. In practice, a sliding window approach, sampling at intervals Δ where $\Delta < \tau$, provides more precise estimates:

$$\sigma(0) \to \sigma(\tau),\ \sigma(\Delta) \to \sigma(\Delta+\tau),\ \sigma(2\Delta) \to \sigma(2\Delta+\tau), \ldots \tag{2.65}$$

This method reduces model uncertainty but may underestimate errors.

Detailed Balance and Prior

Ensuring detailed balance, or microscopic reversibility, is crucial. Each transition from state i to j must be balanced by a transition from j to i. Poorly sampled states can

disrupt this balance. A common method to enforce detailed balance is averaging transitions:

$$\hat{C}_{ij}(\tau) = \frac{C_{ij} + C_{ji}}{2} \tag{2.66}$$

For limited data, this approach biases the transition matrix towards initial conditions. Maximum likelihood methods for reversible matrices can be powerful but often fail with limited data, emphasizing poorly sampled states.

Ergodicity

A valid MSM must be ergodic, meaning the network of states is fully connected. Disconnected components may arise from insufficient sampling. Solutions include discarding smaller components or collecting more data to ensure complete connectivity.

Assessment of the Markov Property Using Implied Timescales:

To evaluate how well the transition matrix adheres to the Markov property, we examined implied timescales (ITS) across various MSM lag times. The ITS of the propagator is defined such that:

$$t_j = -\frac{\tau}{\ln(\lambda_j)} \tag{2.67}$$

remains consistent regardless of the chosen lag time τ. According to equation (3.3), implied timescales should be independent of the lag time at which the MSM is constructed. Given that the transition matrix approximates the transfer operator, the ITS of the estimated transition matrix should follow this property.

We tested this by estimating transition matrices at different lag times τ_{MSM} and comparing their implied timescales. The Markov property is considered satisfied once the ITS as a function of τ_{MSM} converge. Therefore, the chosen MSM lag time τ_{MSM} must be sufficiently large to ensure the Markov property is met, yet as short as possible to maintain computational efficiency.

Additionally, the first and longest implied timescale is typically underestimated. This means that a larger initial timescale indicates a better MSM. Thus, an MSM with converged and substantial implied timescales will best satisfy the Markov property.

Coarse-Graining the Model:

Coarse-graining microstate models by merging rapidly mixing microstates into larger macrostates offers several benefits. It is often possible to create mesoscale models that retain the quantitative predictive power of the original microstate models while being significantly more compact. Additionally, these simplified models are easier to understand, which helps in developing intuition about the system and generating hypotheses, even if they lose some quantitative precision. The primary questions in constructing these coarse-grained models are determining which microstates to combine and deciding on the number of macrostates to form. Several methods have been developed to address these questions, including Perron Cluster Cluster Analysis (PCCA), its more robust version PCCA+, Super-level-set hierarchical clustering (SHC), and the Bayesian agglomerative clustering engine (BACE).

Bibliography

[1] Schlick, T, Collepardo-Guevara, R, Halvorsen, L. A, Jung, S, & Xiao, X. (2011) Biomolecular modeling and simulation: a field coming of age. *Quarterly Reviews of Biophysics* **44**, 191–228.

[2] Metropolis, N, Rosenbluth, A. W, Rosenbluth, M. N, Teller, A. H, & Teller, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.

[3] Alder, B. J & Wainwright, T. E. (1959) Studies in molecular dynamics. i. general method. *The Journal of Chemical Physics* **31**, 459–466.

[4] Rahman, A. (1964) Correlations in the motion of atoms in liquid argon. *Physical Review* **136**, A405–A411.

[5] McCammon, J. A, Gelin, B. R, & Karplus, M. (1977) Dynamics of folded proteins. *Nature* **267**, 585–590.

[6] Schaefer, H. F. (1986) Methylene: A paradigm for computational quantum chemistry. *Science* **231**, 1100–1107.

[7] Maddox, J. (1988) Statistical mechanics by numbers. *Nature* **334**, 561–561.

[8] Check Hayden, E. (2010) Human genome at ten: Life is complicated. *Nature* **464**, 664–667.

[9] Lindorff-Larsen, K, Piana, S, Dror, R. O, & Shaw, D. E. (2011) How fast-folding proteins fold. *Science* **334**, 517–520.

[10] Perilla, J. R & Schulten, K. (2017) Physical properties of the hiv-1 capsid from all-atom molecular dynamics simulations. *Nature Communications* **8**.

[11] Acharya, A, Agarwal, R, Baker, M. B, Baudry, J, Bhowmik, D, Boehm, S, Byler, K. G, Chen, S. Y, Coates, L, Cooper, C. J, Demerdash, O, Daidone, I, Eblen, J. D, Ellingson, S, Forli, S, Glaser, J, Gumbart, J. C, Gunnels, J, Hernandez, O, Irle, S, Kneller, D. W, Kovalevsky, A, Larkin, J, Lawrence, T. J, LeGrand, S, Liu, S.-H, Mitchell, J, Park, G, Parks, J, Pavlova, A, Petridis, L, Poole, D, Pouchard, L, Ramanathan, A, Rogers, D. M, Santos-Martins, D, Scheinberg, A, Sedova, A, Shen, Y, Smith, J. C, Smith, M. D, Soto, C, Tsaris, A, Thavappiragasam, M, Tillack, A. F, Vermaas, J. V, Vuong, V. Q, Yin, J, Yoo, S, Zahran, M, & Zanetti-Polzi, L. (2020) Supercomputer-based ensemble docking drug discovery pipeline with application to covid-19. *Journal of Chemical Information and Modeling* **60**, 5832–5852.

[12] Lindorff-Larsen, K, Piana, S, Palmo, K, Maragakis, P, Klepeis, J. L, Dror, R. O, & Shaw, D. E. (2010) Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics* **78**, 1950–1958.

[13] Maier, J. A, Martinez, C, Kasavajhala, K, Wickstrom, L, Hauser, K. E, & Simmerling, C. (2015) ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation* **11**, 3696–3713.

[14] Huang, J, Rauscher, S, Nawrocki, G, Ran, T, Feig, M, de Groot, B. L, Grubmüller, H, & MacKerell, A. D. (2016) Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73.

[15] Ewald, P. P. (1921) Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik* **369**, 253–287.

[16] Essmann, U, Perera, L, Berkowitz, M. L, Darden, T, Lee, H, & Pedersen, L. G. (1995) A smooth particle mesh ewald method. *The Journal of Chemical Physics* **103**, 8577–8593.

[17] Berendsen, H. J. C, Postma, J. P. M, van Gunsteren, W. F, DiNola, A, & Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690.

[18] Bussi, G, Donadio, D, & Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**.

[19] Andersen, H. C. (1980) Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics* **72**, 2384–2393.

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

[21] Hoover, W. G. (1985) Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **31**, 1695–1697.

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

[23] Molgedey, L & Schuster, H. G. (1994) Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **72**, 3634–3637.

[24] Pérez-Hernández, G, Paul, F, Giorgino, T, De Fabritiis, G, & Noé, F. (2013) Identification of slow molecular order parameters for markov model construction. *The Journal of Chemical Physics* **139**.

[25] Schwantes, C. R & Pande, V. S. (2013) Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. *Journal of Chemical Theory and Computation* **9**, 2000–2009.

[26] Gao, J. (1991) A priori computation of a solvent-enhanced sn2 reaction profile in water: the menshutkin reaction. *Journal of the American Chemical Society* **113**, 7796–7797.

[27] Huang, N, Banavali, N. K, & MacKerell, A. D. (2002) Protein-facilitated base flipping in dna by cytosine-5-methyltransferase. *Proceedings of the National Academy of Sciences* **100**, 68–73.

[28] Allen, T. W, Andersen, O. S, & Roux, B. (2003) Energetics of ion conduction through the gramicidin channel. *Proceedings of the National Academy of Sciences* **101**, 117–122.

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

[30] Kästner, J. (2011) Umbrella sampling. *WIREs Computational Molecular Science* **1**, 932–942.

[31] Isralewitz, B, Gao, M, & Schulten, K. (2001) Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology* **11**, 224–230.

[32] Liao, Q. (2020) *Enhanced sampling and free energy calculations for protein simulations.* (Elsevier), p. 177–213.

[33] Torrie, G & Valleau, J. (1977) Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **23**, 187–199.

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

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

[36] Zhu, F & Hummer, G. (2011) Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry* **33**, 453–465.

[37] Rajamani, R, Naidoo, K. J, & Gao, J. (2003) Implementation of an adaptive umbrella sampling method for the calculation of multidimensional potential of mean force of chemical reactions in solution. *Journal of Computational Chemistry* **24**, 1775–1781.

[38] Bartels, C & Karplus, M. (1998) Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy. *The Journal of Physical Chemistry B* **102**, 865–880.

[39] Meng, Y & Roux, B. (2015) Efficient determination of free energy landscapes in multiple dimensions from biased umbrella sampling simulations using linear regression. *Journal of Chemical Theory and Computation* **11**, 3523–3529.

[40] Barducci, A, Bonomi, M, & Parrinello, M. (2011) Metadynamics. *WIREs Computational Molecular Science* **1**, 826–843.

[41] Valsson, O, Tiwary, P, & Parrinello, M. (2016) Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annual Review of Physical Chemistry* **67**, 159–184.

[42] Kumar, P. P, Kalinichev, A. G, & Kirkpatrick, R. J. (2007) Dissociation of carbonic acid: Gas phase energetics and mechanism from ab initio metadynamics simulations. *The Journal of Chemical Physics* **126**.

[43] Petersen, L, Ardèvol, A, Rovira, C, & Reilly, P. J. (2009) Mechanism of cellulose hydrolysis by inverting gh8 endoglucanases: A qm/mm metadynamics study. *The Journal of Physical Chemistry B* **113**, 7331–7339.

[44] Alfonso-Prieto, M, Biarnés, X, Vidossich, P, & Rovira, C. (2009) The molecular mechanism of the catalase reaction. *Journal of the American Chemical Society* **131**, 11751–11761.

[45] Fiorin, G, Pastore, A, Carloni, P, & Parrinello, M. (2006) Using metadynamics to understand the mechanism of calmodulin/target recognition at atomic detail. *Biophysical Journal* **91**, 2768–2777.

[46] Izrailev, S, Stepaniants, S, Isralewitz, B, Kosztin, D, Lu, H, Molnar, F, Wriggers, W, & Schulten, K. (1999) *Steered Molecular Dynamics*. (Springer Berlin Heidelberg), p. 39–65.

[47] Do, P.-C, Lee, E. H, & Le, L. (2018) Steered molecular dynamics simulation in rational drug design. *Journal of Chemical Information and Modeling* **58**, 1473–1482.

[48] Jarzynski, C. (1997) Nonequilibrium equality for free energy differences. *Physical Review Letters* **78**, 2690–2693.

[49] Zwanzig, R. (1983) From classical dynamics to continuous time random walks. *Journal of Statistical Physics* **30**, 255–262.

[50] Diez, G, Nagel, D, & Stock, G. (2022) Correlation-based feature selection to identify functional dynamics in proteins. *Journal of Chemical Theory and Computation* **18**, 5079–5088.

[51] Ahalawat, N, Sahil, M, & Mondal, J. (2023) Resolving protein conformational plasticity and substrate binding via machine learning. *Journal of Chemical Theory and Computation* **19**, 2644–2657.

3

Role of Low Populated Metastable Protein Conformations in Regulation of Biological Function

"Nothing in life is to be feared; it is only to be understood."

Marie Curie

3.1 Introduction

Toxin-antitoxin (TA) modules are genetic elements widely present in bacterial genomes[1]. These modules assist bacteria in surviving various stress conditions, including those induced by antibiotics.The HipBA system, along with its homologs, is an example of type II TA modules found in numerous bacteria and archaea[2]. In this system, the antitoxin HipB neutralizes the toxin HipA and also binds to four operator

regions upstream, acting as a regulator of the HipBA operon[3]. Under stressful conditions, HipA dissociates from the HipBA complex and phosphorylates a conserved serine residue in glutamyl-tRNA synthetase (GluRS)[4,5]. This phosphorylation event renders GluRS unable to charge $tRNA^{Glu}$,leading to the accumulation of uncharged $tRNA^{Glu}$ in the cell. This accumulation activates RelA to synthesize (p)ppGpp, triggering a stringent response that results in dormancy and persistence. When stress conditions subside, the toxin HipA needs to be inactivated, which occurs either through its binding with HipB or via autophosphorylation[6].

In *Escherichia coli* (E. coli), the autophosphorylation site of HipA is Ser150, located within an ATP-binding "P-loop" (151VAGAQEKT; Figure 6.5), which is typically buried inside the protein core[7]. Upon phosphorylation, the P-loop is expelled from the core [Figure 6.5 (a)], positioning the phosphate group of pSer150 to partially occupy the γ-phosphate site of ATP in the HipA structure. This positions the side chain of pSer150 to interact with Asp309, the proposed catalytic base[8]. Thus, HipA autophosphorylation acts as an intramolecular "pseudo-product" inhibitor. A similar P-loop expulsion is observed in the phosphorylated state of HipA from Shewanella oneidensis, where Ser147 undergoes phosphorylation[9].

For E. coli HipA to undergo phosphorylation, the Ser150-containing P-loop must transition from an "in-state" to an "out-state" [Figure 6.5 (b)]. This transition implies that the equilibrium between these states, although heavily favoring the "in-state," allows for some degree of "out-state" presence. Phosphorylation occurs exclusively in the "out-state" (where Ser150 is exposed), thus shifting the equilibrium towards this state. However, experimental or computational evidence for the existence of an "out-state" in unphosphorylated HipA is currently lacking.

The identification of minor conformations in equilibrium with a protein's native state can be indirectly achieved by lowering the free energy of these minor states through the use of chaotropic agents like urea and extrapolating back to no-urea conditions using linear free energy relationships[10]. Such minor conformations might include stable intermediates, such as the molten globule[11], which exhibit slight tertiary structural differences but significantly looser tertiary/quaternary contacts , similar to the pro-

posed HipA "out-state".

Computational techniques are crucial for identifying metastable states of HipA. While unbiased molecular dynamics (MD) simulations effectively explore structural fluctuations around a protein's native state, metastable states located far from the native conformation can be identified using appropriate sampling methods. This can be achieved by employing biased MD simulation techniques, such as steered MD[12].

In this study, we identify the metastable states of HipA using computational methods. Through multiple MD simulations, we pinpoint intermediates along the P-loop ejection pathway of HipA that are solvent-exposed and phosphorylation-compliant. The discovery of a phosphorylation-competent metastable state suggests a mechanism for autophosphorylation: the metastable state of HipA, characterized by an ejected P-loop with solvent-exposed S150, is the phosphorylation target. Phosphorylation reduces the population of this metastable state, thereby shifting the equilibrium (natively folded state ↔ metastable state) towards the right, ultimately leading to the phosphorylation of the entire natively folded HipA population. This model aligns with mechanisms proposed in recent studies on the phosphorylation of buried amino acids[13–15]. The results presented in this chapter, were previously published in *Pandey et al, (2023), Biochemistry, 62(5), 989*

3.2 Computational Details

System Preparation

We utilized the crystal structures of wild-type (WT) and serine-phosphorylated (SP) HipA proteins (PDB ID: 3DNU and 3TPE, respectively) to perform our modeling and simulation studies. The primary distinction between these two conformations lies in the position of the P-loop: in the WT structure, the P-loop is buried within the protein core, whereas in the SP structure, the loop is exposed outwardly [see Figure 6.5 (a)]. To investigate whether the exposed loop conformation can also occur in the WT, we mutated the phosphorylated serine (pS150) back to a normal serine

residue in the phosphorylated HipA structure. Consequently, we prepared the WT system in two extreme conformations: the buried P-loop from the WT structure and the exposed P-loop from the phosphorylated structure. These conformations are referred to as Conf-WT and Conf-SP, respectively. Missing residues in the crystal structures were modeled using Pdbfixer software[16]. Additionally, the Asp309 residue in the WT structure (PDB ID: 3DNU) was originally mutated to Gln and was reverted to Asp, while in the SP structure (PDB ID: 3TPE), Arg243 was modified to Glu.

Steered MD Simulation

We employed the Steered MD (SMD) technique to generate intermediate structures between Conf-WT and Conf-SP. The root mean square deviation (RMSD) relative to the target structure was used as the pulling coordinate for SMD. The pulling force constant was varied between 15,000 and 25,000 kJ mol^{-1} nm^{-2} depending on the conformational space region being sampled. SMD simulations were conducted using the PLUMED code (version 2.0.2)[17] implemented in AMBER[18] .

Unbiased MD Simulation

Representative configurations were extracted from the SMD path connecting Conf-WT and Conf-SP. Subsequently, 25 unbiased MD trajectories (each 500 ns) were generated around this pathway, resulting in a cumulative 12.5 μs of trajectories covering the entire conformational space. MD simulations were carried out using the AMBER package[18] with the amber14sb force field[19] . All structures were solvated using the TIP3P water model[20] and simulated under periodic boundary conditions. Energy minimization was performed using the steepest descent algorithm. The simulations were run in the NPT ensemble using a modified Berendsen thermostat[21] at 300 K and the Parrinello–Rahman barostat[22] at 1 bar pressure. Long-range electrostatic interactions were calculated with the particle mesh Ewald summation method[23] , employing a grid spacing of 0.16 nm and fourth-order cubic interpolation. A cutoff distance of 1 nm was used for short-range electrostatic and van der Waals interactions,

with an integration time step set to 2 fs.

3.3 Results and Discussion

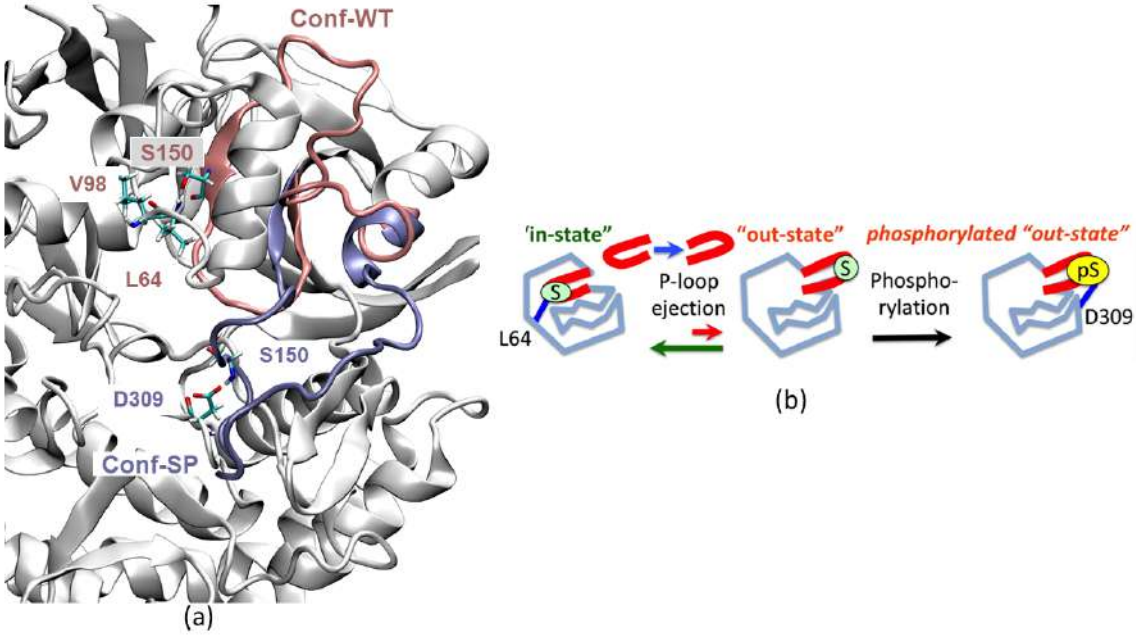

Figure 3.1: (a) Superimposed loop conformations in HipA with Ser150 in an unphosphorylated state (pink; Conf-WT; PDB ID: 3DNU) and a phosphorylated state (blue; Conf-SP; PDB ID: 3TPE). In Conf-WT, S150 establishes hydrogen bonds with L64 and V98, while in Conf-SP, S150 forms a hydrogen bond with D309. (b) Schematic diagram of our hypothesis.

3.3.1 MD Simulations of P-Loop Buried and Exposed HipA

In Figure 6.5 (a), the crystal structure depicts the wild-type (WT) configuration (pink) and the phosphorylated state with S150 (Serine 150) phosphorylation (blue). In the WT structure, the S150 residue is buried within the protein core, while in the phosphorylated system, the P-loop is exposed. This raises a fundamental question: how does the S150 residue undergo phosphorylation when it is situated in a buried state? Our proposed hypothesis, illustrated in Figure 6.5 (b), posits the existence of

at least one exposed conformation of the P-loop that is amenable to phosphorylation. We prepared the WT system in both WT and SP conformation, which we call Conf-WT and Conf-SP respectively. Then we have performed 2 microsecond long normal md simulation both from Conf-WT and Conf-SP. Despite the extensive simulations, no significant conformational changes were observed, suggesting that both states represent distinct minima separated by a large energy barrier. Subsequently, biased MD simulations were performed along the P-loop ejection pathway.

3.3.2 MD Simulations Along the P-Loop Ejection Pathway

To characterize the transition pathway between Conf-WT and Conf-SP, we performed Steered Molecular Dynamics (SMD) using rmsd values relative to target structures as the pulling coordinate [Figure 6.19 (a)]. The initial and final structures of the SMD run matched well with the corresponding crystal structures [Figure 6.19 (b)]. Since SMD primarily involves linear pulling, it does not necessarily represent the minimum free energy pathway connecting the two states.

We extracted several intermediate conformations from the SMD trajectory at intervals and used them as starting points for 25 independent unbiased MD simulations, each 500 ns long, resulting in a cumulative trajectory of 12.5 μs. This extensive trajectory covers a significant region of conformational space [Figure 6.19 (c)], with initial configurations marked by cyan dots and the overall spread shown by black dots in the conformational space defined by RMSD-WT and RMSD-SP.

3.3.3 Principal Component Analysis (PCA)

Although rmsd was used as the pulling coordinate to generate intermediate loop conformations, it may not be the most suitable metric for characterizing conformational states due to its inherent degeneracy at large values. PCA, which reduces the dimensionality of the conformational space and yields collective variables called principal components (PCs), offers a more realistic reaction coordinate. We performed PCA on the protein's backbone atoms, excluding overall translational and rotational motion,

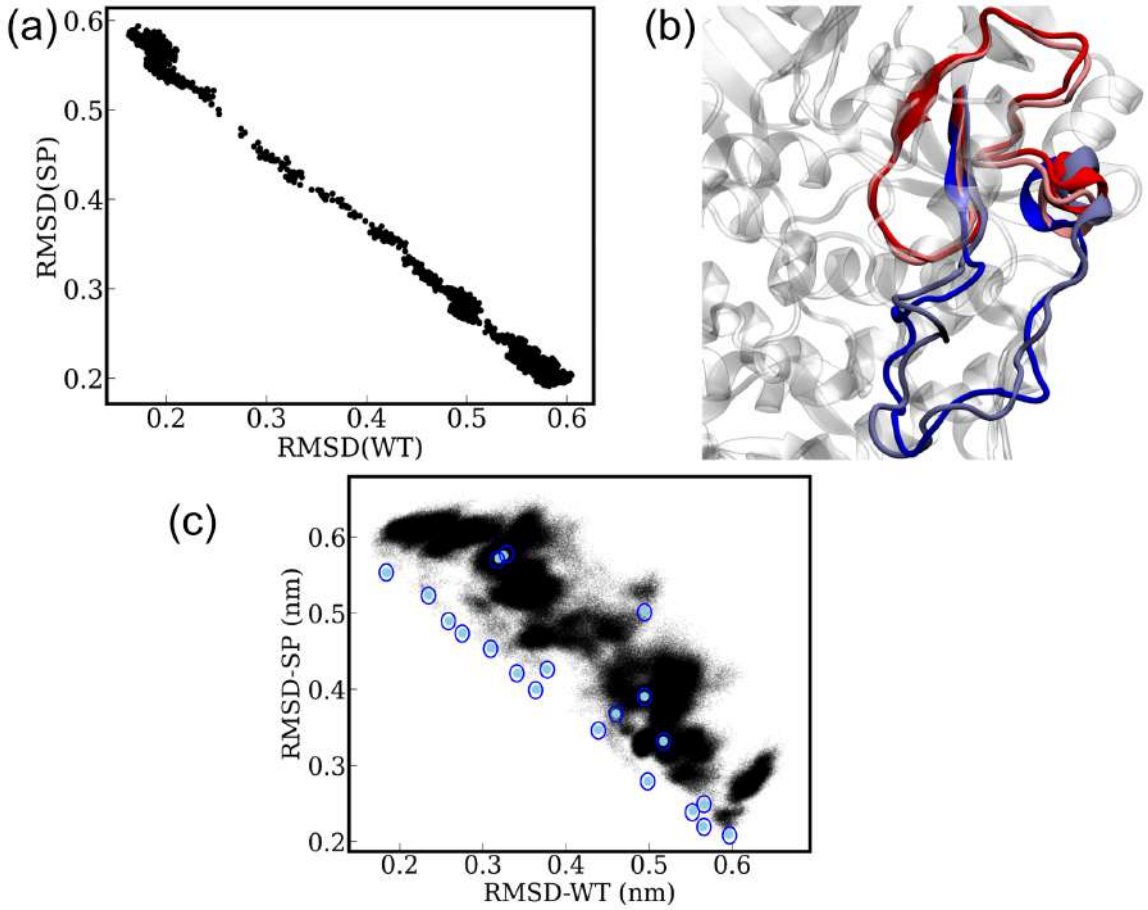

Figure 3.2: (a) Conformations generated along the SMD pathway from Conf-WT to Conf-SP, projected on the 2D space defined by RMSD relative to the WT structure (X-axis) and the SP structure (Y-axis). (b) Superposition of protein structures before and after the SMD run: original crystal structures (post-modeling) are shown in red (WT) and blue (SP), while the initial and final structures from the SMD run are depicted in pink and sky blue, respectively. (c) Conformational sampling in the combined unbiased MD runs initiated from the configurations obtained from the SMD trajectory: cyan dots represent the starting structures for the individual unbiased simulations.

using the cumulative 12.5 μs trajectory.

The first two principal components (PC1 and PC2), accounting for 48.6% and 7.4% of total mean square fluctuation respectively, describe the conformational space explored by the P-loop. Residue-wise RMSF for PC1 and PC2 [Figure 4.8(a)] indicates that both components peak around the P-loop region, with PC1 peaking at residue

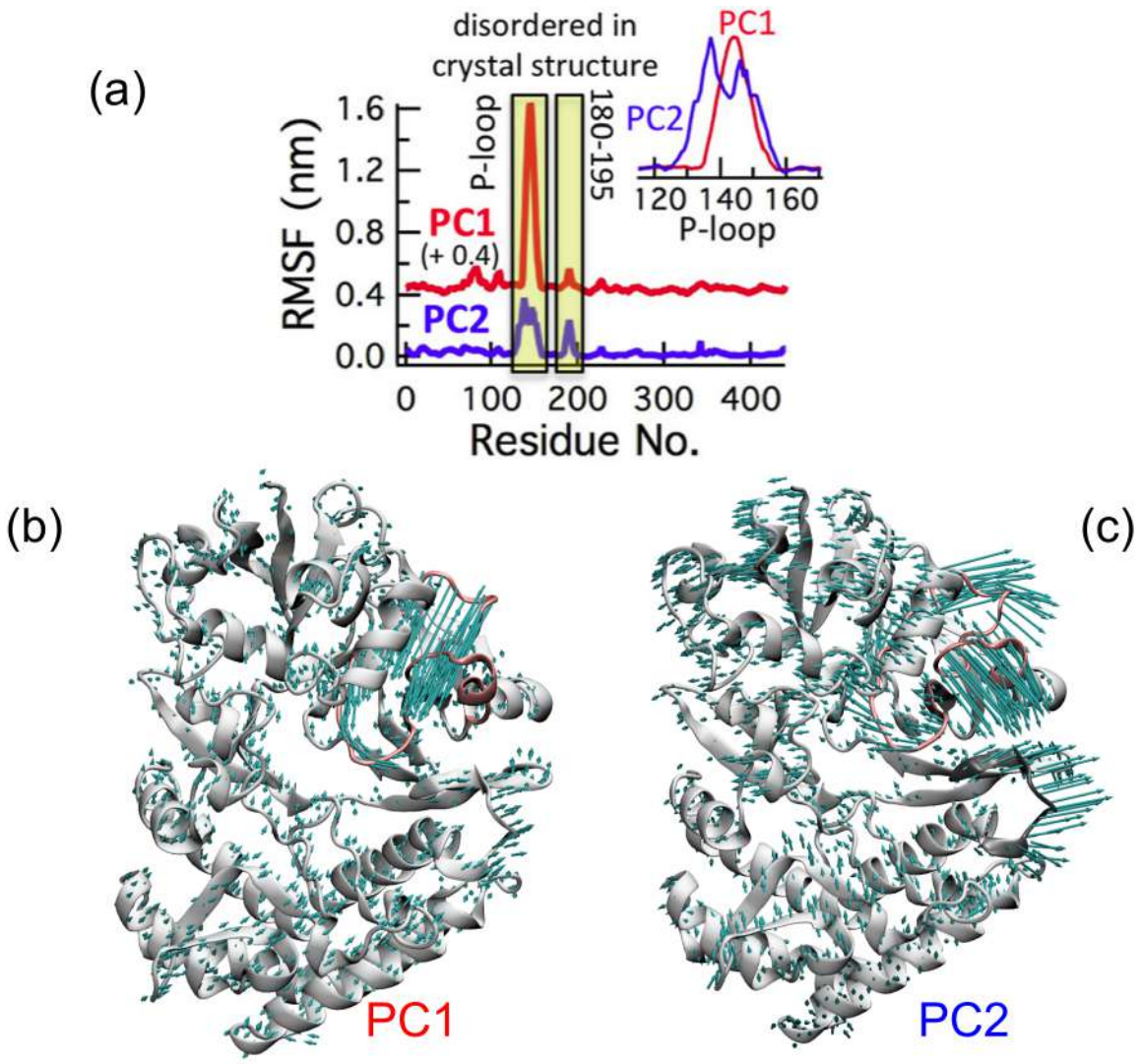

Figure 3.3: (a) Residue-wise root mean square fluctuations (RMSF) corresponding to the first two principal components (PC1 and PC2). The RMSF for PC1 is offset by 0.4 nm for clarity. (b) Porcupine plot visualizing the eigenvectors corresponding to PC1. (c) Porcupine plot visualizing the eigenvectors corresponding to PC2.

144 and PC2 peaking at residues 137 and 146. This suggests that the structural variation is concentrated around the loop. Visualizing the eigenvectors for PC1 and PC2 with "porcupine plots" [Figure 4.8 (b,c)], we observe that PC1 corresponds to a downward movement of the P-loop, while PC2 reflects a sideways movement of the loop and the 180-195 region.

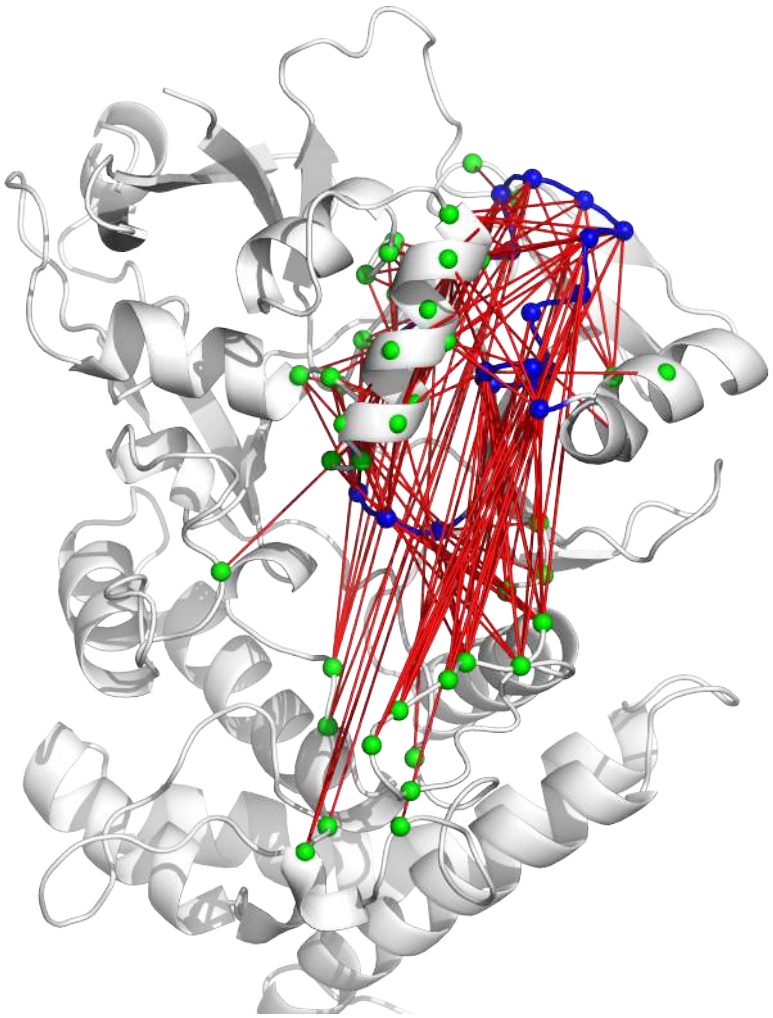

Figure 3.4: Set of C_α-C_α distances used as features. Red lines indicate the selected C_α-C_α distances. The P-loop is shown in blue. Blue spheres represent residues within the P-loop, while green spheres represent residues in the rest of the protein.

3.3.4 Free Energy Surface from a Markov State Model

Using the PyEMMA software, we performed MSM analysis to compute the free energy landscape projected on the principal components. We identified contacts between the P-loop and the rest of the protein based on Cα atom distances within 6Å, which formed the feature set for building the MSM (Figure 6.7). PCA reduced the dimensionality of these distances, projecting on the top two principal components [Figure 3.5

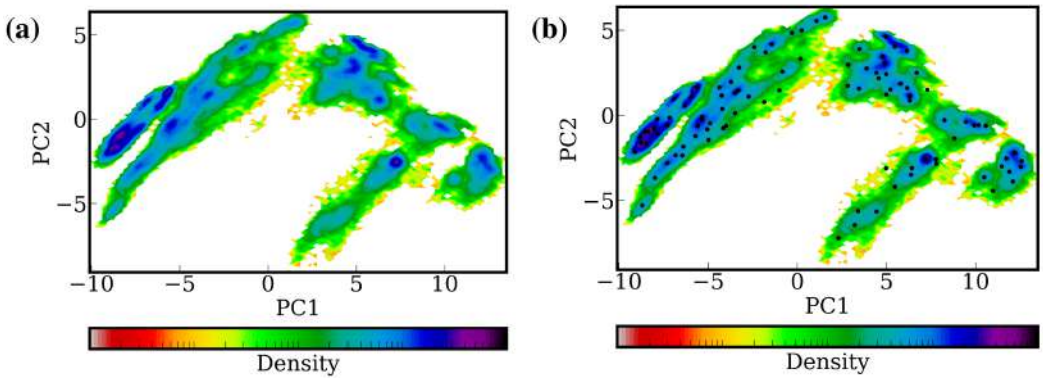

Figure 3.5: (a) PCA probability density plot. (b) K-means cluster centers displayed on the PC1-PC2 conformational space.

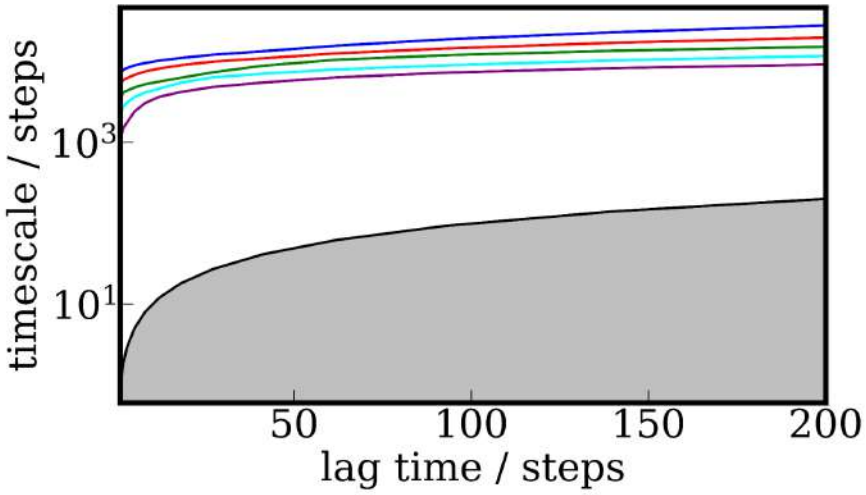

Figure 3.6: Lag Time selection for MSM. Implied relaxation time scales for the first ten eigenvalues calculated from the transition matrix at different lag times. All relaxation time scales become approximately constant beyond step 50. A lag time of 50 steps (i.e. 5 ns) was used to build the MSM.

(a)]. K-means clustering with 100 centers [Figure 3.5(b)] and implied timescale analysis (Figure 3.6) indicated that the system behaved in a Markovian manner after 5 ns (50 steps). We used the Perron-cluster cluster analysis (PCCA++) method to group microstates into four macrostates [Figure 3.7 (a)]. The free energy surface (FES) was derived from the stationary probability distribution of each macrostate.The MSM was validated using the CK-test shown in Figure 3.7 (b).

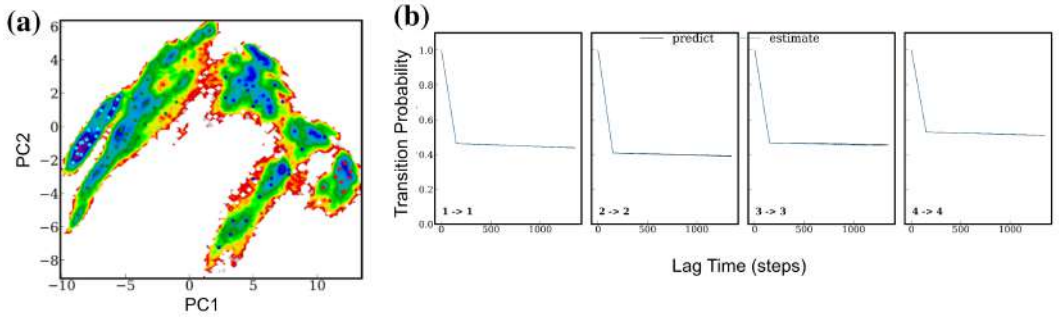

Figure 3.7: The microstates were grouped by the PCCA++ method into four metastable states, with macrostates colored as 1 (red), 2 (blue), 3 (green), and 4 (cyan). Validation of the MSM was performed using the Chapman-Kolmogorov (CK) test, comparing MSM predictions (black line) with observed trajectories (blue dotted line).

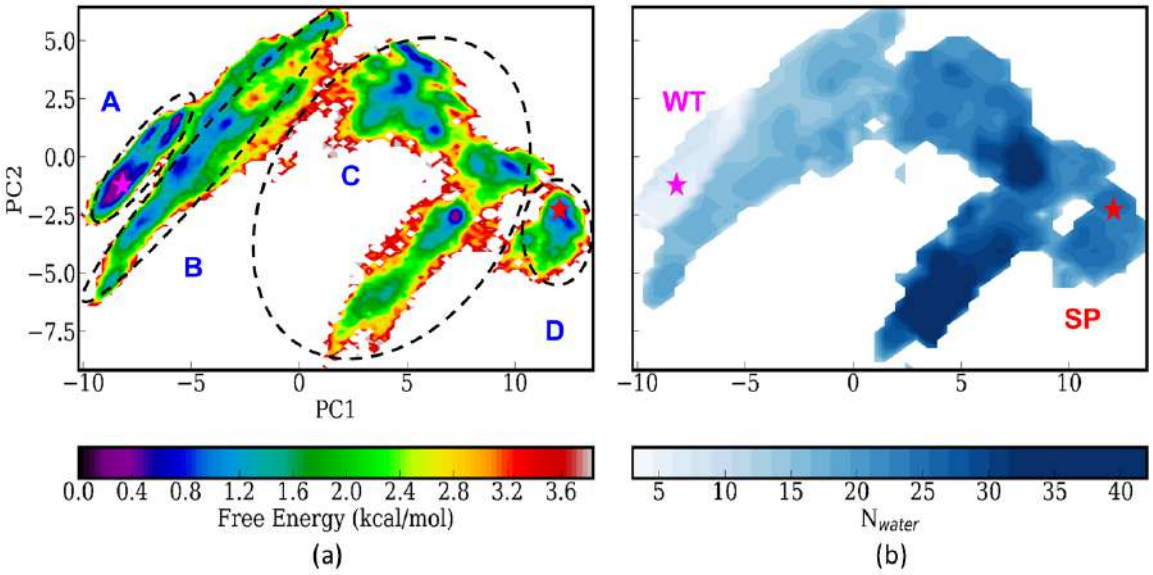

Figure 3.8: (a) Free energy surface (FES) of the P-loop conformation projected onto the PC1 and PC2 space, computed using the set of loop-protein Cα–Cα distances utilized in building the MSM. (b) Solvent exposure of the S150 residue visualized on the same FES (panel A), with the color bar indicating the average number of water molecules within a 0.6 nm distance of S150. Pink and red stars mark the positions of Conf-WT and Conf-SP, respectively.

The FES on the (PC1, PC2) space [Figure 6.13 (a)] reveals multiple minima along the pathway connecting the two end states. The conformational change primarily

Residue Pairs	A (%)	B (%)	C (%)	D (%)
L64-S150	64.69	0.0	0.0	0
E92-R148	43.31	0.0	0.0	0
R78-M140	37.72	0.0	0.0	0
S150-K157	0.0	31.67	0.0	0
E127-R148	0.0	47.11	0.0	0
R148-T158	0.0	28.16	0.0	0
K157-D332	0.0	34.15	0.0	27.15
R74-E156	0.0	0.0	56.89	66.47
R148-D270	0.0	0.0	73.65	0
R78-D135	0.0	0.0	29.34	0
D135-R188	0.0	0.0	0.0	43.31
S150-D309	0.0	0.0	0.0	35.53

Table 3.1: Average Hydrogen Bond Occupancy (Percentage) for clusters (A-D) (as shown in Figure 3.9)

occurs along PC1 in a monotonous manner, with higher PC1 values indicating a more ejected P-loop. In contrast, the pathway involving PC2 is non-monotonous, initially increasing and then decreasing back to the starting value along the P-loop ejection pathway. Thus, PC1 captures the P-loop ejection coordinate, while PC2 acts as a trigger or bottleneck for this process, similar to PC1/PC2 roles in βT7 loop dynamics in tubulin.

Four major clusters of minima (or metastable states) on the FES are marked as A, B, C, and D in Figure 6.13 (a). Cluster A represents the global minimum corresponding to the native buried conformation of the P-loop (Conf-WT), with other metastable states being more solvent-exposed. To characterize these states in terms of solvent exposure, we computed the average number of water molecules around the S150 residue, which needs to be solvent-exposed for phosphorylation. Figure 6.13 (b) shows the correlation between structural transition and solvent exposure along the pathway. Cluster C shows the highest solvation of S150, followed by Cluster D.

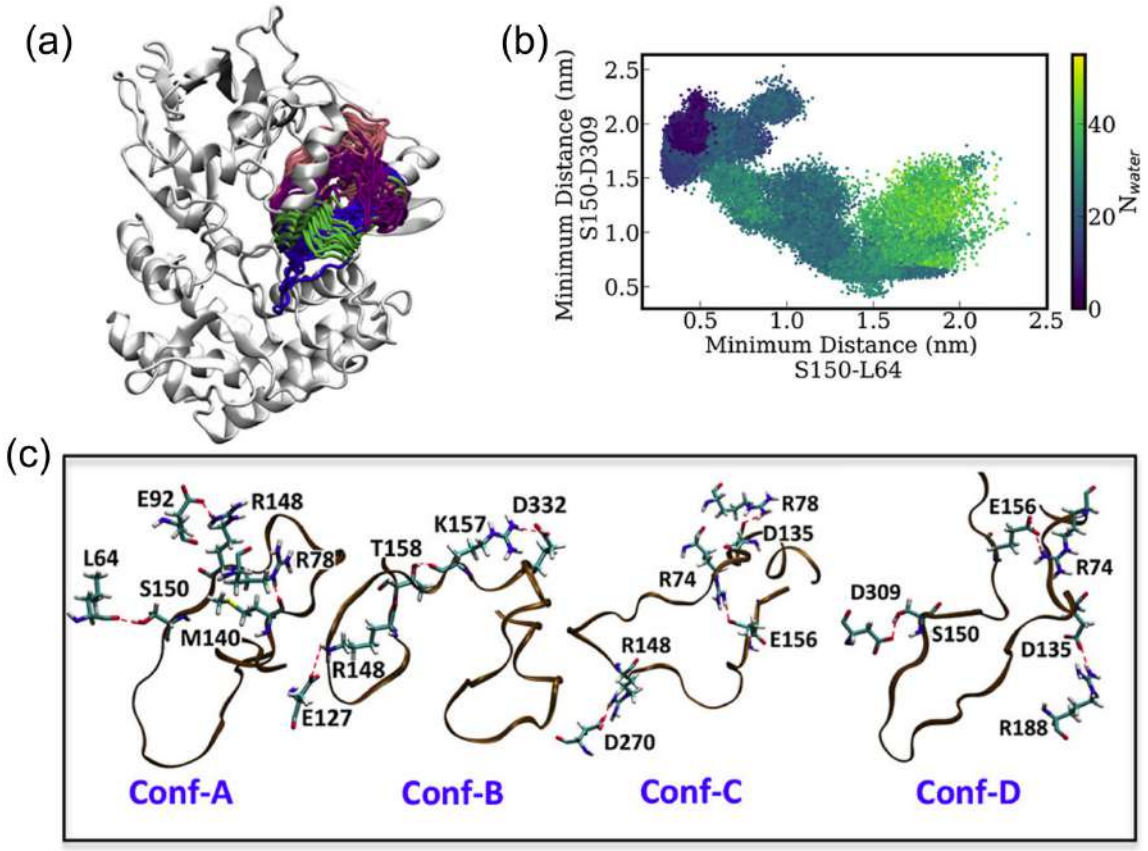

Figure 3.9: Loop conformation in the four clusters from panel (a): cluster-A (pink), cluster-B (purple), cluster-C (lime), and cluster-D (blue). (b) Specific contact patterns in representative loop conformations from clusters A–D.

3.3.5 Specific Interactions Along the P-Loop Ejection Pathway

Figure 3.9 (a) illustrates the structural ensemble of the P-loop in various metastable states (clusters A-D), highlighting the diverse loop structures depending on the state. To understand the molecular stability of these conformations, we analyzed hydrogen bond occupancy between P-loop residues and the rest of the protein. Distinct sets of contacts, consisting of hydrogen bonds and salt bridges, stabilize each metastable loop conformation [Figure 3.9 (b) and Table 3.1]. These interactions include L64-

S150, E92-R148, and R78-M140 in cluster A; S150-K157, E127-R148, R148-T158, and K157-D332 in cluster B; R74-E156, R148-D270, and R78-D135 in cluster C; and K157-D332, R74-E156, D135-R188, and S150-D309 in cluster D. These results elucidate the specific interactions necessary for the transition from Conf-WT to Conf-SP across metastable states.

Crystal structures indicate that the P-loop buried state (Conf-WT) is stabilized by a hydrogen bond between Ser150 and Leu64. In the phosphorylated state with the P-loop exposed (Conf-SP), Ser150 forms a hydrogen bond with Asp309. We projected the combined trajectory on the conformational space defined by the minimum distances between (i) Ser150 and Leu64, and (ii) Ser150 and Asp309. This projection [Figure 3.9 (c)] clearly shows the heterogeneity and spread in the conformational space during the transition from Conf-WT to Conf-SP, with high-density regions indicating locally metastable states.

Our computational analyses confirm the presence of several low-lying metastable conformational states of the P-loop, which are significantly solvent-exposed and phosphorylation-competent. We elucidated the molecular interactions stabilizing these states, consistent with experimental findings that the WT system can exist in one or more metastable, solvent-exposed phosphorylation-competent states.

3.4 Summary

In summary, our computational analyses have unveiled multiple low-lying metastable states of the P-loop in the HipA protein. These states exhibit significant solvent exposure and are competent for phosphorylation. By examining the molecular details of specific contacts and interactions, we gained insights into how these metastable states are stabilized. These findings are in agreement with experimental observations, reinforcing the existence of one or more metastable, solvent-exposed, phosphorylation-competent states in the wild-type (WT) system.

Our results suggest that transient surface exposure could be a generalized mechanism for the phosphorylation of buried residues in proteins. This mechanism not

only enhances our understanding of the phosphorylation dynamics of HipA but also contributes to the growing body of evidence supporting the prevalence of such mechanisms across diverse proteins.

By elucidating the intricate details of P-loop dynamics and the transient exposure of Ser150, our study bridges the gap between static crystal structures and dynamic protein behavior. This work establishes a foundation for further exploration of phosphorylation mechanisms within the context of protein dynamics, providing a critical link to understanding how transient conformational states can facilitate regulatory processes in proteins.

Bibliography

[1] Page, R & Peti, W. (2016) Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nature Chemical Biology* **12**, 208–214.

[2] Gerdes, K, Bærentsen, R, & Brodersen, D. E. (2021) Phylogeny reveals novel hipa-homologous kinase families and toxin-antitoxin gene organizations. *mBio* **12**.

[3] Black, D. S, Kelly, A. J, Mardis, M. J, & Moyed, H. S. (1991) Structure and organization of hip, an operon that affects lethality due to inhibition of peptidoglycan or dna synthesis. *Journal of Bacteriology* **173**, 5732–5739.

[4] Kaspy, I, Rotem, E, Weiss, N, Ronin, I, Balaban, N. Q, & Glaser, G. (2013) Hipa-mediated antibiotic persistence via phosphorylation of the glutamyl-trna-synthetase. *Nature Communications* **4**.

[5] Germain, E, Castro-Roa, D, Zenkin, N, & Gerdes, K. (2013) Molecular mechanism of bacterial persistence by hipa. *Molecular Cell* **52**, 248–254.

[6] Correia, F. F, D'Onofrio, A, Rejtar, T, Li, L, Karger, B. L, Makarova, K, Koonin, E. V, & Lewis, K. (2006) Kinase activity of overexpressed hipa is required for growth arrest and multidrug tolerance in escherichia coli. *Journal of Bacteriology* **188**, 8360–8367.

[7] Schumacher, M. A, Piro, K. M, Xu, W, Hansen, S, Lewis, K, & Brennan, R. G. (2009) Molecular mechanisms of hipa-mediated multidrug tolerance and its neutralization by hipb. *Science* **323**, 396–401.

[8] Schumacher, M, Min, J, Link, T, Guan, Z, Xu, W, Ahn, Y.-H, Soderblom, E, Kurie, J, Evdokimov, A, Moseley, M, Lewis, K, & Brennan, R. (2012) Role of unusual p loop ejection and autophosphorylation in hipa-mediated persistence and multidrug tolerance. *Cell Reports* **2**, 518–525.

[9] Wen, Y, Sobott, F, & Devreese, B. (2016) Atp and autophosphorylation driven conformational changes of hipa kinase revealed by ion mobility and crosslinking mass spectrometry. *Analytical and Bioanalytical Chemistry* **408**, 5925–5933.

[10] Greenfield, N. J. (2006) Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism. *Nature Protocols* **1**, 2733–2741.

[11] Judy, E & Kishore, N. (2019) A look back at the molten globule state of proteins: thermodynamic aspects. *Biophysical Reviews* **11**, 365–375.

[12] Do, P.-C, Lee, E. H, & Le, L. (2018) Steered molecular dynamics simulation in rational drug design. *Journal of Chemical Information and Modeling* **58**, 1473–1482.

[13] Henriques, J & Lindorff-Larsen, K. (2020) Protein dynamics enables phosphorylation of buried residues in cdk2/cyclin-a-bound p27. *Biophysical Journal* **119**, 2010–2018.

[14] Orioli, S, Henning Hansen, C. G, & Lindorff-Larsen, K. (2022) Transient exposure of a buried phosphorylation site in an autoinhibited protein. *Biophysical Journal* **121**, 91–101.

[15] Golla, H, Kannan, A, Gopi, S, Murugan, S, Perumalsamy, L. R, & Naganathan, A. N. (2022) Structural–energetic basis for coupling between equilibrium fluctuations and phosphorylation in a protein native ensemble. *ACS Central Science* **8**, 282–293.

[16] Eastman, P, Swails, J, Chodera, J. D, McGibbon, R. T, Zhao, Y, Beauchamp, K. A, Wang, L.-P, Simmonett, A. C, Harrigan, M. P, Stern, C. D, Wiewiora, R. P, Brooks, B. R, & Pande, V. S. (2017) Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **13**, e1005659.

[17] Bonomi, M, Branduardi, D, Bussi, G, Camilloni, C, Provasi, D, Raiteri, P, Donadio, D, Marinelli, F, Pietrucci, F, Broglia, R. A, & Parrinello, M. (2009) Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **180**, 1961–1972.

[18] Case, D. A, Cheatham, T. E, Darden, T, Gohlke, H, Luo, R, Merz, K. M, Onufriev, A, Simmerling, C, Wang, B, & Woods, R. J. (2005) The amber biomolecular simulation programs. *Journal of Computational Chemistry* **26**, 1668–1688.

[19] Maier, J. A, Martinez, C, Kasavajhala, K, Wickstrom, L, Hauser, K. E, & Simmerling, C. (2015) ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation* **11**, 3696–3713.

[20] Jorgensen, W. L. (1981) Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *Journal of the American Chemical Society* **103**, 335–340.

[21] Berendsen, H. J. C, Postma, J. P. M, van Gunsteren, W. F, DiNola, A, & Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690.

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

[23] Essmann, U, Perera, L, Berkowitz, M. L, Darden, T, Lee, H, & Pedersen, L. G. (1995) A smooth particle mesh ewald method. *The Journal of Chemical Physics* **103**, 8577–8593.

4

Molecular Mechanism of Dissociation of RhoA GTPase by Phosphorylation of Ser34 of RhoGDI

"The best way to have a good idea is to have lots of ideas."

Linus Pauling

4.1 Introduction

Rho GTPases, including RhoA, Rac1, and Cdc42, are pivotal members of the Ras superfamily of signaling G-proteins, crucially involved in various cellular processes such as actin dynamics, cell adhesion, and gene transcription[1–5]. These proteins share structural features in their G-domains, notably the switch I and switch II regions facilitating nucleotide exchange[6–8]. Functioning as molecular switches, they regulate

cellular activities by toggling between an active GTP-bound state and an inactive GDP-bound state[8,9]. Upon activation, Rho GTPases interact with downstream effectors, orchestrating processes like cytoskeletal rearrangement and gene expression. The intricate interplay between effectors and Rho GTPases has been elucidated through structural and mutational analyses[10,11].

The Rho family GTPase signaling cycle is tightly regulated by three major proteins: GTPase activating protein, guanine nucleotide exchange factor (GEF), and Rho-specific guanine nucleotide dissociation inhibitor (RhoGDI)[10]. GTPase activating protein promotes GTP hydrolysis, terminating the signaling cascade, while GEF activates monomeric Rho GTPases by facilitating GDP release and GTP binding[11–15]. RhoGDI, on the other hand, maintains Rho GTPases in an inactive state by isolating them from the membrane, where they exert their functions.

RhoGDIs exert critical control over Rho family GTPases[16], acting as "invisible hand" modulating their activation state within the cell[17]. They form complexes with prenylated Rho GTPases, regulating their cytoplasmic pool[18–20]. Structural studies have delineated two key interaction regions between RhoA and RhoGDI[21–23]:(a) the N-terminal region of RhoGDI, adopting a helix-turn-helix motif, interacts with switch I and II regions of RhoA, inhibiting GDP dissociation; and (b) the C-terminal region, folding into a β sandwich structure, interacts with the prenyl moiety of RhoA.

PTM Site	Kinase	Effect
Ser34	PKCα	Promotes Dissociation of RhoA
Ser96	PKCα	Promotes Dissociation of RhoA
Ser101	PKA1	Promotes Dissociation of Rac1
Ser174	PKA1	Promotes Dissociation of Rac1
Tyr156	FER	Promotes Dissociation of Rac1

Table 4.1: Summary of the phosphorylation code based on available experimental data, indicating the phosphorylation or post-translational modification (PTM) site and their respective effects on specific Rho GTPases

Activation of RhoGTPases requrir release of RhoGTPase from the complex[24] and

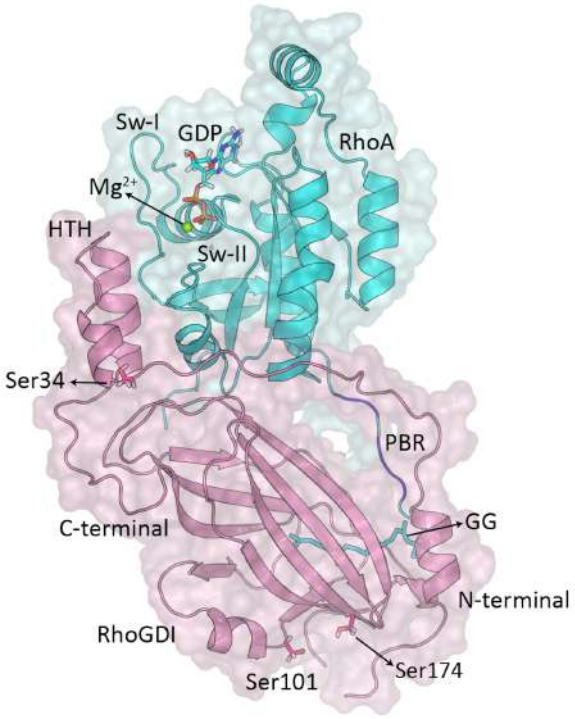

Figure 4.1: Crystal structure of the RhoA-RhoGDI complex depicted with RhoA (residues 1–190) in cyan and RhoGDI (residues 1–204) in pink. The polybasic region (PBR) of RhoA is highlighted in blue. The sites of phosphorylation (Ser34, Ser101, and Ser174), GDP, and the geranylgeranyl (GG) moiety are illustrated in stick representation. Additionally, Mg^{2+} is denoted as a green sphere.

the phosphorylation of RhoGDI emerges as a crucial posttranslational modification facilitating the dissociation of Rho GTPases from the complex[25–27]. Phosphorylation events at specific residues, such as Ser34 for RhoA[28],Ser96 for RhoA[29] and Ser101/Ser174 for Rac1[24,26], lead to selective release of the respective GTPases, suggesting the existence of an "unique phosphorylation code" governing their dissociation. The details about specificity is given in Table 4.1. It is speculated that phosphorylation of Ser34 disrupts its interaction with Arg68 of RhoA (or Arg66 for Rac1 or Cdc42), resulting in dissociation[30–32]. However, as this residue is conserved across most GTPases, the precise mechanism underlying this specificity remains unclear. However, the exact molecular mechanism underlying this specificity remains unclear. Numerous NMR investigations have demonstrated that two N-terminal segments of

RhoGDI (residues 9–20 and 36–58) exhibit considerable disorder in the absence of Rho GTPases[33–36]. Specifically, residues 36–58 of the N-terminal region adopts an HTH conformation upon interaction with Rho GTPases, a configuration stabilized by the establishment of multiple hydrogen bonds involving conserved residues from switch I (Thr37 and Val38) of RhoA (Thr35 and Val36 for Rac1 and Cdc42) and RhoGDI (Asp45 and Ser47)[37]. The former N-terminal region (residues 9–20) can adopt either helical or random coil conformations[38]. Recent crystallographic data indicate a helical structure (residues 10–15) in RhoA (PDB: 4F38)[39] and Cdc42 (PDB: 1DOA)[22], whereas for Rac1, it remains in a random coil conformation (PDB: 1HH4)[23]. This chapter delves into the molecular mechanisms underlying the phosphorylation-induced dissociation of Rho GTPases from RhoGDIs.

In this study, we utilize atomistic molecular dynamics (MD) simulations spanning several microseconds to probe the conformational dynamics and energetics of the RhoA-RhoGDI complex in both its wild-type (WT) form and two distinct phosphorylated states of RhoGDI. Our primary objective is to ascertain the existence of a phosphorylation code by demonstrating that phosphorylation at the RhoA-specific site, Ser34, diminishes the binding energy between RhoA and RhoGDI, unlike phosphorylation at the Rac1-specific sites, Ser101/Ser174. Subsequently, we conduct a comprehensive examination of the structural and energetic aspects to delineate the mechanistic pathway through which RhoA-specific phosphorylation allosterically influences a remote site, leading to a reduction in binding affinity. Ultimately, our aim is to uncover the underlying molecular thermodynamic basis for the specificity of the phosphorylation code governing the RhoA-RhoGDI interaction. The results presented in this chapter, were previously publised in *Sinha et al Biophysical Journal, (2023), 123(1), 57*

4.2 Computational Details

Modeling and Simulation

For modeling and simulation, we utilized the crystal structure of the RhoA-RhoGDI complex (PDB: 4F38) (Figure 4.1). The GNP molecule present in the crystal structure was replaced with GDP using the PyMOL program[40]. The *pdbfixer* utility of the OpenMM package[41] was employed to model the missing residues and atoms. The PyTMs plugin of PyMOL was utilized to prepare the phosphorylated system. Three different systems were constructed for our study: (1) WT: RhoA-GDP in complex with WT RhoGDI, (2) SP34: RhoA-GDP in complex with phosphorylated Ser34 (pS34) RhoGDI, and (3) SP101/174: RhoA-GDP in complex with phosphorylated Ser101/Ser174 (pS101/pS174) RhoGDI.

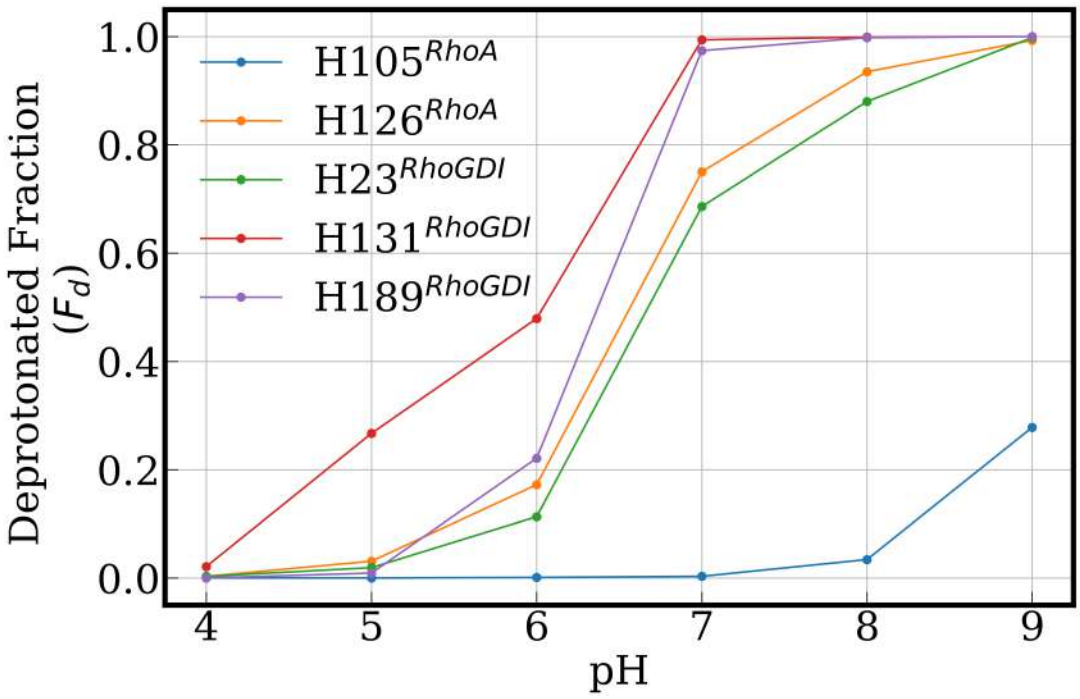

Figure 4.2: Plot illustrating the proportion of deprotonated histidine residues against varying pH levels. At a pH of 7, only His105 retains notable protonation, while all other histidine residues remain deprotonated.

Constant pH MD

A 20 ns constant pH MD simulation of the WT RhoA-RhoGDI complex was performed to determine the protonation states of His amino acids. The constant pH MD simulations were conducted using Amber18 software[42] with the amber14sb force field[43]. The titration graph is shown in the supporting material (Figure 4.2). Only His105 remained protonated at pH 7; therefore, His105 was kept in the protonated state for all systems.

MD Simulations

Four independent 1 μs long MD simulations were performed for each of the three systems using GROMACS 2019.6 software[44]. The CHARMM all-atom force field (version 36m) was employed, along with the TIP3P water model[45,46]. Energy minimization of the solvated proteins was achieved using the steepest descent algorithm. A modified Berendsen thermostat[47] and Parrinello-Rahman barostat[48] were utilized in an NPT ensemble at 300 K and 1 bar, respectively. Long-range electrostatic interactions were computed using the particle mesh Ewald summation method[49] with a grid spacing of 0.16 nm and fourth-order cubic interpolation. The cutoff distance for short-range electrostatic and van der Waals interactions was set to 1 nm. All covalent bonds were constrained using the LINCS algorithm[50], and a time step of 2 fs was employed. 0.15 M salt concentration was used to mimic the actual system.

Electrostatic Interaction Energy Calculations:

Phosphorylation introduces an additional negative charge to the system. To investigate how this perturbation affects structural modifications at a distant site, we computed the perturbation/change in the average interaction energy for each residue upon phosphorylation. Pairwise changes in interaction energy and H-bond occupancy were analyzed to elucidate how the signal propagates from the site of phosphorylation to the protein-protein interaction (PPI) interface.

Binding Energy Calculations:

The binding energy between RhoA and RhoGDI was calculated using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method[51]. The gmx_MMPBSA software was employed to calculate the binding energy[52].

Principal-Component Analysis:

Principal-component analysis (PCA) was performed on backbone coordinates of a combined initial 1 μs trajectory to visualize and interpret the conformational landscape sampled by the proteins.

4.3 Results and Discussion

4.3.1 MM-PBSA Analysis of Binding Energy between RhoA and RhoGDI Proteins

Drawing from the concept of a phosphorylation code as discussed earlier, we compared the binding energy between RhoA and RhoGDI across three distinct systems: 1) WT: RhoA-GDP in complex with WT RhoGDI, 2) SP34: RhoA-GDP in complex with phosphorylated Ser34 (pS34) RhoGDI, and 3) SP101/174: RhoA-GDP in complex with phosphorylated Ser101/Ser174 (pS101/pS174) RhoGDI. Upon RhoA-specific phosphorylation (SP34 system), we anticipated a weakening of the RhoA-RhoGDI interaction, contrasting with the lack of impact observed for Rac1-specific phosphorylation (SP101/174 system). A decrease in binding affinity (indicated by less negative values of the binding energy) suggests a diminished interaction between RhoA and RhoGDI. Figure 4.3(a) illustrates the probability distribution of binding energy for the different systems. With Ser34 phosphorylation, the distribution shifts towards more positive values, signifying a weakened protein-protein interaction (PPI), whereas for Ser101/Ser174 phosphorylation, the binding energy distribution remains comparable to the WT type system. Decomposing the binding energy residue-wise

for the WT system reveals two primary binding regions [Figure 4.3(b)]: 1) switch II of RhoA and HTH motif of RhoGDI, and 2) polybasic region (PBR) of RhoA and N-terminal region (residues 9–25) of RhoGDI. Mapping the residues on the structure indicates that these major binding regions contain more residues with favorable contributions to binding energy compared to those with unfavorable contributions [Figure 4.3 (b and c)]. Figure 4.4 depicts the residue-wise change in binding energy

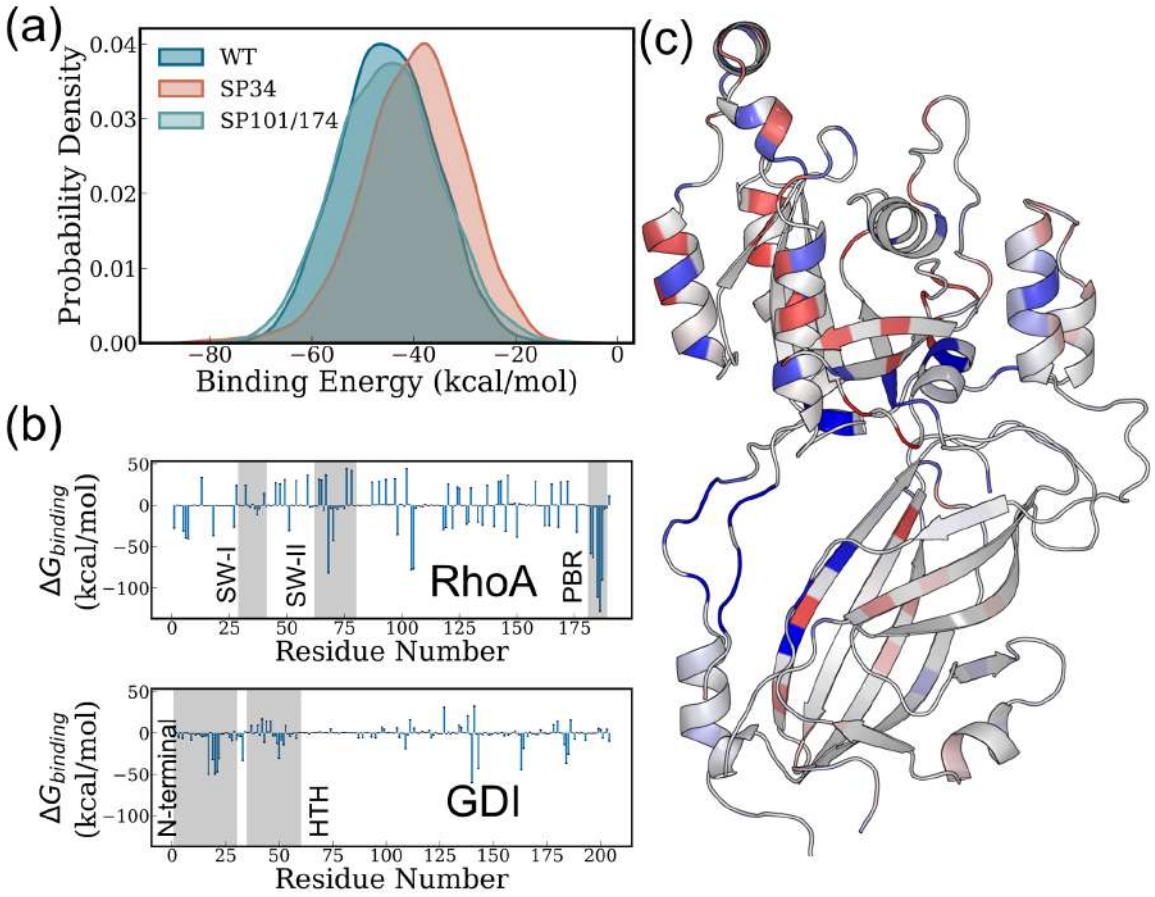

Figure 4.3: (a) Comparative analysis of the binding energy distribution among the WT, SP34, and SP101/174 systems for the RhoA-RhoGDI complex using MM-PBSA. (b) Decomposition of binding energy at the residue level within the WT system for both RhoA and RhoGDI proteins. (c) Structural representation of the WT system highlighting residues contributing favorably ($\Delta G_{binding} < 0$) and unfavorably ($\Delta G_{binding} > 0$) to the binding energy. Residues with favorable contributions are depicted in blue, while those with unfavorable contributions are depicted in red.

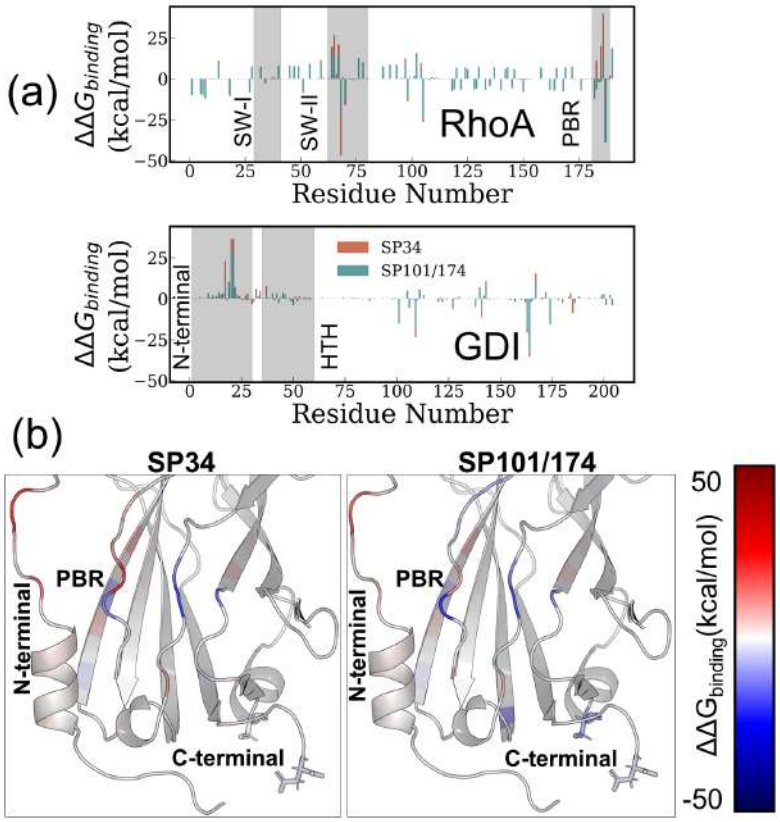

Figure 4.4: (a) Analysis of residue-wise contributions to the change in binding energy for the SP34 and SP101/174 systems. Significant changes are observed primarily in the polybasic region (PBR) of RhoA and the N-terminal region of RhoGDI, along with alterations near the switch regions of RhoA and the HTH region of RhoGDI. (b) Representation of the PBR/N-terminal region interaction in the RhoA-RhoGDI complex, with each residue color-coded based on its contribution. Residues S101 and S174 are depicted in stick representation. The weakening of the interaction between the PBR and N-terminal is more pronounced in the SP34 system, while the strengthening of the interaction between the N-terminal and C-terminal is more notable in the SP101/174 system.

for the phosphorylated systems. Notably, the interaction between the PBR and the N-terminal region weakens in both SP34 and SP101/174 phosphorylation scenarios [Figure 4.4 (b)], with the weakening more pronounced for SP34 phosphorylation. For the SP101/174 system, this weakening is compensated by more negative contributions from phosphorylated Ser residues (pS101 and pS174), maintaining the binding energy almost unchanged. This observation corroborates previous experimental find-

ings[28] and reinforces the notion of a phosphorylation code. Our hypothesis posits

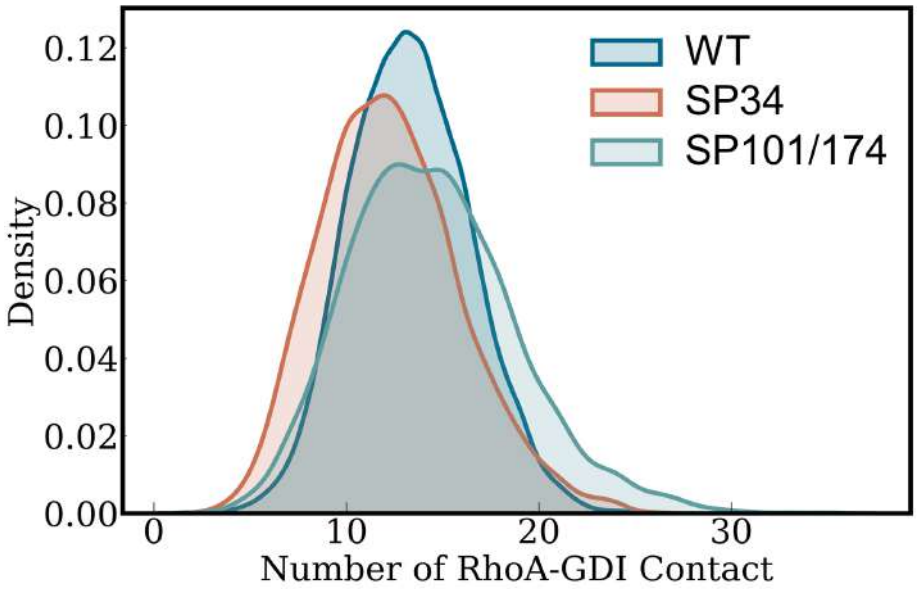

Figure 4.5: Comparison of the total number of contacts between RhoA and RhoGDI across different phosphorylation states. The analysis indicates a reduction in the total number of RhoA-RhoGDI contacts for the SP34 system compared to the WT, whereas the SP101/174 phosphorylation state shows a marginal increase in these contacts.

that SP101/174 phosphorylation strengthens the interaction between the positively charged PBR of RhoA and the C-terminal region of RhoGDI. To validate this, we quantified the total number of contacts between RhoA and RhoGDI for the WT, SP34, and SP101/174 systems (Figure 4.5). The data reveal a decrease in contacts between RhoA and RhoGDI for the SP34 system, but not for the SP101/174 system. Subsequently, the number of contacts and the short-range interaction energy between PBR/N-terminal, PBR/C-terminal, and N-terminal/C-terminal were determined. Results indicate a decrease in both contacts and interaction energy between PBR and the N-terminal region for both phosphorylation scenarios, with a more significant decrease observed in the SP34 system. Conversely, the number of PBR/C-terminal contacts and interaction energy increase for both phosphorylation scenarios, with a more pronounced increase in SP101/174 phosphorylation, in line with our initial hypothesis.

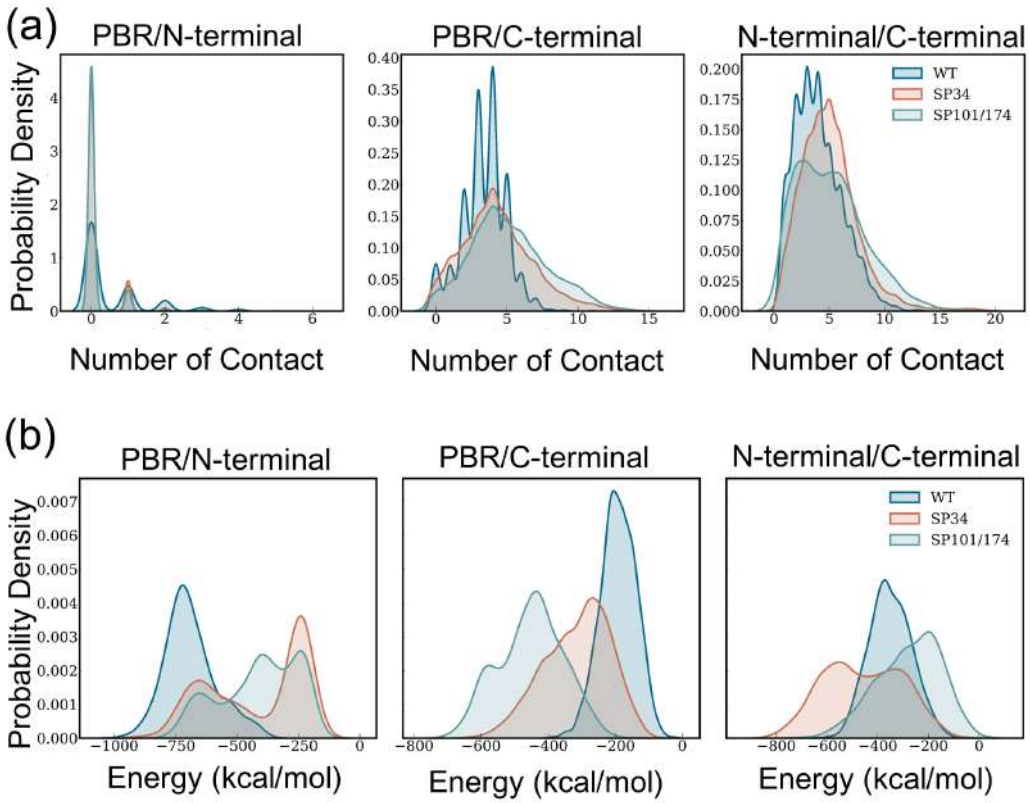

Figure 4.6: (a) Comparison of the number of contacts and (b) the interaction energy among the PBR/N-terminal, PBR/C-terminal, and N-terminal/C-terminal regions for the WT, SP34, and SP101/174 systems.

Having established that phosphorylation at Ser34 reduces the RhoA-RhoGDI interaction, subsequent sections will focus on unraveling the molecular mechanism underlying this effect in the SP34 system.

4.3.2 Structural Changes due to Ser34 Phosphorylation:

The anticipated reduction in binding affinity between RhoA and RhoGDI post-phosphorylation suggests accompanying structural alterations affecting the protein-protein interaction (PPI) interface. Root mean-square deviation (RMSD) computations for both RhoA and RhoGDI, referenced to the crystal structure, were conducted to monitor any such structural deviations. While RhoA's RMSD remains relatively

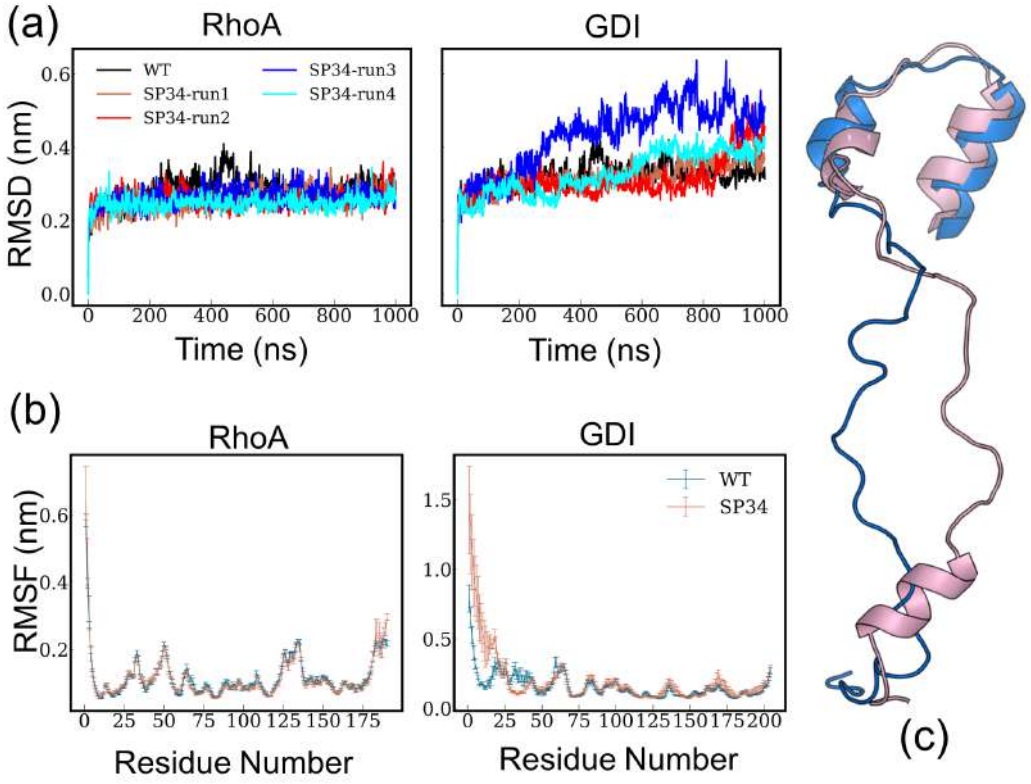

Figure 4.7: (a) Time series of RMSD with respect to the crystal structure for both WT (black) and four independent simulations for the SP34 system (pink, red, blue, and cyan) separately for the RhoA and RhoGDI proteins. Error bars represent the standard deviation computed from four independent trajectories. (b) Residue-wise RMSF for RhoA and RhoGDI separately for the WT and SP34 systems. Error bars denote the standard deviation computed from four independent trajectories. (c) Superimposed structures of the N-terminal region for the WT (pink) and SP34 (blue) systems.

stable, significant changes are observed for RhoGDI, noted across four independent simulations of the SP34 system [Figure 4.7(a)]. Visual inspection reveals three notable structural changes post-phosphorylation: 1) melting of the N-terminal helix, 2) conformational alteration of the N-terminal loop, and 3) outward movement of RhoA's geranylgeranyl group from RhoGDI's hydrophobic cavity. Residue-wise root mean-square fluctuation (RMSF) comparisons between WT and phosphorylated (SP) systems elucidated notable changes primarily in RhoA's polybasic region (PBR) and

RhoGDI's N-terminal region, alongside minor variations in RhoA's switch I and switch II loops [Figure 4.7(b, c)].

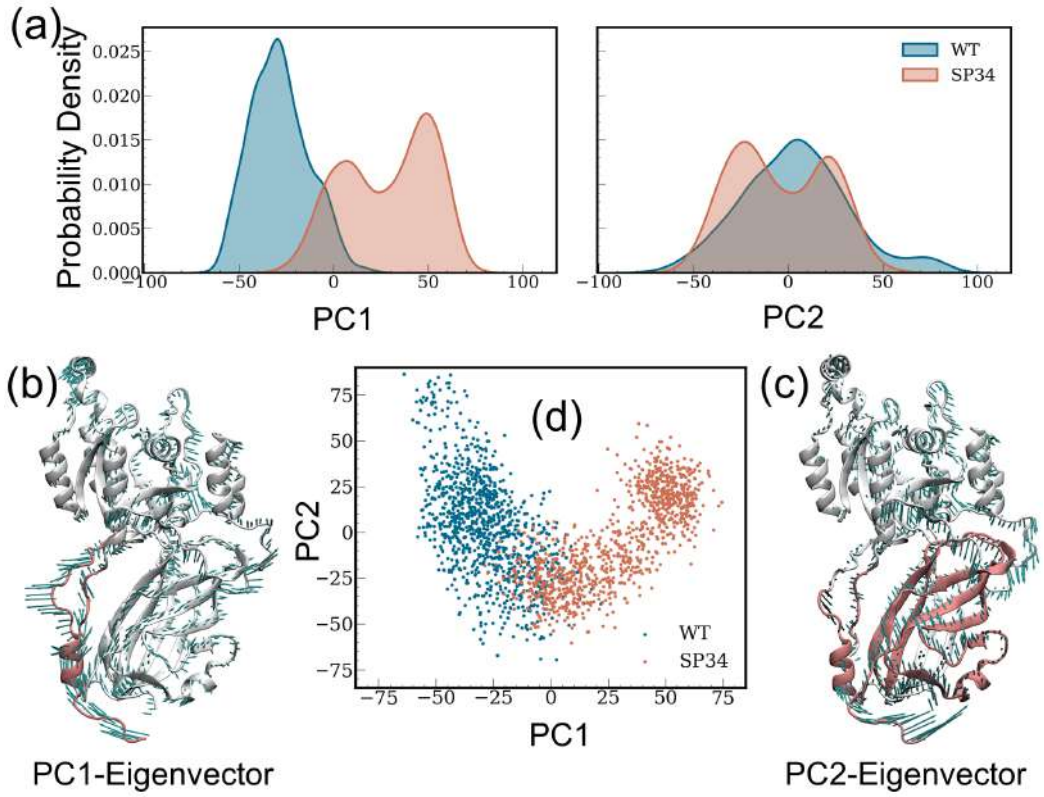

Figure 4.8: Principal-component analysis (PCA) results. (a) Distribution plots of PC1 and PC2 for the WT and SP34 systems. Projection of the eigenvectors of (b) PC1 and (c) PC2 onto the protein structure of the RhoA-RhoGDI complex. (d) Projection of the first two principal components, PC1 and PC2, for the WT and SP34 systems.

To comprehensively characterize phosphorylation-induced conformational ensemble modulation and dynamics, principal-component analysis (PCA) was performed on combined WT and SP34 trajectories. PC1 and PC2 captured the major differences in protein conformation, with PC1 illustrating an outward movement of RhoGDI's N-terminal region and subsequent PBR displacement, while PC2 depicted an upward shift in RhoGDI's C-terminal β sandwich (Figure 4.8). PC2 also showcased the

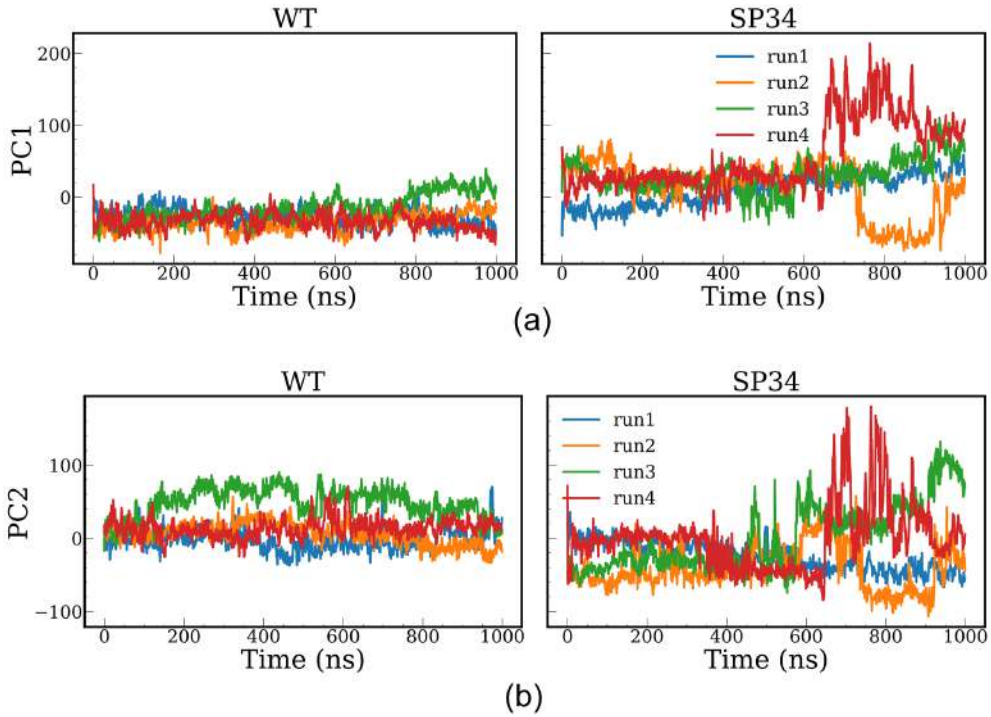

Figure 4.9: Time series plots of PC1 and PC2 for the WT and SP34 systems across all independent simulations. Panels (a) and (b) depict PC1 and PC2 for the WT and SP34, respectively. Minimal fluctuations in both PC1 and PC2 are observed in the WT system, indicating a well-equilibrated simulation. Conversely, the SP34 system displays substantial variations in both PC1 and PC2, implying significant conformational changes and suggesting that these simulations are in a non-equilibrium state.

N-terminus of RhoGDI moving towards the C-terminal β sandwich. These dynamic modes reflect significant conformational changes, particularly evident in the SP34 system, as depicted by minimal fluctuations in WT simulations but substantial variations in both PC1 and PC2 for SP34 (Figure 4.9). These findings suggest that while WT simulations reached equilibrium, SP34 simulations remained in a nonequilibrium state.

Two major structural rearrangements upon phosphorylation were observed: 1) RhoA's N-terminal region moving away from RhoGDI's PBR, and 2) loosening of RhoA's geranylgeranyl hydrophobic tail from RhoGDI's cavity. Monitoring the number of contacts between these regions revealed a decrease in contacts between

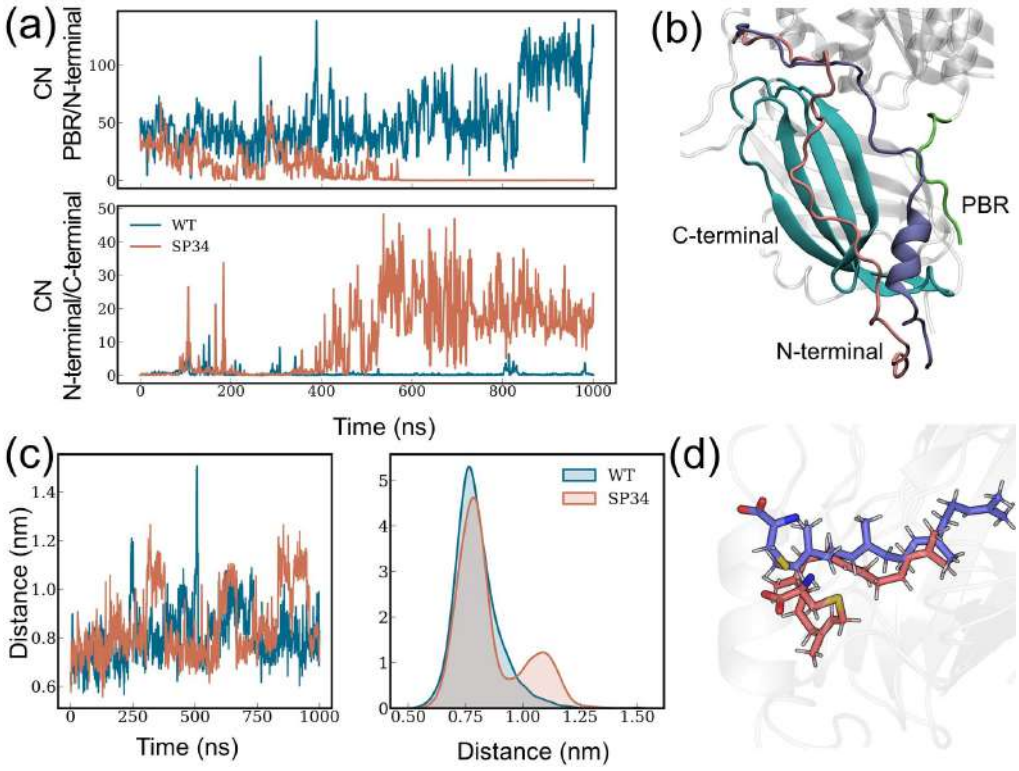

Figure 4.10: Coordination number (CN) analysis. (a) Time series of CNs between the N-terminal region of RhoGDI and RhoA PBR (upper panel), and between the N-terminal and C-terminal regions of RhoGDI (lower panel). (b) Snapshot illustrating the movement of the N-terminal region of RhoGDI away from RhoA PBR. (c) Time series of the distance (left panel) and probability distribution of the distance (right panel) between the last carbon atom of the geranylgeranyl group and the center of mass (COM) of the hydrophobic cavity. (d) Snapshot highlighting the outward movement of the geranylgeranyl group, with WT depicted in sky blue and SP34 in salmon.

RhoGDI's N-terminal region and RhoA's PBR, alongside an increase in contacts between RhoGDI's N-terminal and C-terminal regions in the phosphorylated SP34 system [Figure 4.10(a)]. This suggests movement of RhoGDI's N-terminal region away from the PBR but towards its C-terminal region [Figure 4.10(b)], leading to decreased RhoA-RhoGDI interaction and specific RhoA release from RhoGDI. Similarly, calculations of the geranylgeranyl tail's distance from RhoGDI's hydrophobic cavity indicated an increased distance post-phosphorylation, suggesting a tendency for RhoA's geranylgeranyl group to exit RhoGDI's hydrophobic cavity [Figure 4.10(c,d)].

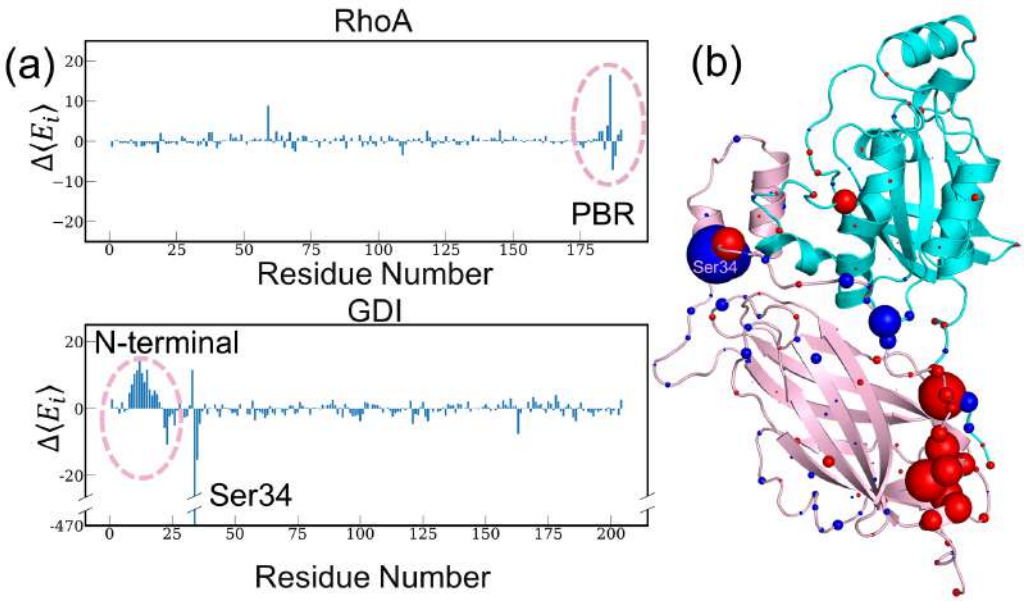

Figure 4.11: (a) Residue-specific alteration in interaction energy ($\Delta\langle E_i\rangle$) depicted for RhoA and RhoGDI. (b) Each residue's Cα atom depicted on the protein structure, with sphere size reflecting the absolute value of $\Delta\langle E_i\rangle$ and sphere color indicating the direction of the energy change. Blue and red signify negative and positive energy changes, respectively, with the most pronounced alteration occurring at the phosphorylation site.

4.3.3 Propagation of Signal and Mechanism of Dissociation:

The introduction of extra negative charges through phosphorylation initiates a local perturbation within the electrostatic interaction network. This redistribution of electrostatic interactions, involving the rewiring of hydrogen bonds and salt bridges, akin to a domino effect, can serve as a dominant mechanism of allostery, as previously demonstrated[53–55]. Additionally, electrostatic interaction serves as a sensitive indicator of even subtle conformational changes, thus warranting the tracking of residue-wise changes in average interaction energy to comprehend how the energetic perturbation emanating from the phosphorylation site propagates to the protein-protein interaction (PPI) interface, weakening it. This change in residue-wise average interaction energy

$$\Delta\langle E_i\rangle = \langle E_i^{SP}\rangle - \langle E_i^{WT}\rangle \tag{4.1}$$

where $\langle E_i^{SP} \rangle$ and $\langle E_i^{WT} \rangle$ represent the average nonbonded interaction energy of the i-th residue with its environment (protein, water, and ions) for wild type and ser34 phosphorylated system respectively. It is pivotal in understanding the propagation of the energetic perturbation. As anticipated, the most substantial change occurs at the phosphorylation site, namely Ser34. Notably, significant alterations in interaction energy are observed near the polybasic region (PBR) and N-terminal regions, crucial for stabilizing the PPI [Figure 4.11(a)]. Visualizing these residues with notable interaction energy perturbation underscores the dramatic long-range (allosteric) effect caused by phosphorylation, leading to substantial structural and energetic rearrangements in the PBR and N-terminal region [Figure 4.11(b)]. The subsequent inquiry delves into understanding how the energetic perturbation initiated by Ser34 phosphorylation propagates over a considerable distance ($\geqslant$3.5 nm) from the phosphorylation site to RhoA's PBR. To trace this propagation pathway, an examination of the under-

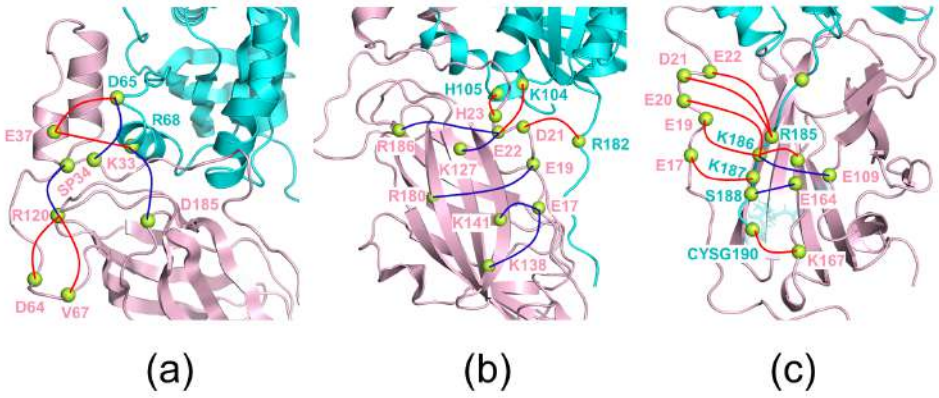

Figure 4.12: Mechanistic picture of signal propagation from the site of phosphorylation (SP34) to PBR in terms of change in residue pairwise average interaction energy ($\Delta\langle\Delta E_{ij}\rangle$) for (a) the HTH region, (b) the C-terminal β sandwich, and (c) the N-terminal region of RhoGDI. Blue and red lines indicate the residue pairs where the interaction becomes stronger ($\Delta\langle E_{ij}\rangle < 0$) and weaker ($\Delta\langle E_{ij}\rangle > 0$), respectively, upon phosphorylation.

lying interaction energy network's rewiring is undertaken. This involves computing the residue pairwise average interaction energy and its change upon phosphorylation:

$$\Delta\langle E_{ij}\rangle = \langle E_{ij}^{SP}\rangle - \langle E_{ij}^{WT}\rangle \tag{4.2}$$

The resultant network of residue pairs with significant changes in average interaction energy illustrates the propagation of perturbation, facilitating the visualization of molecular events and local rearrangements connecting the phosphorylation site (Ser34) to the distal PBR. Upon phosphorylation at Ser34, a negative charge is introduced at pS34, forming a salt bridge with Arg120 of RhoGDI. Subsequently, Arg68 of RhoA, previously favorably interacting with Glu37, forms an intermolecular hydrogen bond with Asp185 of RhoGDI. This leads to increased flexibility of RhoGDI's N-terminal region, causing the N-terminal region to move towards the C-terminal region of RhoGDI. Concurrently, the PBR detaches from the N-terminal region and establishes new contacts with the C-terminal region. Notably, the hydrogen bond between CYSG190^{RhoA} and Lys167^{RhoGDI}, crucial for the geranylgeranyl group's release from the hydrophobic cavity, breaks upon phosphorylation. E_{ij} values for the pairs involved in the allosteric pathway are given in Table 6.9 and the distance distribution of these pairs for WT and SP systems are given in Figure 4.13

Additionally, the formation of a stronger interaction between Ser188^{RhoA} and Glu163^{RhoGDI} post-phosphorylation suggests an intermediate state towards the eventual dissociation of the RhoA-RhoGDI complex. This result is in agreement with the fact that phosphorylation of Ser188^{RhoA} stabilizes the RhoA-RhoGDI complex[56].These mechanistic insights underscore the intricate interplay of phosphorylation-induced changes and their profound effects on the stability and dynamics of the RhoA-RhoGDI complex.

4.4 Conclusion

This study elucidates the existence of a "phosphorylation code" governing the selective dissociation of the RhoA-RhoGDI complex. Through extensive microsecond-long MD simulation trajectories, we present a comprehensive microscopic and mechanistic understanding of the initial phases of phosphorylation. Our findings reveal structural alterations, particularly the weakening of the protein-protein interaction (PPI), occurring specifically upon phosphorylation at the RhoA-specific site (Ser34). MM-PBSA

Residue Pairs	$\langle E_{ij}^{WT} \rangle$ **(kcal/mol)**	$\langle E_{ij}^{SP} \rangle$ **(kcal/mol)**	$\Delta\langle E_{ij} \rangle = \langle E_{ij}^{SP} \rangle - \langle E_{ij}^{WT} \rangle$ **(kcal/mol)**
S34^{RhoGDI}-R120^{RhoGDI}	0.9	-115.8	-116.7
E22^{RhoGDI}-K127^{RhoGDI}	-24.1	-75.0	-50.9
K186^{RhoA}-E109^{RhoGDI}	-11.8	-50.9	-39.1
E22^{RhoGDI}-K186^{RhoGDI}	-12.0	-41.1	-29.1
R68^{RhoA}-R185^{RhoGDI}	-70.2	-97.5	-27.3
H105^{RhoA}-E22^{RhoGDI}	-18.3	-44.6	-26.3
E17^{RhoGDI}-K141^{RhoGDI}	-46.3	-62.6	-16.3
E17^{RhoGDI}-K138^{RhoGDI}	-21.9	-36.5	-14.5
S188^{RhoA}-E164^{RhoGDI}	1.2	-3.7	-4.9
E19^{RhoGDI}-R180^{RhoGDI}	-11.4	-15.9	-4.5
D65^{RhoA}-K33^{RhoGDI}	-63.9	-67.5	-3.6
R68^{RhoA}-E337^{RhoGDI}	-26.4	-22.5	3.9
V64^{RhoGDI}-R120^{RhoGDI}	-30.1	-23.3	6.8
R68^{RhoA}-K33^{RhoGDI}	18.2	25.0	6.8
K186^{RhoA}-E163^{RhoGDI}	-38.0	-31.0	7.0
CYSG190^{RhoA}-K167^{RhoGDI}	-42.0	-34.1	7.9
R182^{RhoA}-D21^{RhoGDI}	-16.4	-7.4	9.0
D65^{RhoA}-E37^{RhoGDI}	28.7	38.1	9.4
K187^{RhoA}-E17^{RhoGDI}	-36.6	-24.9	11.7
K186^{RhoA}-E19^{RhoGDI}	-24.6	-8.0	16.6
R185^{RhoA}-E21^{RhoGDI}	-36.6	-17.5	19.1
R185^{RhoA}-D22^{RhoGDI}	-40.1	-12.9	27.2
K104^{RhoA}-E21^{RhoGDI}	-46.4	-17.9	28.5
K186^{RhoA}-E20^{RhoGDI}	-55.3	-5.5	49.8

Table 4.2: $\Delta\langle E_{ij} \rangle$ values representing the interaction energy changes for all residue pairs involved in the allosteric network. All values are reported in kcal/mol

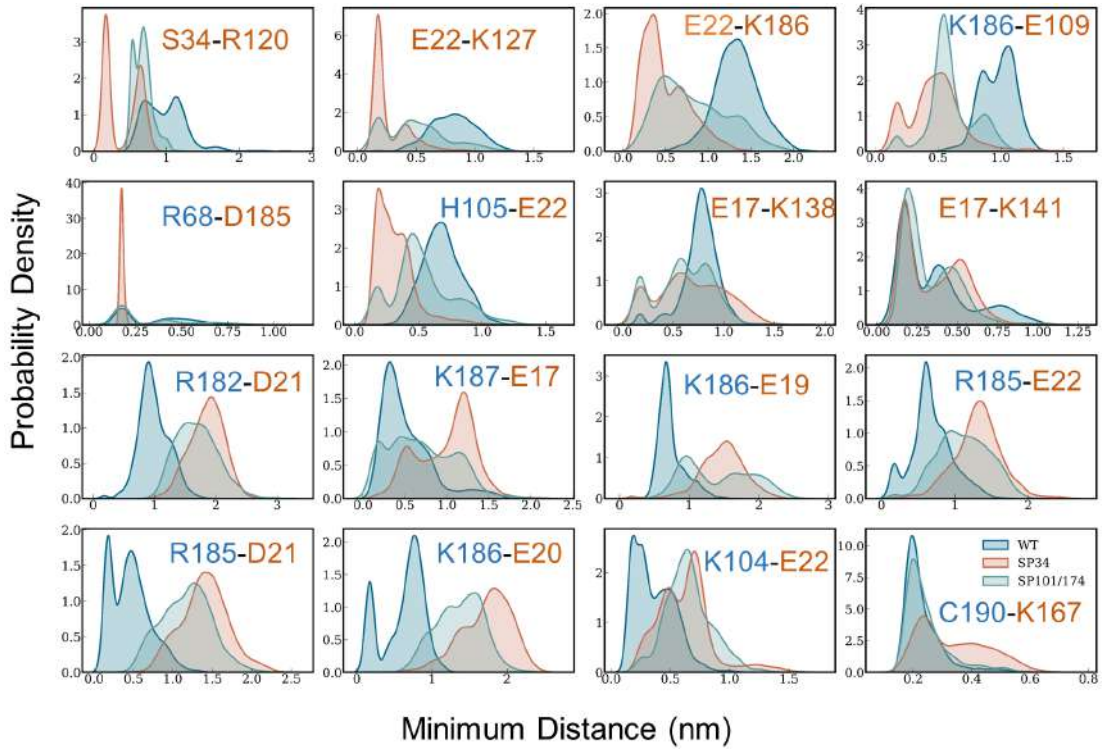

Figure 4.13: Distribution of minimum distances between residues involved in the signal propagation network for three distinct systems: WT, SP34, and SP101/174. RhoA residues are depicted in blue, while RhoGDI residues are represented in orange.

binding energy calculations unequivocally demonstrate a significant decrease in PPI strength following phosphorylation at Ser34, contrasting with the minor impact observed for phosphorylation at Ser101/Ser174, presumed to release Rac1 instead of RhoA.

The robust binding between RhoA and RhoGDI primarily arises from the strong interaction between the negatively charged N-terminal region of RhoGDI and the positively charged polybasic region (PBR) of RhoA. Phosphorylation at Ser34 weakens the interaction between the helix-turn-helix (HTH) region of RhoGDI and the switch region of RhoA, thereby enhancing flexibility in the N-terminal loop, which subsequently moves away from the PBR. Major structural changes include the displacement of the N-terminal region of RhoGDI from the PBR of RhoA towards its C-terminal region, along with the extrication of the geranylgeranyl group from RhoGDI's hy-

drophobic cavity.

We utilize perturbations in residue-wise (and pairwise) interaction energy as indicators of the structural rearrangements responsible for propagating the energetic perturbation induced by phosphorylation. Our proposed mechanistic model delineates the propagation of the distant signal from the phosphorylation site to the PBR and buried geranylgeranyl group through the rearrangement and rewiring of hydrogen bonds and salt bridges (charge-charge interactions). This underscores the pivotal role of local and specific electrostatic interactions in driving the observed allosteric response.

Bibliography

[1] BISHOP, A. L & HALL, A. (2000) Rho gtpases and their effector proteins. *Biochemical Journal* **348**, 241–255.

[2] Kaibuchi, K, Kuroda, S, & Amano, M. (1999) Regulation of the cytoskeleton and cell adhesion by the rho family gtpases in mammalian cells. *Annual Review of Biochemistry* **68**, 459–486.

[3] Raftopoulou, M & Hall, A. (2004) Cell migration: Rho gtpases lead the way. *Developmental Biology* **265**, 23–32.

[4] Lu, S, Jang, H, Muratcioglu, S, Gursoy, A, Keskin, O, Nussinov, R, & Zhang, J. (2016) Ras conformational ensembles, allostery, and signaling. *Chemical Reviews* **116**, 6607–6665.

[5] Wennerberg, K, Rossman, K. L, & Der, C. J. (2005) The ras superfamily at a glance. *Journal of Cell Science* **118**, 843–846.

[6] Nethe, M & Hordijk, P. L. (2010) The role of ubiquitylation and degradation in rhogtpase signalling. *Journal of Cell Science* **123**, 4011–4018.

[7] Kumawat, A, Chakrabarty, S, & Kulkarni, K. (2017) Nucleotide dependent switching in rho gtpase: Conformational heterogeneity and competing molecular interactions. *Scientific Reports* **7**.

[8] Gorfe, A. A, Grant, B. J, & McCammon, J. A. (2008) Mapping the nucleotide and isoform-dependent structural and dynamical features of ras proteins. *Structure* **16**, 885–896.

[9] Vetter, I. R & Wittinghofer, A. (2001) The guanine nucleotide-binding switch in three dimensions. *Science* **294**, 1299–1304.

[10] Cherfils, J & Zeghouf, M. (2013) Regulation of small gtpases by gefs, gaps, and gdis. *Physiological Reviews* **93**, 269–309.

[11] Lowy, D. R & Willumsen, B. M. (1993) Function and regulation of ras. *Annual Review of Biochemistry* **62**, 851–891.

[12] Hall, A. (2012) Rho family gtpases. *Biochemical Society Transactions* **40**, 1378–1382.

[13] Etienne-Manneville, S & Hall, A. (2002) Rho gtpases in cell biology. *Nature* **420**, 629–635.

[14] DerMardirossian, C & Bokoch, G. M. (2005) Gdis: central regulatory molecules in rho gtpase activation. *Trends in Cell Biology* **15**, 356–363.

[15] Falkenberg, C. V & Loew, L. M. (2013) Computational analysis of rho gtpase cycling. *PLoS Computational Biology* **9**, e1002831.

[16] Cho, Kim, Baek, Kim, & Lee. (2019) Regulation of rho gtpases by rhogdis in human cancers. *Cells* **8**, 1037.

[17] Garcia-Mata, R, Boulter, E, & Burridge, K. (2011) The "invisible hand": regulation of rho gtpases by rhogdis. *Nature Reviews Molecular Cell Biology* **12**, 493–504.

[18] COX, A & DER, C. (1992) Protein prenylation: more than just glue? *Current Opinion in Cell Biology* **4**, 1008–1016.

[19] Michaelson, D, Ali, W, Chiu, V. K, Bergo, M, Silletti, J, Wright, L, Young, S. G, & Philips, M. (2005) Postprenylationcaaxprocessing is required for proper localization of ras but not rho gtpases. *Molecular Biology of the Cell* **16**, 1606–1616.

[20] Winter-Vann, A. M & Casey, P. J. (2005) Post-prenylation-processing enzymes as new targets in oncogenesis. *Nature Reviews Cancer* **5**, 405–412.

[21] Scheffzek, K, Stephan, I, Jensen, O. N, Illenberger, D, & Gierschik, P. (2000). *Nature Structural Biology* **7**, 122–126.

[22] Hoffman, G. R, Nassar, N, & Cerione, R. A. (2000) Structure of the rho family gtp-binding protein cdc42 in complex with the multifunctional regulator rhogdi. *Cell* **100**, 345–356.

[23] Grizot, S, Fauré, J, Fieschi, F, Vignais, P. V, Dagher, M.-C, & Pebay-Peyroula, E. (2001) Crystal structure of the rac1–rhogdi complex involved in nadph oxidase activation,. *Biochemistry* **40**, 10007–10013.

[24] Li, R & Zheng, Y. (1997) Residues of the rho family gtpases rho and cdc42 that specify sensitivity to dbl-like guanine nucleotide exchange factors. *Journal of Biological Chemistry* **272**, 4671–4679.

[25] Qiao, J, Holian, O, Lee, B.-S, Huang, F, Zhang, J, & Lum, H. (2008) Phosphorylation of gtp dissociation inhibitor by pka negatively regulates rhoa. *American Journal of Physiology-Cell Physiology* **295**, C1161–C1168.

[26] DerMardirossian, C, Schnelzer, A, & Bokoch, G. M. (2004) Phosphorylation of rhogdi by pak1 mediates dissociation of rac gtpase. *Molecular Cell* **15**, 117–127.

[27] DerMardirossian, C, Rocklin, G, Seo, J.-Y, & Bokoch, G. M. (2006) Phosphorylation of rhogdi by src regulates rho gtpase binding and cytosol-membrane cycling. *Molecular Biology of the Cell* **17**, 4760–4768.

[28] Dovas, A, Choi, Y, Yoneda, A, Multhaupt, H. A, Kwon, S.-H, Kang, D, Oh, E.-S, & Couchman, J. R. (2010) Serine 34 phosphorylation of rho guanine dissociation inhibitor (rhogdiα) links signaling from conventional protein kinase c to rhogtpase in cell adhesion. *Journal of Biological Chemistry* **285**, 23296–23308.

[29] Knezevic, N, Roy, A, Timblin, B, Konstantoulaki, M, Sharma, T, Malik, A. B, & Mehta, D. (2007) Gdi-1 phosphorylation switch at serine 96 induces rhoa activation and increased endothelial permeability. *Molecular and Cellular Biology* **27**, 6323–6333.

[30] GANDHI, P. N, GIBSON, R. M, TONG, X, MIYOSHI, J, TAKAI, Y, KONIECZKOWSKI, M, SEDOR, J. R, & WILSON-DELFOSSE, A. L. (2004) An activating mutant of rac1 that fails to interact with rho gdp-dissociation inhibitor stimulates membrane ruffling in mammalian cells. *Biochemical Journal* **378**, 409–419.

[31] Gibson, R. M, Gandhi, P. N, Tong, X, Miyoshi, J, Takai, Y, Konieczkowski, M, Sedor, J. R, & Wilson-Delfosse, A. L. (2004) An activating mutant of cdc42 that fails to interact with rho gdp-dissociation inhibitor localizes to the plasma membrane and mediates actin reorganization. *Experimental Cell Research* **301**, 211–222.

[32] GIBSON, R. M & WILSON-DELFOSSE, A. L. (2001) Rhogdi-binding-defective mutant of cdc42hs targets to membranes and activates filopodia formation but does not cycle with the cytosol of mammalian cells. *Biochemical Journal* **359**, 285–294.

[33] Golovanov, A. P, Chuang, T.-H, DerMardirossian, C, Barsukov, I, Hawkins, D, Badii, R, Bokoch, G. M, Lian, L.-Y, & Roberts, G. C. (2001) Structure-activity relationships in flexible protein domains: regulation of rho gtpases by rhogdi and d4 gdi. *Journal of Molecular Biology* **305**, 121–135.

[34] Lian, L.-Y, Barsukov, I, Golovanov, A. P, Hawkins, D. I, Badii, R, Sze, K.-H, Keep, N. H, Bokoch, G. M, & Roberts, G. C. (2000) Mapping the binding site for the gtp-binding protein rac-1 on its inhibitor rhogdi-1. *Structure* **8**, 47–56.

[35] Gosser, Y. Q, Nomanbhoy, T. K, Aghazadeh, B, Manor, D, Combs, C, Cerione, R. A, & Rosen, M. K. (1997) C-terminal binding domain of rho gdp-dissociation inhibitor directs n-terminal inhibitory peptide to gtpases. *Nature* **387**, 814–819.

[36] Ueyama, T, Son, J, Kobayashi, T, Hamada, T, Nakamura, T, Sakaguchi, H, Shirafuji, T, & Saito, N. (2013) Negative charges in the flexible n-terminal domain of rho gdp-dissociation inhibitors (rhogdis) regulate the targeting of the rhogdi–rac1 complex to membranes. *The Journal of Immunology* **191**, 2560–2569.

[37] Keep, N. H, Barnes, M, Barsukov, I, Badii, R, Lian, L.-Y, Segal, A. W, Moody, P. C, & Roberts, G. C. (1997) A modulator of rho family g proteins, rhogdi, binds these g proteins via an immunoglobulin-like domain and a flexible n-terminal arm. *Structure* **5**, 623–633.

[38] Golovanov, A. P, Hawkins, D, Barsukov, I, Badii, R, Bokoch, G. M, Lian, L, & Roberts, G. C. K. (2001) Structural consequences of site-directed mutagenesis in flexible protein domains: Nmr characterization of the l(55, 56)s mutant of rhogdi. *European Journal of Biochemistry* **268**, 2253–2260.

[39] Tnimov, Z, Guo, Z, Gambin, Y, Nguyen, U. T, Wu, Y.-W, Abankwa, D, Stigter, A, Collins, B. M, Waldmann, H, Goody, R. S, & Alexandrov, K. (2012) Quantitative analysis of prenylated rhoa interaction with its chaperone, rhogdi. *Journal of Biological Chemistry* **287**, 26549–26562.

[40] Schrödinger, LLC. (2015) The PyMOL molecular graphics system, version 1.8.

[41] Eastman, P, Swails, J, Chodera, J. D, McGibbon, R. T, Zhao, Y, Beauchamp, K. A, Wang, L.-P, Simmonett, A. C, Harrigan, M. P, Stern, C. D, Wiewiora, R. P, Brooks, B. R, & Pande, V. S. (2017) Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **13**, e1005659.

[42] Case, D. A, Cheatham, T. E, Darden, T, Gohlke, H, Luo, R, Merz, K. M, Onufriev, A, Simmerling, C, Wang, B, & Woods, R. J. (2005) The amber biomolecular simulation programs. *Journal of Computational Chemistry* **26**, 1668–1688.

[43] Maier, J. A, Martinez, C, Kasavajhala, K, Wickstrom, L, Hauser, K. E, & Simmerling, C. (2015) ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation* **11**, 3696–3713.

[44] Abraham, M. J, Murtola, T, Schulz, R, Páll, S, Smith, J. C, Hess, B, & Lindahl, E. (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25.

[45] Klauda, J. B, Venable, R. M, Freites, J. A, O'Connor, J. W, Tobias, D. J, Mondragon-Ramirez, C, Vorobyov, I, MacKerell, A. D, & Pastor, R. W. (2010) Update of the charmm all-atom additive force field for lipids: Validation on six lipid types. *The Journal of Physical Chemistry B* **114**, 7830–7843.

[46] Huang, J, Rauscher, S, Nawrocki, G, Ran, T, Feig, M, de Groot, B. L, Grubmüller, H, & MacKerell, A. D. (2016) Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73.

[47] Bussi, G, Donadio, D, & Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**.

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

[49] Essmann, U, Perera, L, Berkowitz, M. L, Darden, T, Lee, H, & Pedersen, L. G. (1995) A smooth particle mesh ewald method. *The Journal of Chemical Physics* **103**, 8577–8593.

[50] Hess, B, Bekker, H, Berendsen, H. J. C, & Fraaije, J. G. E. M. (1997) Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463–1472.

[51] Baker, N. A, Sept, D, Joseph, S, Holst, M. J, & McCammon, J. A. (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences* **98**, 10037–10041.

[52] Valdés-Tresanco, M. S, Valdés-Tresanco, M. E, Valiente, P. A, & Moreno, E. (2021) gmx_mmpbsa: A new tool to perform end-state free energy calculations with gromacs. *Journal of Chemical Theory and Computation* **17**, 6281–6291.

[53] Kumawat, A & Chakrabarty, S. (2017) Hidden electrostatic basis of dynamic allostery in a pdz domain. *Proceedings of the National Academy of Sciences* **114**.

[54] Liu, J & Nussinov, R. (2017) Energetic redistribution in allostery to execute protein function. *Proceedings of the National Academy of Sciences* **114**, 7480–7482.

[55] Kumawat, A & Chakrabarty, S. (2020) Protonation-induced dynamic allostery in pdz domain: Evidence of perturbation-independent universal response network. *The Journal of Physical Chemistry Letters* **11**, 9026–9031.

[56] Ellerbroek, S. M, Wennerberg, K, & Burridge, K. (2003) Serine phosphorylation negatively regulates rhoa in vivo. *Journal of Biological Chemistry* **278**, 19023–19031.

5

Role of Protonation State of Histidine in Phosphorylation Induced Dissociation of Rac1

"Shall I refuse my dinner because I do not fully understand the process of digestion?"

Oliver Heaviside

5.1 Introduction

Despite belonging to the same family, RhoA and Rac1 exhibit distinct functional and structural characteristics that underpin their specific roles in cellular processes. RhoA is primarily involved in the formation of stress fibers and focal adhesions, contributing to cell contraction and motility[1,2]. Rac1, in contrast, regulates the formation of lamellipodia and membrane ruffles, promoting cell spreading and migration[3,4]. Our

findings, as illustrated in Figure 5.1, indicate that the major differences in sequence and structure between RhoA and Rac1 are located in the switch I (SW-1) and poly-basic region (PBR). The SW-1 region is crucial for interaction with GEFs and GAPs, while the PBR influences membrane association and subcellular localization. These structural differences result in distinct interaction profiles with regulatory proteinslike RhoGDI and effectors, thereby driving specific cellular outcomes for RhoA and Rac1. In the previous chapter, we detailed the regulatory mechanisms of Rho GTPases by RhoGDI, emphasizing the role of the "phosphorylation code" in modulating their interactions and activity. We think this minor difference in sequence and structure play a major role in specificity of the "phosphorylation code."

This study investigates the impact of phosphorylation on the Rac1-RhoGDI complex, focusing on the role of histidine protonation states. Using constant pH molecular dynamics (CpHMD) simulations, the research explored how phosphorylation at Ser101 and Ser174 in RhoGDI alters the protonation states of key histidine residues. These changes were linked to structural disruptions and increased flexibility in the Rac1-RhoGDI complex. The histidine residues, particularly HIS23 and HIS189, showed significant changes in protonation states due to phosphorylation, impacting the stability and binding affinity of the complex. Unbiased molecular dynamics (MD) simulations, incorporating corrected protonation states, revealed increased dynamic behavior in the phosphorylated state, particularly in the N-terminal region of RhoGDI and the SW-1 region of Rac1. Binding energy calculations confirmed that this increased flexibility reduces the binding energy between Rac1 and RhoGDI, thereby facilitating dissociation. Overall, this study highlights the critical role of histidine protonation states in phosphorylation-induced dissociation of the Rac1-RhoGDI complex, advancing our understanding about the regulatory mechanism of the "phosphorylation code".

5.2 Methodology

Structural Modeling and Simulation Preparation

The crystal structure of the Rac1-RhoGDI complex (PDB ID: 1HH4) was used as the basis for modeling and simulation (Figure 5.2). Missing residues and atoms were modeled using the pdbfixer utility of the OpenMM package[5]. The phosphorylated system was prepared using the PyTMs plugin of PyMOL[6]. Two different systems were constructed for this study:

1. WT: Rac1-GDP in complex with RhoGDI.

2. SP: Rac1-GDP in complex with phosphorylated Ser101 (pSer101) and Ser174 (pSer174) in RhoGDI.

Constant pH MD Simulations

To determine the protonation states of titratable residues (ASP, GLU, ARG, LYS, and HIS), we performed 1000 ns constant pH molecular dynamics (CpHMD) simulations for both WT and SP systems. The multisite representation was used to model the protonation states of these residues:

- Two λ-coordinates for ASP and GLU residues.

- Three λ-coordinates for the three protonation states of the imidazole side chain in HIS residues.

The CpHMD simulations were conducted using GROMACS software[7] with the Charmm36m force field and TIP3P water model[8,9]. The protonation states of ASP, GLU, ARG, and LYS residues remained unchanged upon phosphorylation, while HIS residues exhibited significant changes.

Unbiased Molecular Dynamics Simulations

Five independent 1 μs long MD simulations were conducted for each system, utilizing corrected protonation states from the CpHMD simulations. These simulations employed GROMACS 2022.5[10] software with the updated CHARMM all-atom force field (version 36m) and the TIP3P water model[8,9]. The protocol included energy minimization of the solvated proteins using the steepest descent algorithm, followed by production runs in the NPT ensemble with a modified Berendsen thermostat[11] at 300 K and a Parrinello-Rahman barostat[12] at 1 bar. Long-range electrostatic interactions were computed using the Particle Mesh Ewald summation[13] method with a grid spacing of 0.16 nm and fourth-order cubic interpolation. A cutoff distance of 1 nm was applied for short-range electrostatic and van der Waals interactions, and all covalent bonds were constrained using the LINCS algorithm[14].

Binding Energy Calculations

The binding energy between Rac1 and RhoGDI was calculated using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method[15]. The gmx_MMPBSA software[16] was used for this purpose. Detailed methodology for MM-PBSA calculations is provided in the methodology chapter.

5.3 Results and Discussions

5.3.1 Impact of Phosphorylation of Protonation State of Histidine

To understand the impact of phosphorylation on the protonation states of histidine residues within the Rac1-RhoGDI complex, we performed 1 μs constant pH molecular dynamics (CpHMD) simulations. Histidine residues can exist in three protonation states: (i) both δ and ϵ nitrogens are protonated (λ_1), (ii) δ nitrogen is protonated (λ_2), and (iii) ϵ nitrogen is protonated (λ_3).

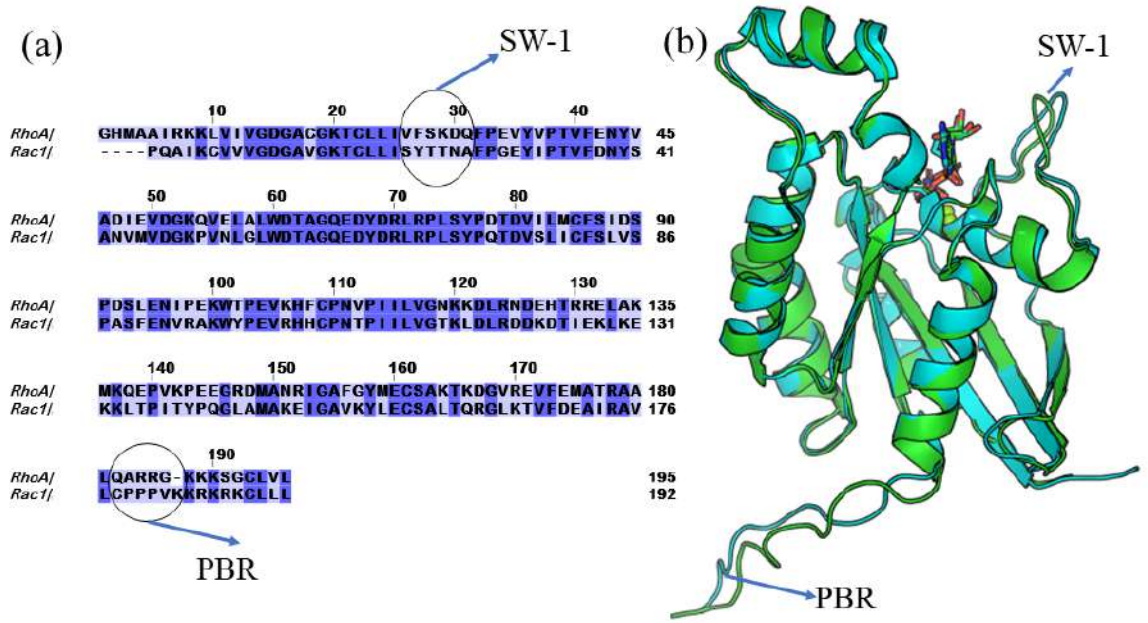

Figure 5.1: (a) Sequence alignment between RhoA and Rac1. Conserved residues are highlighted in blue, with notable differences in the Switch 1 (SW-1) and Polybasic Region (PBR) circled. (b) Superimposed crystal structure of RhoA (cyan) and Rac1 (green), showing structural dissimilarities in the SW-1 and PBR regions. These regions are crucial for interactions with the RhoGDI protein and may contribute to the specificity of the phosphorylation code.

Figure 5.3 (a) shows the population distribution of these states for both the wild-type (WT) and the phosphorylated (SP) systems. The data indicate that phosphorylation significantly alters the protonation distribution, particularly for HIS23 and H189 of RhoGDI. In the phosphorylated system (SP), HIS23 and H189 predominantly remain in the λ_1 state, where both δ and ϵ nitrogens are protonated. In contrast, in the WT system, HIS23 mainly exists in the λ_2 state (δ protonated), and HIS189 predominantly adopts the λ_3 state (ϵ protonated).

These findings suggest that phosphorylation induces a shift in the protonation state equilibrium of specific histidine residues. This shift could be attributed to the electrostatic and structural changes induced by the addition of phosphate groups, which may alter the local environment and the pK_a values of these histidine residues.

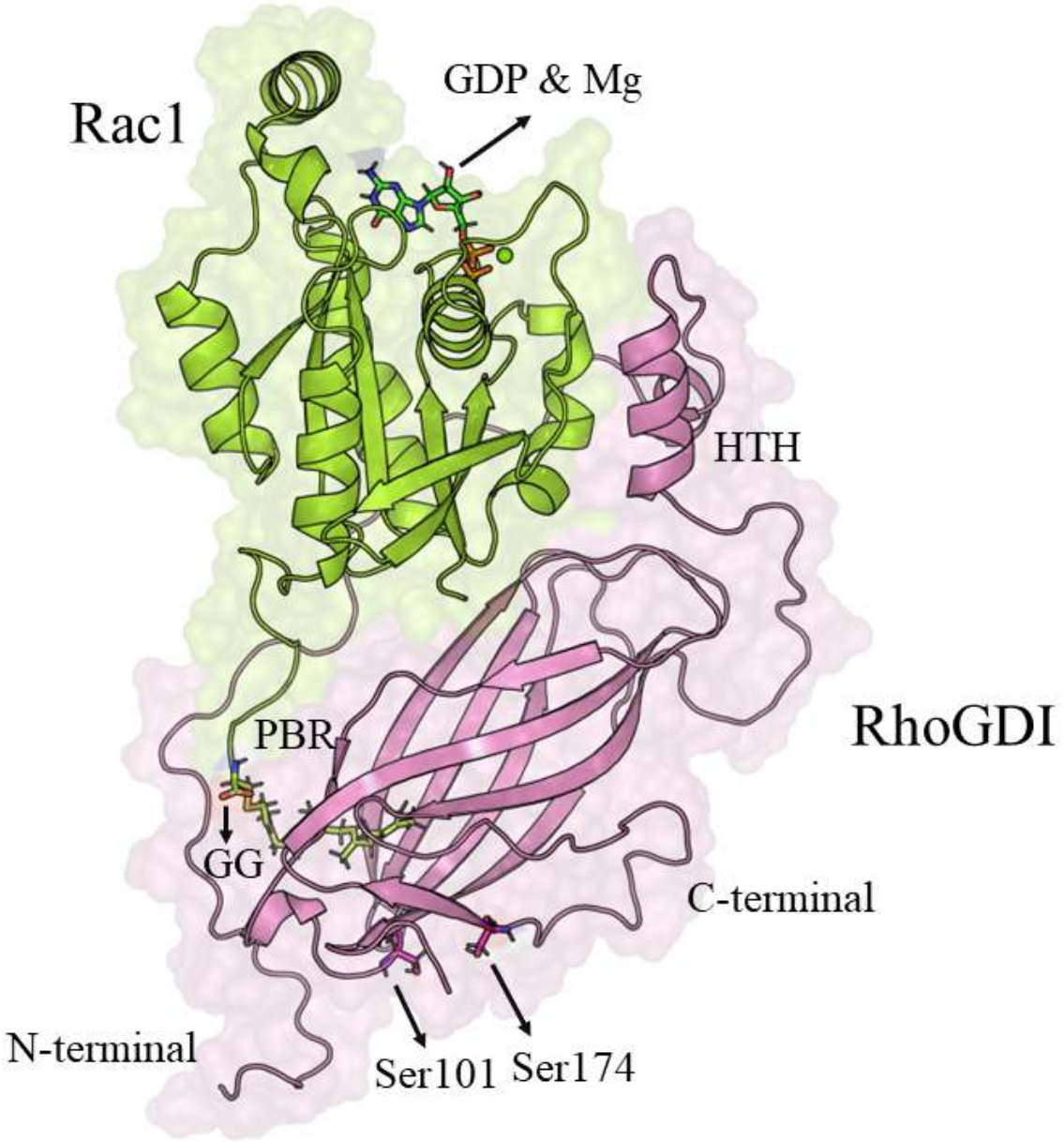

Figure 5.2: Crystal structure of the Rac1-RhoGDI complex (PDB ID: 1HH4). Rac1 (residues 1–189) is shown in lime and RhoGDI (residues 1–204) in pink. The phosphorylation sites (Ser101 and Ser174), GDP, and the geranylgeranyl (GG) moiety are depicted in stick representation. Mg^{2+} is illustrated as a green sphere. Important regions such as PBR (Rac1), N–terminal, C–terminal, and HTH (GDI) are highlighted.

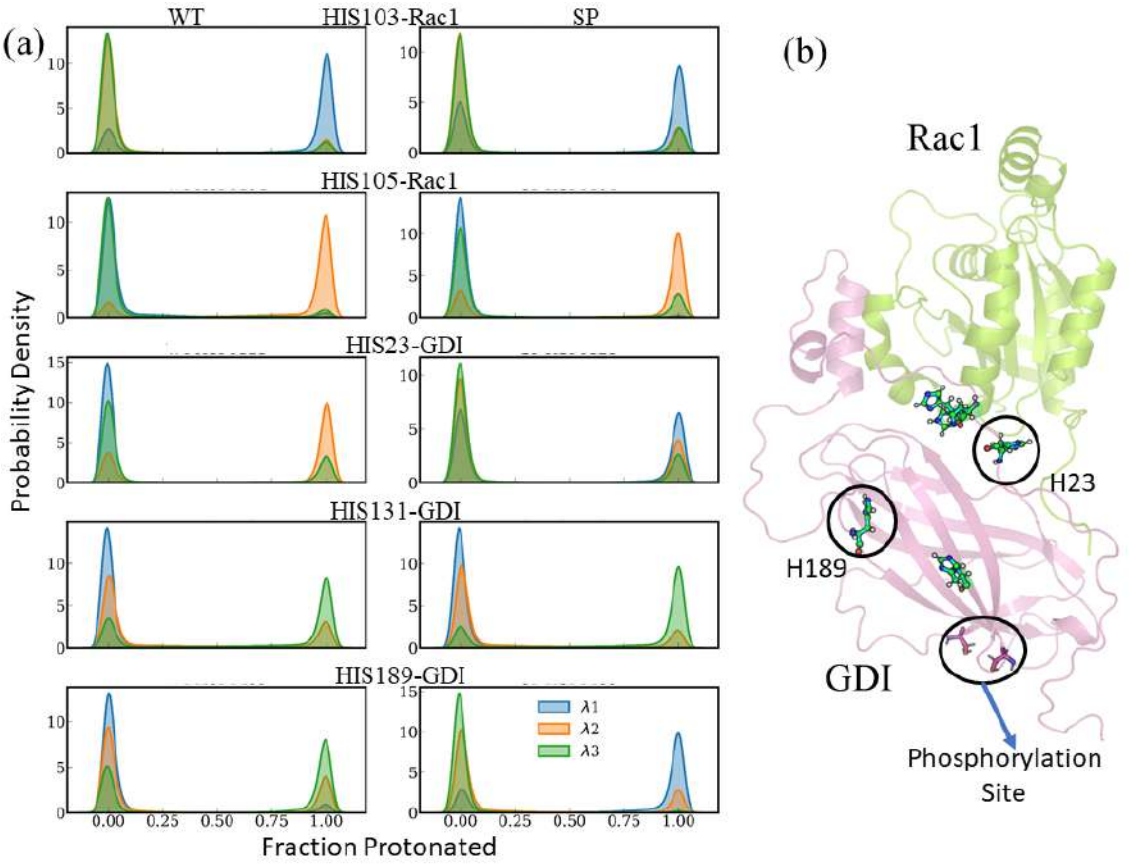

Figure 5.3: Histidine can remain in three different protonation states: (i) both δ and ϵ nitrogens are protonated (represented by λ_1), (ii) δ nitrogen is protonated (represented by λ_2), and (iii) ϵ nitrogen is protonated (represented by λ_3). (a) Population distribution of these states for both WT and SP systems. The distribution plot shows that phosphorylation changes the population distribution of the different histidine protonation states, particularly for HIS23 and H189 of RhoGDI. In the phosphorylated system, both HIS23 and H189 remain in λ_1 (i.e., both δ and ϵ nitrogens are protonated), whereas in the WT system, HIS23 remains δ protonated and HIS189 remains ϵ protonated. (b) Positions of the HIS residues on the protein surface. The phosphorylated residues and HIS23 and H189 residues are highlighted. Rac1 is shown in lime and RhoGDI in pink.

We have performed multiple simulations (five independent simulation each of which is 1μs long) for both WT and phosphorylated (SP) systems with corrected protonation states obtained from constant pH molecular dynamics simulations. The resulting data were analyzed to elucidate the molecular mechanisms underlying dissociation.

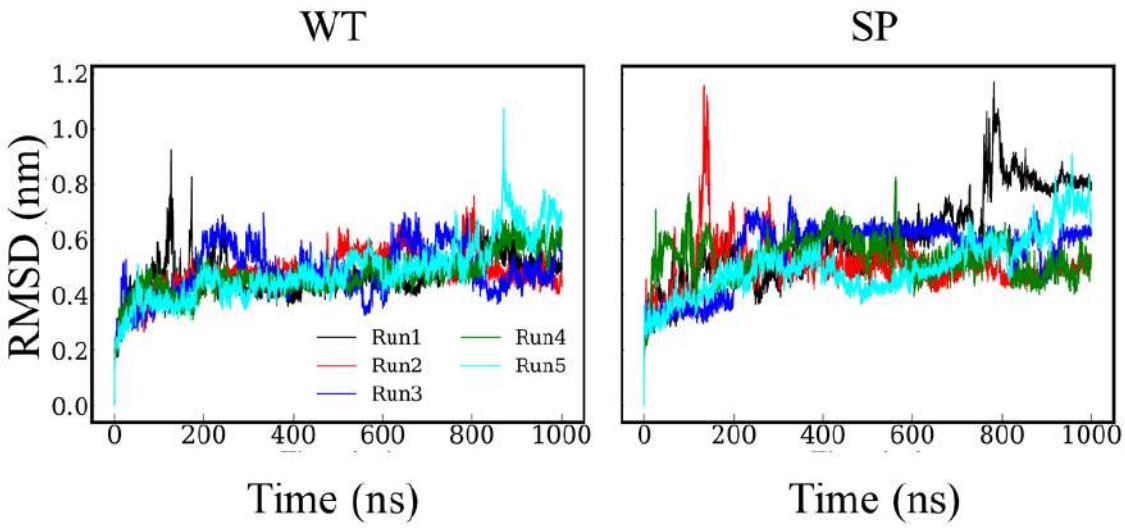

Figure 5.4: Root Mean Square Deviation (RMSD) analyses for five independent runs. For SP the RMSD is slightly higher indicating less structural stabilization.

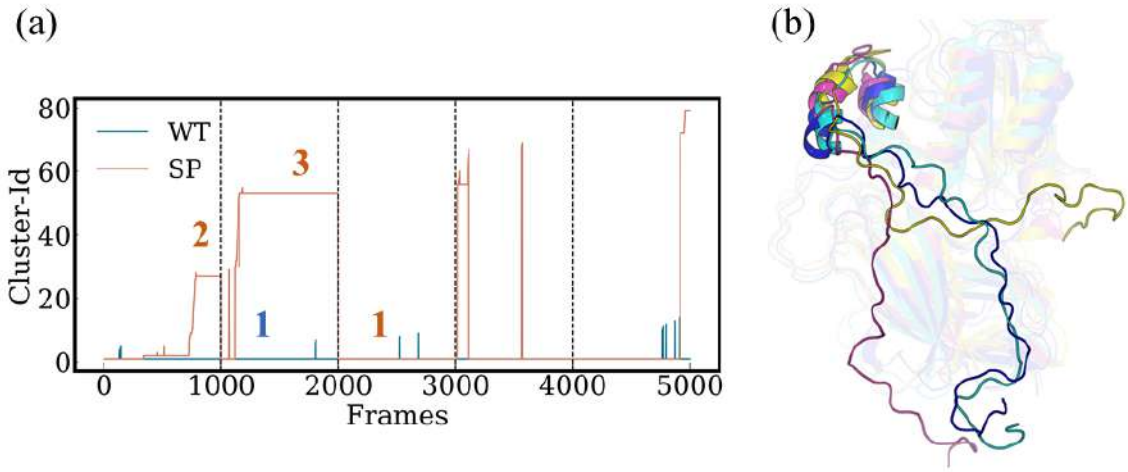

Figure 5.5: (a) Cluster analysis of WT and SP systems. The y-axis represents the cluster ID, and the x-axis represents the frames. Each frame corresponds to 1 ns. For the WT system (blue), there is only one major cluster, indicating structural stability. For the SP system (red), three major clusters are observed, indicating greater structural variability upon phosphorylation. (b) Superimposed structures of the WT major cluster (blue) and the top three clusters for the SP system (cyan, pink, yellow), illustrating the structural disruption caused by phosphorylation.

5.3.2 Phosphorylation-Induced Structural and Dynamical Changes in the Rac1-RhoGDI Complex

To evaluate the structural stability of the Rac1-RhoGDI complex, root mean square deviation (RMSD) analyses were performed on both WT and phosphorylated (SP) systems. Figure 5.4 illustrates the RMSD fluctuations across five independent runs. The WT system exhibits smaller fluctuations compared to the SP system, indicating a higher structural stability in the absence of phosphorylation.

The clustering analysis of the WT and SP systems, as shown in Figure 5.5 a, reveals significant differences in their structural stability. Using 1000 frames (saved at 1 ns intervals) from each of the five independent simulations, and an RMSD cut-off of 0.2 nm, we observed distinct clustering patterns. For WT system there is only one major cluster throughout the simulation, indicating a stable structure. In contrast, the SP system exhibits three major clusters, suggesting that phosphorylation introduces significant structural variability.

This observation aligns with the RMSD analysis, where the SP system demonstrated larger RMSD fluctuations compared to the WT system(Figure 5.4). These findings suggest that phosphorylation at Ser101 and Ser174 significantly disrupts the structural integrity of the Rac1-GDI complex, leading to increased conformational diversity.

The superimposed structures [Figure 5.5 (b)] further illustrate the extent of this disruption. The WT major cluster (blue) maintains a relatively consistent conformation, while the top three clusters of the SP system (cyan, pink, yellow) show considerable deviation. This structural disruption likely contributes to the observed decrease in binding energy and the increased flexibility of the N-terminal of GDI, facilitating dissociation. To understand how phosphorylation alters the dynamics of the Rac1-GDI complex, we calculated the residue-wise change in RMSF. The analysis, shown in Figure 5.6 (a), indicates that phosphorylation significantly increases the overall flex-

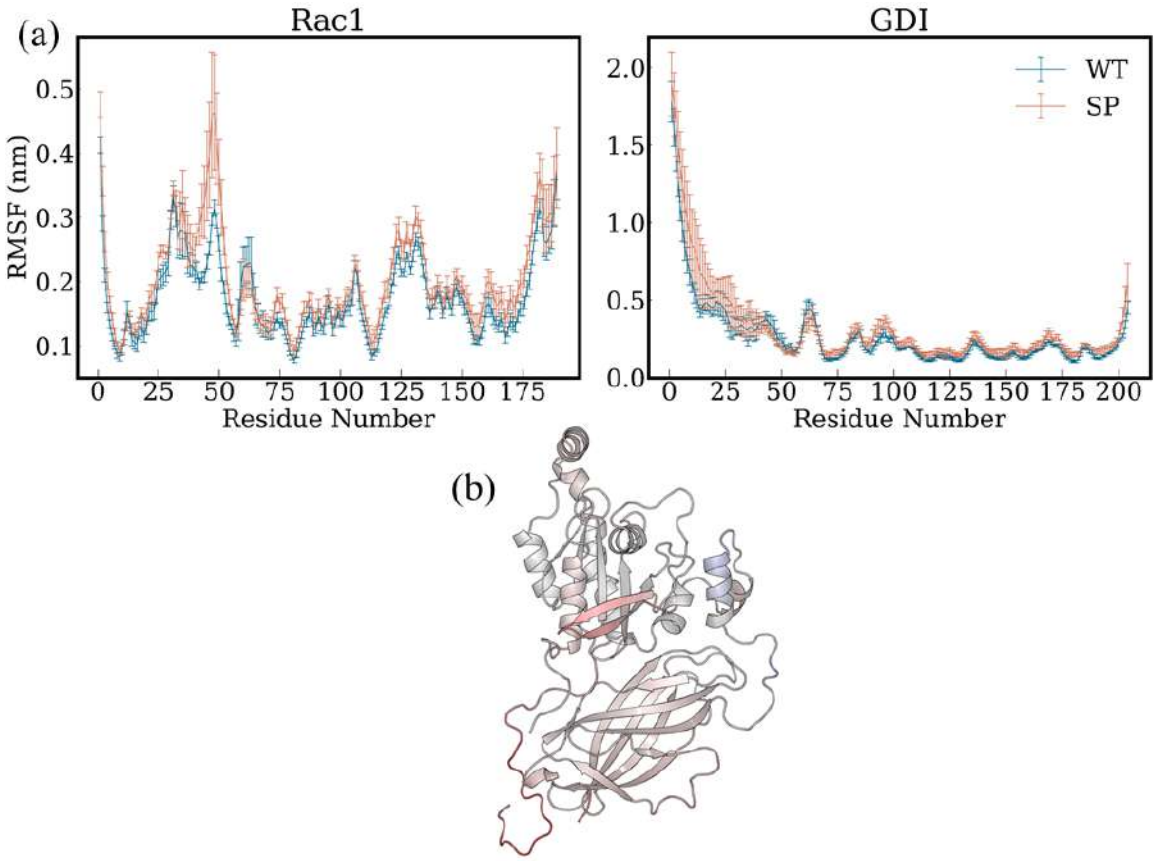

Figure 5.6: (a) RMSF analysis for Rac1 (left) and GDI (right) in WT (blue) and SP (red) systems. The y-axis represents RMSF (nm) and the x-axis represents residue number. Phosphorylation increases the flexibility of residues, particularly in the N-terminal region of RhoGDI (residues 1-30) and the SW-1 region of Rac1 (residues 20-40). (b) The protein structure colored based on ΔRMSF (ΔRMSF = RMSF_{SP} - RMSF_{WT}). Red indicates an increase in flexibility (ΔRMSF > 0), and blue indicates increased rigidity (ΔRMSF < 0).

ibility of the system. This increase in flexibility is most prominent in the N-terminal region of RhoGDI (residues 1-30) and the SW-1 region of Rac1 (residues 20-40).
In Figure 5.6 (b), the protein structure is colored according to ΔRMSF values. Red regions, indicating positive ΔRMSF values, show increased flexibility, while blue regions, indicating negative ΔRMSF values, show increased rigidity after phosphorylation. The overall trend suggests that phosphorylation causes a general increase in the dynamic behavior of the Rac1-GDI complex.

The observed increase in flexibility, particularly in critical regions such as the N-terminal of RhoGDI and SW-1 of Rac1, may facilitate the dissociation of the complex. This increased flexibility could be attributed to changes in the electrostatic network of the system induced by phosphorylation. Such changes likely alter the protonation states of histidine residues, further contributing to the dissociation.

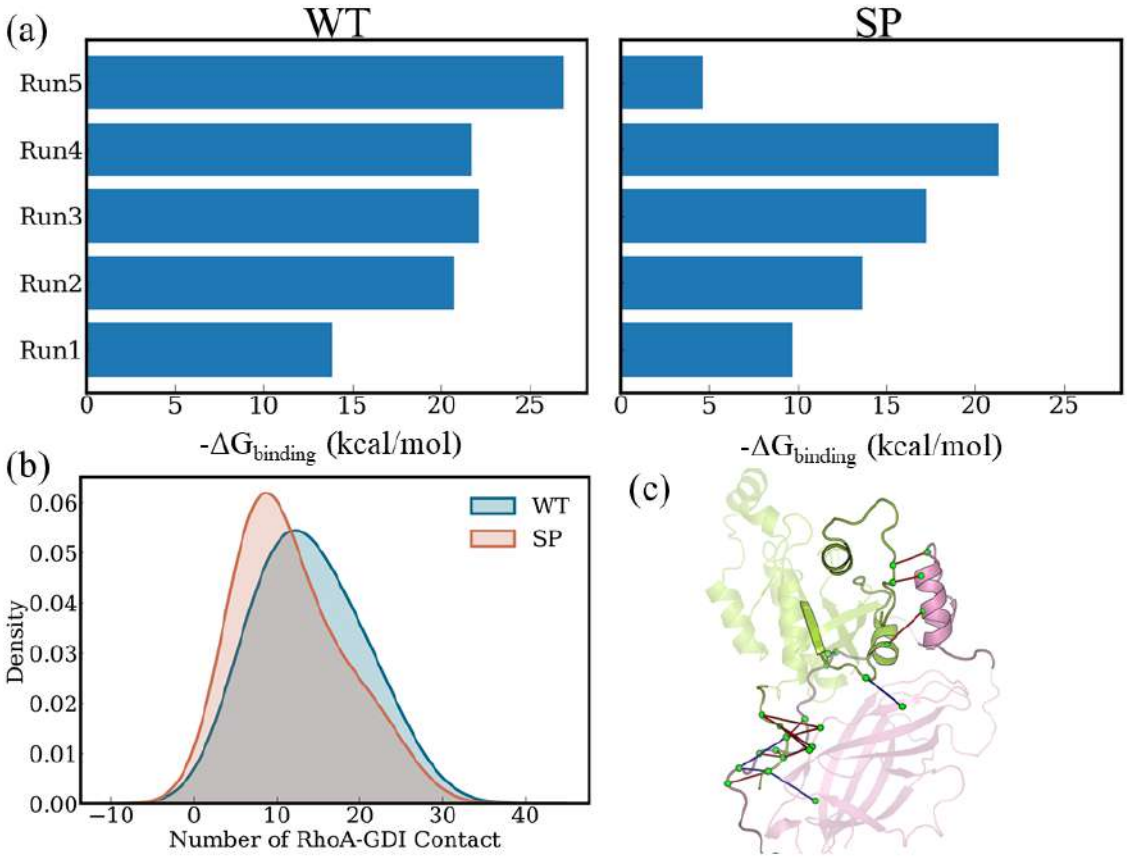

Figure 5.7: (a) Binding energy ($-\Delta G_{binding}$) for five independent simulations of WT (left) and SP (right) systems. Phosphorylation significantly decreases binding energy. (b) Distribution of the number of contacts between Rac1 and GDI, showing fewer contacts in the SP system. (c) Hydrogen bond occupancy change (ΔHB). Red lines indicate decreased occupancy, and blue lines indicate increased occupancy after phosphorylation.

5.3.3 Changes in Binding Energy and Interactions between Rac1 and GDI

To determine whether phosphorylation affects the binding energy between Rac1 and GDI, we calculated the binding energy for both WT and SP systems using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method. Figure 5.7 (a) shows the binding energy for five independent simulations, revealing a significant decrease in binding energy upon phosphorylation. This decrease is supported by Figure 5.7 (b), which shows a reduction in the number of contacts between Rac1 and GDI in the SP system compared to the WT system.

Hydrogen bond analysis further elucidates the mechanism behind this decreased binding energy. Figure 5.7 (c) shows the change in hydrogen bond occupancy (ΔHB) between Rac1 and GDI, with red lines indicating decreased occupancy and blue lines indicating increased occupancy after phosphorylation. The significant reduction in hydrogen bonds in the phosphorylated system suggests that these weakened interactions contribute to the decreased binding energy.

Figure 6.7 represents all the Rac1-GDI contacts that form during simulations, illustrating the extensive network of interactions that stabilize the complex. These contacts shown in this figure were used as features for the TICA analysis. The blue spheres represent the Cα atoms of residues that form contacts, with the red lines indicating the contacts.

To further characterize the conformational changes induced by phosphorylation, we performed time-lagged independent component analysis (TICA) using various Cα-Cα distances within the Rac1-GDI complex that form contacts during the simulations. The free energy surfaces along the two slowest components (IC-1 and IC-2) are presented in Figure 6.13 (a).

The WT system predominantly resides in a single free energy minimum (1), indicating a stable conformational state. In contrast, the SP system exhibits a more diverse conformational landscape, populating three distinct minima (1, 2, and 3). This highlights the increased conformational flexibility introduced by phosphorylation.

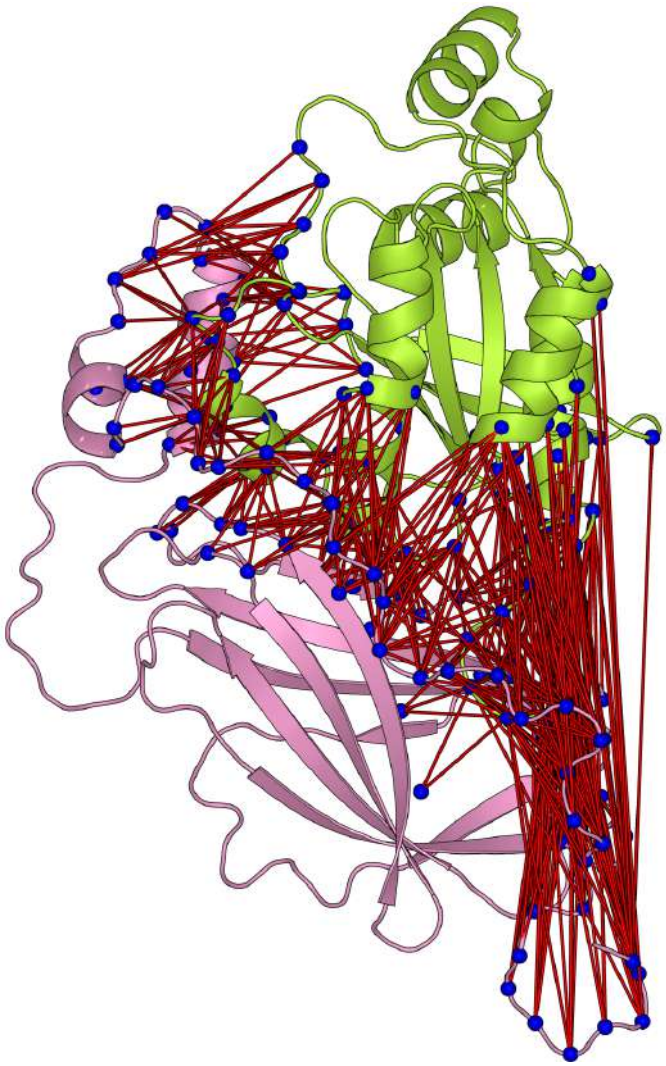

Figure 5.8: The figure illustrates the contacts formed between Rac1 (green) and GDI (pink) during the simulation. The blue spheres represent the Cα atoms of the residues involved in these contacts. All the Cα-Cα distances depicted are used as a feature set for performing time-lagged independent component analysis (TICA). The red lines denote the interactions, highlighting the extensive network of contacts that stabilize the Rac1-GDI complex.

To understand the structural differences between these minima, we extracted representative snapshots from each of the three minima for the SP system and compared them with the WT structure. Figure 6.13 (b) shows these structural snapshots, with green, cyan, and magenta representing the structures from minima 1, 2, and 3, re-

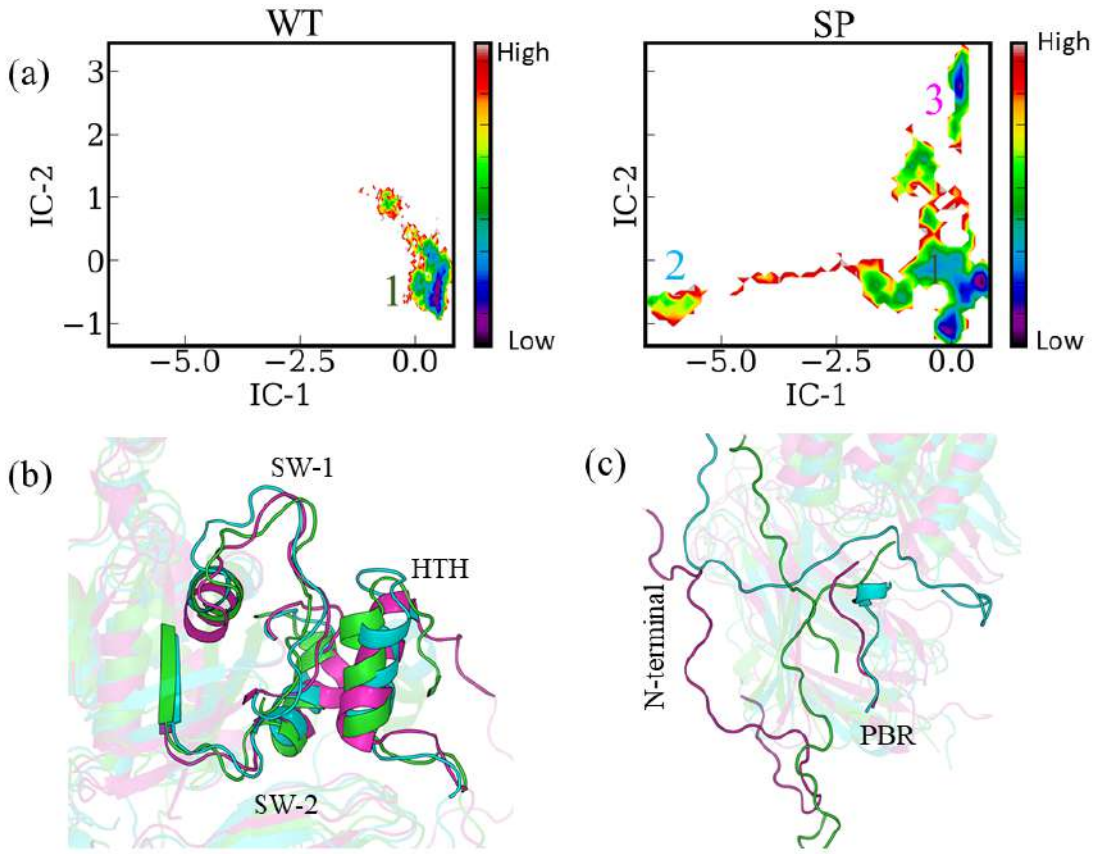

Figure 5.9: (a) Free energy surfaces along the first two slowest components (IC-1 and IC-2) for WT and SP systems. The energy is color-coded from low (blue) to high (red). The WT system predominantly occupies a single minima (1), while the SP system populates three distinct minima (1, 2, 3). (b) Structural snapshot showing structural variation in the switch (SW1, SW2) region and HTH region. (c) Snapshot showing structural variation near the N-terminal region.

spectively. The WT system's structure aligns closely with minima 1 of the SP system, indicating that the unphosphorylated complex is more structurally constrained.

Significant structural deviations are observed in the SP system, particularly in the SW1 and SW2 regions of Rac1, and the N-terminal region of GDI. In minima 2 and 3, the SW1 region moves away from the HTH region, and the N-terminal region of GDI shifts away from the PBR region. These movements are indicative of the destabilizing effects of phosphorylation.

The increased flexibility and structural deviations observed in the SP system correlate

with the reduced binding energy between Rac1 and GDI. The displacement of key regions involved in protein-protein interactions likely contributes to the decrease in binding affinity, as observed in the binding energy calculations and contact analyses.

5.4 Conclusion

This study provides detailed insights into the molecular mechanisms governing the phosphorylation-induced dissociation of Rac1 from RhoGDI. Through extensive constant pH molecular dynamics (CpHMD) simulations, we elucidated how phosphorylation at Ser101 and Ser174 of RhoGDI influences the protonation states of key histidine residues and subsequently affects the structural integrity of the Rac1-RhoGDI complex.

Our findings demonstrate that phosphorylation significantly alters the protonation states of histidine residues, particularly HIS23 and HIS189, which leads to substantial conformational changes and increased flexibility within the complex. These structural perturbations, especially in the N-terminal region of RhoGDI and the SW-1 region of Rac1, contribute to a decrease in binding affinity between Rac1 and RhoGDI.

The binding energy calculations using the MM-PBSA method corroborated that phosphorylation-induced changes reduce the binding energy, thus facilitating the release of Rac1 from RhoGDI. This work underscores the critical role of histidine protonation states in mediating the effects of phosphorylation on protein-protein interactions.

By providing a comprehensive understanding of the structural and thermodynamic consequences of phosphorylation in the Rac1-RhoGDI complex, this research advances our knowledge of the regulatory mechanisms involved in Rac1 signaling. These insights pave the way for developing therapeutic strategies aimed at modulating Rac1 activity, with significant implications for diseases involving aberrant Rac1 signaling pathways, such as cancer and other pathological conditions.

Bibliography

[1] Hall, A. (1998) Rho gtpases and the actin cytoskeleton. *Science* **279**, 509–514.

[2] Ridley, A. J & Hall, A. (1992) The small gtp-binding protein rho regulates the assembly of focal adhesions and actin stress fibers in response to growth factors. *Cell* **70**, 389–399.

[3] Ridley, A. J, Paterson, H. F, Johnston, C. L, Diekmann, D, & Hall, A. (1992) The small gtp-binding protein rac regulates growth factor-induced membrane ruffling. *Cell* **70**, 401–410.

[4] Kraynov, V. S, Chamberlain, C, Bokoch, G. M, Schwartz, M. A, Slabaugh, S, & Hahn, K. M. (2000) Localized rac activation dynamics visualized in living cells. *Science* **290**, 333–337.

[5] Eastman, P, Swails, J, Chodera, J. D, McGibbon, R. T, Zhao, Y, Beauchamp, K. A, Wang, L.-P, Simmonett, A. C, Harrigan, M. P, Stern, C. D, Wiewiora, R. P, Brooks, B. R, & Pande, V. S. (2017) Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **13**, e1005659.

[6] Schrödinger, LLC. (2015) The PyMOL molecular graphics system, version 1.8.

[7] Aho, N, Buslaev, P, Jansen, A, Bauer, P, Groenhof, G, & Hess, B. (2022) Scalable constant ph molecular dynamics in gromacs. *Journal of Chemical Theory and Computation* **18**, 6148–6160.

[8] Klauda, J. B, Venable, R. M, Freites, J. A, O'Connor, J. W, Tobias, D. J, Mondragon-Ramirez, C, Vorobyov, I, MacKerell, A. D, & Pastor, R. W. (2010) Update of the charmm all-atom additive force field for lipids: Validation on six lipid types. *The Journal of Physical Chemistry B* **114**, 7830–7843.

[9] Huang, J, Rauscher, S, Nawrocki, G, Ran, T, Feig, M, de Groot, B. L, Grubmüller, H, & MacKerell, A. D. (2016) Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73.

[10] Abraham, M. J, Murtola, T, Schulz, R, Páll, S, Smith, J. C, Hess, B, & Lindahl, E. (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25.

[11] Bussi, G, Donadio, D, & Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**.

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

[13] Essmann, U, Perera, L, Berkowitz, M. L, Darden, T, Lee, H, & Pedersen, L. G. (1995) A smooth particle mesh ewald method. *The Journal of Chemical Physics* **103**, 8577–8593.

[14] Hess, B, Bekker, H, Berendsen, H. J. C, & Fraaije, J. G. E. M. (1997) Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463–1472.

[15] Baker, N. A, Sept, D, Joseph, S, Holst, M. J, & McCammon, J. A. (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences* **98**, 10037–10041.

[16] Valdés-Tresanco, M. S, Valdés-Tresanco, M. E, Valiente, P. A, & Moreno, E. (2021) gmx_mmpbsa: A new tool to perform end-state free energy calculations with gromacs. *Journal of Chemical Theory and Computation* **17**, 6281–6291.

6

Leveraging Bidirectional Nature of Allostery for Drug Discovery

"Science is the poetry of reality."

Richard Dawkins

6.1 Introduction

Proteins serve as essential components of biological processes, orchestrating intricate pathways crucial for sustaining life. With a vast array of proteins controlling diverse physiological functions, protein-protein interactions (PPIs) play a pivotal role in governing these processes. Nevertheless, disruptions in PPIs are associated with various health conditions, highlighting the significance of regulating their dynamics and kinetics for therapeutic interventions[1–5]. Targeting protein-protein interactions (PPIs) is challenging because their large interfaces often lack clearly defined binding sites, making it difficult to identify small-molecule inhibitors. While biologics and

peptidomimetics offer alternatives, the pursuit of orally-administrable small-molecule inhibitors persists. An emerging strategy involves designing allosteric inhibitors that bind distant sites on proteins, altering PPI affinity. These allosteric sites may exist as stable cavities or transient "cryptic sites", necessitating computational approaches for their identification[6]. Allosteric modulation of PPIs represents a novel avenue in drug discovery[7–12]. However, there are many challenges, such as uncertainty about how the allosteric site connects to the functional site and the "cryptic" nature of some sites. Computational methods, such as molecular dynamics simulations, aid in predicting these sites by exploring conformational states and identifying druggable cavities[8,13–17]. This study proposes a conceptual framework leveraging thermodynamic principles to identify allosteric/cryptic sites[8,18,19]. Additionally, proof-of-concept results demonstrate allosteric modulation of a promising drug target for hypercholesterolemia and cardiovascular diseases. Cardiovascular diseases remain a leading global cause of mortality, often attributed to atherosclerosis. Hypercholesterolemia, characterized by elevated LDL-c levels, exacerbates atherosclerosis risk. PCSK9 has emerged as a therapeutic target for hypercholesterolemia, with PCSK9-LDLR interactions contributing to LDLR degradation[20–24]. Targeting PCSK9 or blocking its interactions with LDLR are primary therapeutic goals[25–32].While monoclonal antibodies show promise[33], orally-administrable small-molecule therapeutics offer accessibility, albeit facing challenges due to the large ($\geqslant 500\mathring{A}^2$) and featureless (flat) binding surface of PCSK9-LDLR PPIs. Recent studies report success in allosterically disrupting PCSK9-LDLR PPIs, identifying druggable binding sites distant from the interaction surface[31,34,35]. This study emphasizes the structural plasticity of specific PCSK9 loop regions as potential drug-binding sites, aiming to disrupt PCSK9-LDLR interactions effectively using simple thermodynamic principles.

6.2 Theoretical Background

Bidirectional Nature of Allostery: Interplay Between Conformational and Binding Free Energy

First, we will define the thermodynamic principles that underpin the bidirectional nature of allostery and explore how they can be utilized to locate allosteric hotspots or hidden sites. At its core, this concept suggests that when a substrate or ligand binds at the orthosteric site and induces a conformational change at a distant location on the protein, that distant site must be allosterically linked to the orthosteric site. Consequently, these distinct conformations at the distant site are expected to exhibit different binding affinities for the substrate at the orthosteric site due to the thermodynamic constraints we'll discuss below. Presented in Figure 6.1 is a schematic representation of this concept in a thermodynamic cycle. We posit that the allosteric site can have two distinct conformations (e.g., square: S and triangular: T). The differ-

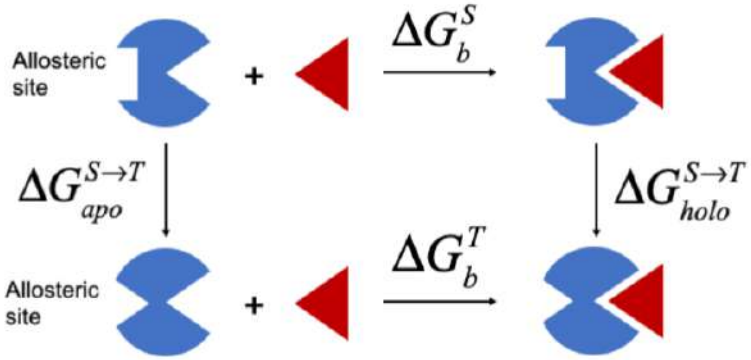

Figure 6.1: A thermodynamic cycle illustrating the substrate (red triangle) binding process to a protein (blue): The protein possesses a distal allosteric site on its left side, capable of adopting two distinct conformations: square (S) and triangular (T). The S and T conformations predominate in the ligand (red triangle) bound and apo states, respectively. The horizontal arrows depict the binding free energy when the allosteric site is in the "S" (ΔG_b^S) or "T" (ΔG_b^T) conformation. Vertical arrows represent the change in conformational free energy transitioning from "S" to "T" state in the absence ($\Delta G_{apo}^{S\to T}$) and presence ($\Delta G_{holo}^{S\to T}$) of the bound substrate.

ence in free energy between these conformations, and thus their relative populations,

hinges on the ligand binding status, represented by $\Delta G_{holo}^{S\rightarrow T}$ and $\Delta G_{apo}^{S\rightarrow T}$, which denote the free energy differences between the S and T states for the holo (ligand-bound) and apo (ligand-free) states, respectively. Additionally, the ligand binding free energy is influenced by the conformational state of the allosteric site, expressed as ΔG_b^S and ΔG_b^T for the S and T states, respectively. This yields the following relation:

$$\Delta G_b^T = \Delta G_b^S + (\Delta G_{holo}^{S\rightarrow T} - \Delta G_{apo}^{S\rightarrow T}) \tag{6.1}$$

It's crucial to recognize that if the T conformation is more stable in the apo state than in the bound state, ensuring that $\Delta G_{holo}^{S\rightarrow T} > \Delta G_{apo}^{S\rightarrow T}$, then $\Delta G_b^T > \Delta G_b^S$. Consequently, the substrate's binding affinity will be reduced for the T conformation, which is more stable in the apo state. The conformational transition from S to T may or may not coincide with a well-defined structural change at the orthosteric binding site. The decrease in binding affinity may arise from either structural modifications or dynamical change (dynamic allostery) at the orthosteric site. Importantly, the thermodynamic cycle guarantees that a change in conformational free energy at the allosteric site due to ligand binding suggests a corresponding difference in binding energy, irrespective of our understanding of the precise mechanism behind the change in binding energy. Therefore, we can confidently devise the subsequent general approach to anticipate or recognize allosteric hotspots across various systems, encompassing challenging protein-protein interactions (PPIs). We execute standard molecular dynamics (MD) simulations, potentially alongside certain enhanced sampling techniques (e.g TMD, Metadynamics) for both the "apo" and "holo" protein states. Subsequently, a comparison of the conformational ensembles between these two sets is conducted. Any noteworthy conformational change observed at a distant location are recognised as potential allosteric sites. Conformations or clusters distinctively prevalent or more abundant in the "apo" system are presumed to exhibit reduced binding affinity for the substrate at the orthosteric site. An inhibitor could be developed to stabilize these "apo"-specific conformations, even in the presence of the substrate (in the "bound" system), thereby diminishing the substrate's binding affinity. While this discourse focuses solely on the conformational

variations (free energy) at the distant site, we will later demonstrate in this chapter that this approach remains effective even in the presence of any designed allosteric inhibitor.

6.3 Computational Methodology

Details of System Preparation and Simulation

We utilized the crystal structure of PCSK9 with PDB ID 4NE9 as the basis for conducting molecular dynamics (MD) simulations. This structure comprises two PCSK9 molecules: one interacting with the EGF-A domain of LDLR and the other existing in the apo state. A list of missing residues in the crystal structure is provided in Table 6.1, which were modeled using the Modeller software[36]. Specifically, for the bound system, the loop 211–222 was modeled, and an ensemble of generated conformations is depicted in Figure 6.2, demonstrating the existence of multiple possible loop conformations. Two representative conformations, namely Conf-A-model and Conf-B-model, were selected for further simulations (Figure 6.3). Validation of these structures was conducted using the ERRAT score[37], with both Conf-A-model and Conf-B-model exhibiting scores exceeding 70, indicating high-quality models suitable for subsequent MD simulations (Table 6.2)[38]. Additionally, validation via the Ramachandran plot revealed that over 90% of residues remained in the most favorable regions (Figure 6.4), analyzed using the PROCHECK program[39]. MD simulations were performed using GROMACS 2019.6[40] with proteins modeled using the Amber99sb-ildn force field[41]. The structures were solvated using the TIP3P water model[42] under periodic boundary conditions. Energy minimization of solvated proteins was achieved using the steepest descent algorithm. NPT ensemble simulations were conducted at 300 K using a velocity rescaling thermostat[43] and Parrinello–Rahman barostat[44] at 1 bar, with a salt concentration of 0.15 M. Long-range electrostatic interactions were calculated using the particle mesh Ewald (PME)

Missing Residues in Apo Structure	Missing Residues in Bound Structure
169-177	169-175
-	213-218
447-453	450-453
573-574	573-574
660-670	660-670

Table 6.1: Missing Residues in the Apo and Bound Structures of Crystal Structure (PDB ID: 4NE9)

Model Structure	ERRAT Score
Conf-A-Model	71.51
Conf-B-Model	76.03

Table 6.2: ERRAT scores for Conf-A and Conf-B model structures

summation method[45], with a cutoff distance of 1 nm for short-range electrostatic and van der Waals interaction. Covalent bonds were constrained using the LINCS algorithm[46], and an integration time step of 2 fs was utilized. Our study employed various types of MD simulation trajectories with distinct objectives. Initially, we ran three replicas of 1–2 μs long trajectories for three systems: apo (without LDLR), bound (with LDLR), and bound structure after removing the LDLR domain from the structure, initializing the apo system in the bound conformation of PCSK9. Conformational clustering of these trajectories revealed significant differences in the loop 211–222 region, which serves as the primary focus of this study. We designate the loop conformation characteristics of the apo and bound states as "Conf-A" and "Conf-B", respectively. Furthermore, we demonstrate that even after removing the ligand starting at the "Conf-B" conformation, PCSK9 remains in this metastable state during our simulation time. Consequently, biased sampling is necessary to generate intermediate conformations between Conf-A and Conf-B for constructing free energy surfaces through Markov State Model.

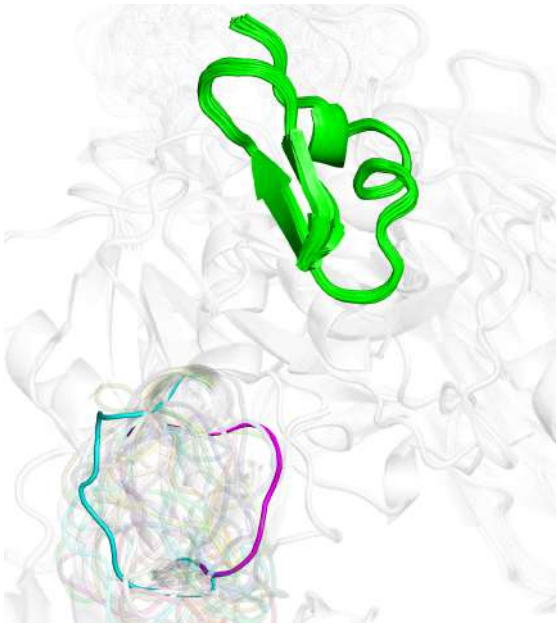

Figure 6.2: The Modeller software was employed to model the missing residues of the 211-222 loop. This loop is capable of adopting multiple conformations, as observed in our MD simulation results. The structures of Conf-A and Conf-B, utilized in subsequent MD simulations, are depicted in cyan and magenta, respectively.

Sampling the Conformational Space using Targeted Molecular Dynamics (TMD)

Targeted Molecular Dynamics (TMD) simulations, a variant of steered MD, utilize the root-mean-square deviation (RMSD) from a reference structure as a collective variable. In this context, TMD was employed to generate intermediary structures between Conf-A and Conf-B, and vice versa. The pulling was executed using the RMSD relative to the target structure as the reaction coordinate, with the pulling force constant adjusted from 10,000 to 18,000 kJ/mol·nm^2. A total of 23 structures were collected along the TMD trajectory, depicted in Figure 6.8, to initiate multiple replicas of standard MD trajectories for sampling the intermediate conformational space, essential for subsequent Markov State Model (MSM) construction.

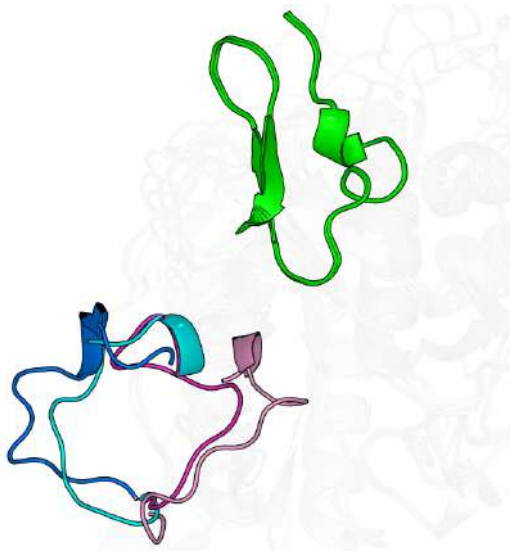

Figure 6.3: The superimposed structures depict modelled loop conformations resembling Conf-A (blue) and Conf-B (pink). The model conformations of Conf-A and Conf-B are illustrated in cyan and magenta, respectively.

Binding Energy Calculation Using Constant Velocity Steered Molecular Dynamics (SMD)

Constant Velocity Steered Molecular Dynamics (SMD) simulations were conducted to compute binding free energy for each metastable state identified from the MSM analysis of the bound system (Conf-A, IM-1, IM-2, Conf-B). From each metastable state, five snapshots were selected to perform constant velocity Steered MD simulations. The reaction coordinate used for pulling was the center of mass distance (d) between the EGF-A domain of LDLR and the "interacting region" of PCSK9, illustrated in Figure 6.19. PCSK9 residues within 6 Å of the EGF-A domain of LDLR during MD simulation were deemed part of the "interacting region".This region contains following residues S153, I154, P155, W156, L158, H193, R194, E195, D224, S225, R237, D238, A239, D367, I369, G370, S372, D374, C375, S376, T377, C378, F379, V380, S381, Q382. A pulling force constant of 5000 kJ/mol·nm^2 was applied, with a pulling

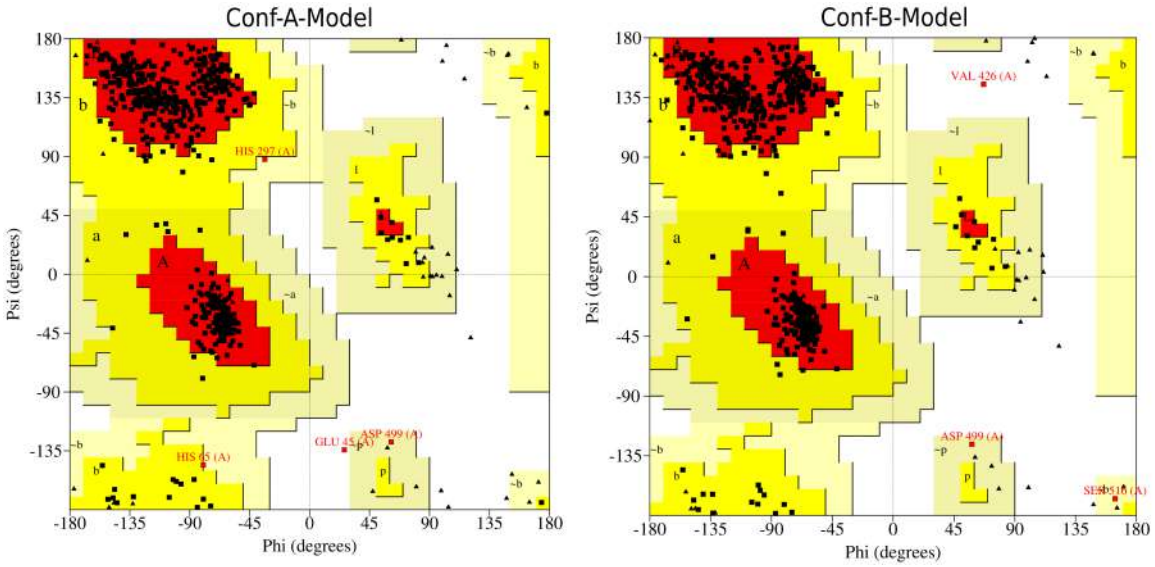

Figure 6.4: Validation of Conf-A and Conf-B model structures using Ramachandran plot : A high-quality model is characterized by over 90% of residues falling within the most favorable regions. Both Conf-A and Conf-B exhibit over 90% residues in the most favorable region, indicating their structural integrity. This validation was conducted using the PROCHECK program.

velocity set at 0.0006 Å/ps. Each of these SMD simulations was 40 ns long. The binding energy was computed using the Jarzynski equality[47,48]:

$$\Delta G = -\beta^{-1} \ln\langle exp^{-\beta W} \rangle \tag{6.2}$$

where W represents the cumulative change in Hamiltonian over time, and $\beta = 1/kT$. Both TMD and SMD simulations were conducted using the PLUMED code (ver. 2.0.2)[49] integrated into GROMACS.

Markov State Model

PyEMMA 2.5 software[50] was used to build the Markov State Model (MSM). As customary, a substantial number of relatively short trajectories (1 μs) encompassing the entire conformational space of interest were essential for MSM construction, linking the conformational substates. From diverse positions in the 2D RMSD space (depicted in Figure 6.8), 23 conformations each for the apo and bound systems

were gathered. One-microsecond trajectories were initiated from each of these conformations, ensuring that the cumulative trajectories (totaling 23 μs for each system) continuously covered the conformational space between Conf-A and Conf-B, as portrayed in Figure 6.8.

To distinguish between the two extreme conformations (Conf-A and Conf-B), 12 loop-protein C_α-C_α distances were employed as the input feature set for MSM construction. The C_α-C_α distance distributions for these residue pairs are illustrated in Figure 6.7. Time-Lagged Independent Component Analysis (TICA)[51–53] was utilized for dimensionality reduction with a lag time of 10 ns (Figure 6.9, upper panel). Subsequently, these lower-dimensional data were discretized using the k-means clustering algorithm into 250 clusters. Based on the implied time scale (relaxation time scale) plot, a lag time of 25 ns was considered for building the final MSM, which reached a plateau at approximately 25 ns (Figure 6.9, lower panel).

Moreover, the time scale plot indicated that three processes were slower than 400 ns for both the apo and bound systems (Figure 6.10, upper panel). Hence, the PCCA++ algorithm was employed to coarse-grain the MSM into 4 metastable states (Figure 6.10, lower panel). For Markov model estimation, a connected data set is indispensable. In Figure 6.11, the transition probability matrix among various microstates for both the apo and bound systems is illustrated, indicating the acquisition of a connected data set for both systems. Finally, the model was validated using the Chapman–Kolmogorov (CK) test (Figure 6.12).

6.3.1 Binding Energy Calculation Using MM-PBSA

One hundred snapshots were obtained for each metastable state from the MSM of the bound state (Figure 6.17). The binding energy between PCSK9 and LDLR was assessed for these metastable states utilizing the MM-PBSA method[54].GMXPBSA[55] software was employed for this purpose.

6.4 Results and Discussion

6.4.1 Binding of LDLR Causes Significant Structural Change in a Distant Loop of PCSK9

PCSK9 is a protein with multiple domains [see Figure 6.5(a)], including a pro-domain (residue 61–152), catalytic domain (residue 153–451), and C-terminal domain (residue 452–682). The catalytic domain primarily interacts with the EGF-A domain of LDLR. The crystal structure utilized in this investigation (PDB ID: 4NE9) encompasses two configurations: (i) PCSK9 bound to the EGF-A domain of LDLR and (ii) the apo form. Following the modeling of the absent residues (as listed in Table 6.1) of the protein and aligning the EGF-A bound and apo structures of PCSK9, it was observed that the loop 211–222 exhibited a substantial difference in conformation between these two states. In Figure 6.5(b), these contrasting conformations are labeled as Conf-A (apo) and Conf-B (bound). To validate that this notable change in loop conformation is not an artifact of homology modeling, we conducted a 2 μs long unbiased molecular dynamics (MD) simulation starting from the respective crystal structures. Additionally, we carried out two independent replica simulations for the same systems. Furthermore, to ascertain whether the EGF-A bound conformation is indeed stable in nature, we removed the EGF-A domain from the bound structure to generate the "apo" state in the originally "bound" conformation (apo state with the loop in Conf-B conformation). Intriguingly, across all three types of trajectories, the loop conformation did not transition between the conformational states, indicating that these are indeed distinct and kinetically resolved at a microsecond time scale. Pairwise hydrogen bond occupancies for these three major clusters are detailed in Table 6.3. The metastability of the loop arises from these specific interactions.Representative structures from the most populated clusters obtained from the apo and bound trajectories are displayed in Figure 6.5(b), where Conf-A and Conf-B represent the conformations in the "apo" and "bound" states, respectively. In Figure 6.5(c), we illustrate the distinct set of

Residue Pairs	Apo in Conf-A (%)	Bound in Conf-B(%)	Apo in Conf-B (%)
E211-R251	63	0	0
E211-R218	79	0	0
E206-R215	59	0	0
R218-D224	60	0	0
D224-R251	29	0	0
E211-C255	31	6	8
E210-K258	45	0	0
E195-A239	2	91	86
E211-R215	30	90	83
E195-D238	3	87	93
D212-Q256	16	75	90
D212-T214	41	91	2
S188-D224	80	0	0
E211-Q256	1	50	12

Table 6.3: Hydrogen bond occupancy between residue pairs in the three most populated clusters derived from three independent MD trajectories for three distinct systems: (i) Apo (Conf-A in the apo state), (ii) Bound (Conf-B in the bound state), and (iii) Apo form achieved by eliminating LDLR (Conf-B in the apo state).

specific interactions (salt bridges and hydrogen bonds) accountable for stabilizing these two distinct conformations. To delineate the conformational space of the loop

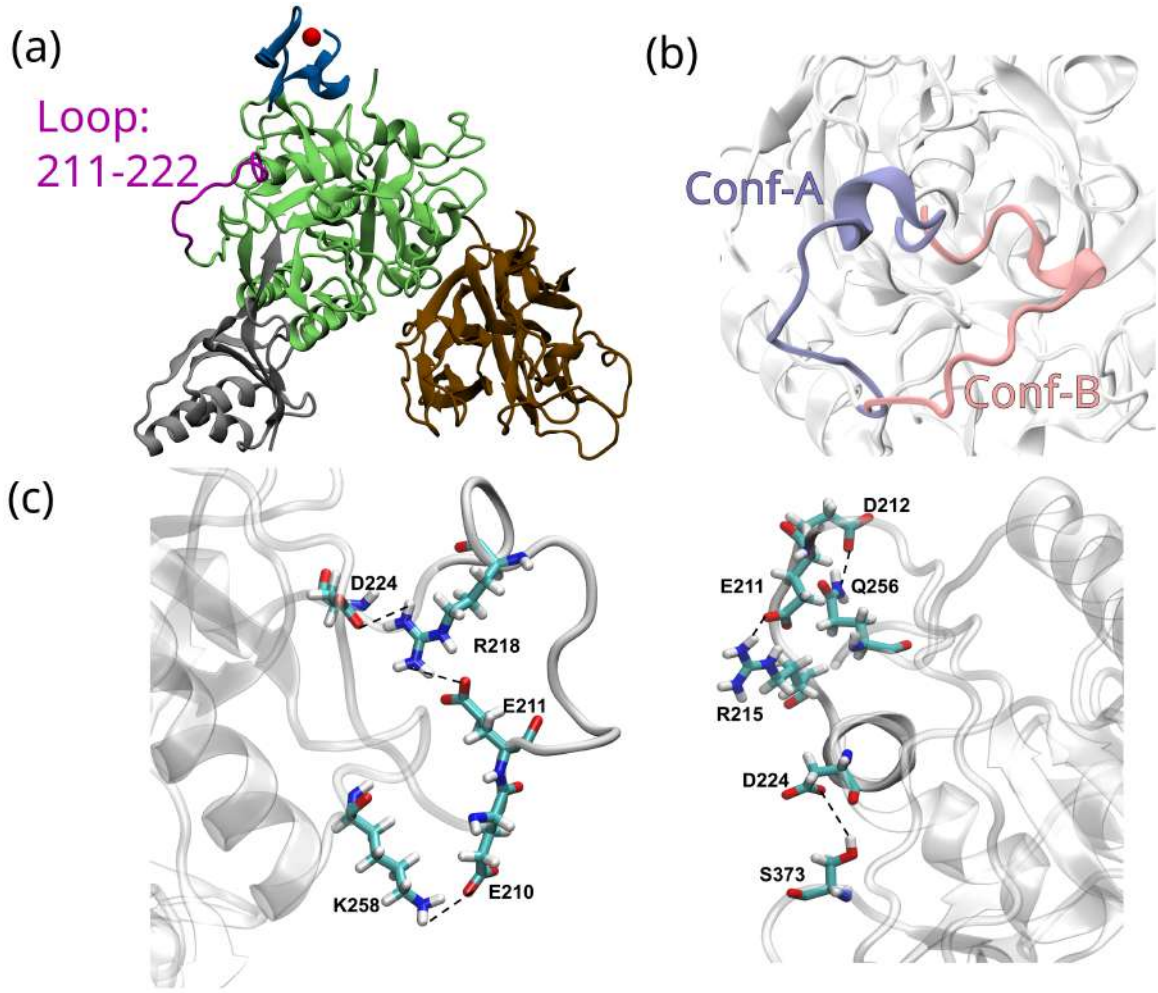

Figure 6.5: (a) Crystal structure of PCSK9 in complex with the EGF-A domain of LDLR (PDB ID: 4NE9): Pro-domain (residues 61–152) shown in silver, Catalytic domain (residues 153–451) in green, C-terminal domain (residues 452–682) in brown, and the EGF-A domain of LDLR in blue. (b) Superposition of representative conformations of loop 211–222 from the most populated clusters extracted from the apo (Conf-A; depicted in ice blue) and EGF-A bound (Conf-B; depicted in pink) trajectories. (c) Visualization of specific interactions (salt bridges and hydrogen bonds) responsible for stabilizing the conformations (Conf-A and Conf-B).

sampled in the MD trajectories, we conducted principal component analysis (PCA). The figure depicts projections of all replica trajectories for all three systems (apo,

EGF-A bound, and starting from the bound conformation after removing the EGF-A domain) onto the first two principal components (PC1 and PC2). It is evident that the conformational space sampled in the "apo" and "bound" states is disparate and widely separated. Remarkably, even after deletion of the EGF-A domain, the

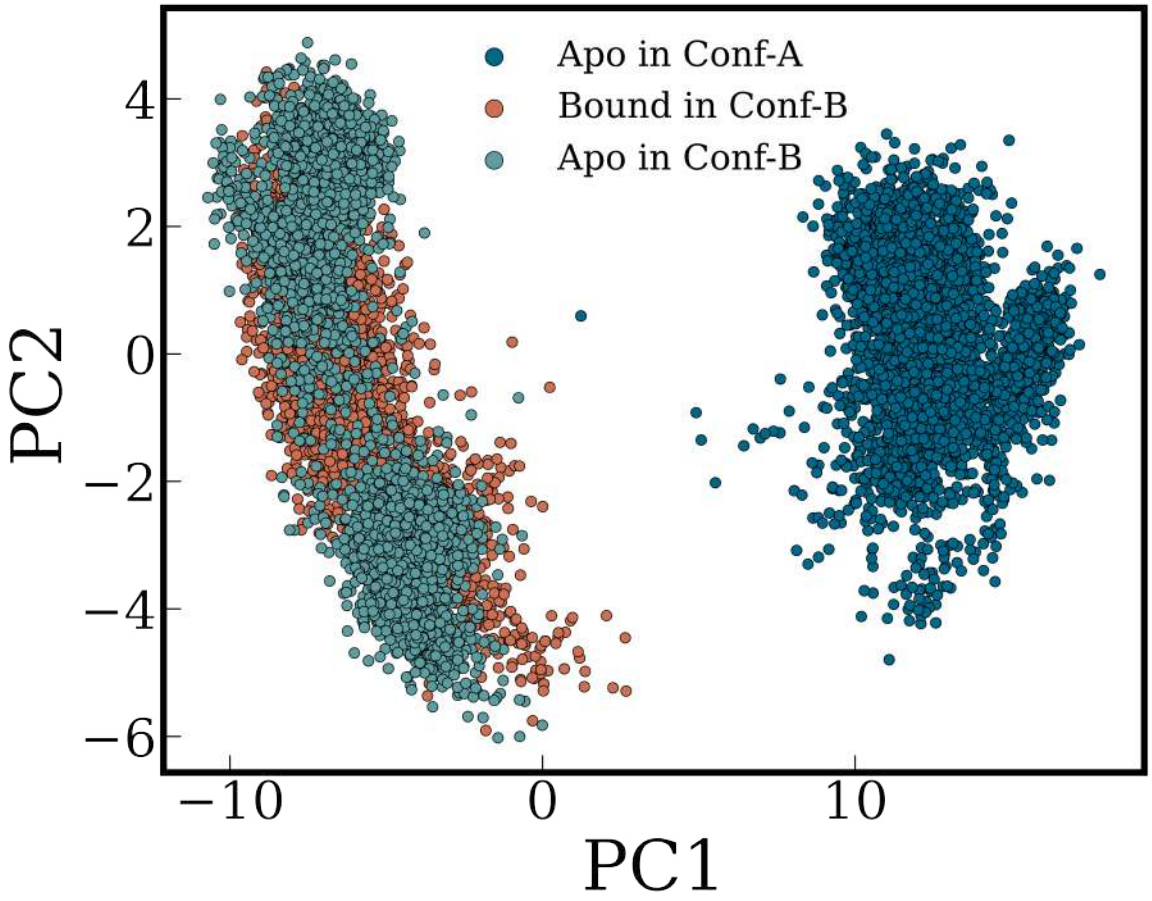

Figure 6.6: Principal Component Analysis (PCA) was employed to characterize the conformational space explored by the loop 211-222 in single MD trajectories for each system. Trajectories projected on the top two PCs (PC1 and PC2) include: (i) apo system (dark blue dots), (ii) bound system (salmon dots), and (iii) apo system in Conf-B (light blue dots) obtained after removing the EGF-A domain from the bound structure.

loop conformation remains close to the Conf-B state throughout the simulation (1–2 μs), without transitioning to the Conf-A conformation, which represents the ground/native state for the apo form. This clearly suggests that the loop 211–222

may exist in at least two distinct conformations separated by a considerable free energy barrier, and binding with EGF-A induces a shift in the population of the conformational space. The free energy barrier between these conformational states must be substantial enough to ensure that the time scale of interconversion would be at least several microseconds, as indicated by the unbiased MD simulations.

6.4.2 Constructing a Markov State Model (MSM) of the Allosteric Site

Based on the aforementioned observations, it can be inferred that the binding of EGF-A is intricately linked to the conformation of a distal site (211–222) through allosteric coupling. However, to assess the comparative stability of the distinct conformational states of the loop in the presence and absence of the EGF-A domain, it is necessary to evaluate the underlying conformational free energy landscape. Given the inherent "disorder" in the loop region, defining an appropriate collective variable (CV) to characterize the sampled conformational space of the loop poses a challenge. In this study, a specific set of C_α-C_α distances (refer to Figure 6.7) was employed as the feature set to construct a Markov State Model (MSM) of the conformational states. MSM has been effectively utilized by various research groups to investigate conformational shifts and allosteric effects[56,57]. The Time-Lagged Independent Component (TIC) projections were utilized to identify the slowest-relaxing collective degrees of freedom for delineating the macrostates and computing the free energy surfaces. The MSM was constructed using 23 independent trajectories (each 1 μs in duration) generated between the extreme structures: Conf-A and Conf-B. Initial configurations for the intermediate structures were generated through targeted MD simulations (refer to Figure 6.8). Further details regarding the construction and validation of the MSM are provided in Figure 6.9-Figure 6.12.

LDLR Binding Influences the Conformational Landscape of Loop 211–222. Figure 6.13 (a) and Figure 6.13 (b) depict the conformational free energy landscapes for

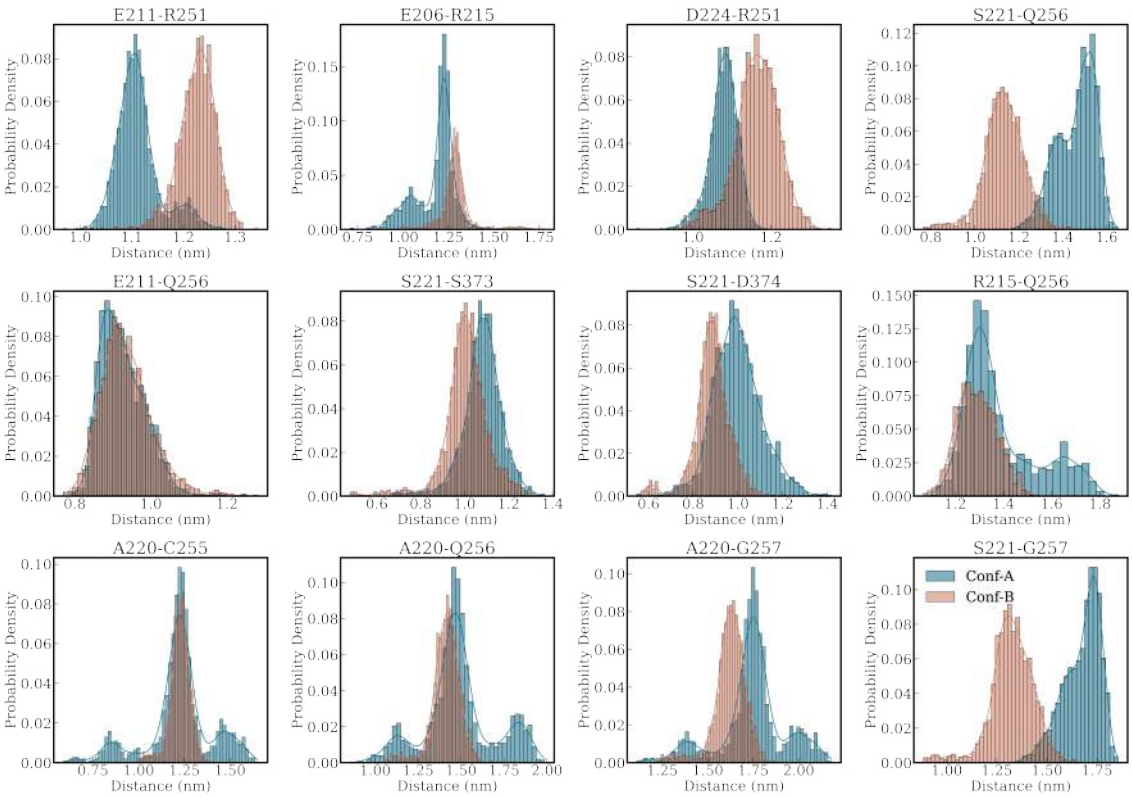

Figure 6.7: Comparison of probability distributions of the C_α-C_α distances for a selection of residue pairs exhibiting notable distinctions between the Conf-A and Conf-B ensembles. These distances effectively discriminate between the diverse conformational states explored by the loop. Consequently, this set of $C_\alpha - C_\alpha$ distances served as input features for constructing the Markov State Model (MSM).

the apo (without EGF-A) and bound (with EGF-A) systems, respectively, projected onto the first two TIC vectors (IC1 and IC2). Notably, both systems exhibit multiple intermediate conformational states. While Conf-A and Conf-B represent the global minima for the apo and bound systems, respectively, the relative stability of specific conformational states undergoes a notable shift upon LDLR binding. The distribution of each metastable conformation is detailed in Table 6.4 (for the Apo system) and Table 6.5 (for the bound system). This suggests a conformational selection model of allostery, wherein various conformational states may coexist in dynamic equilibrium, and ligand binding may induce a shift in the conformational landscape. The

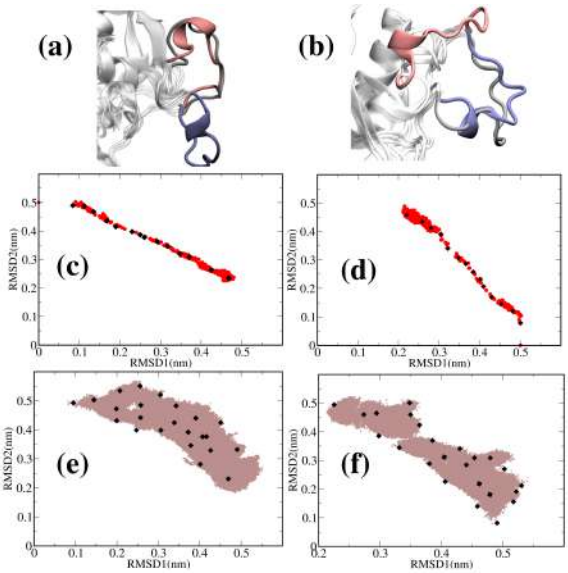

Figure 6.8: (a) and (b) depict the final TMD structure overlaid on the reference structures derived from the alternate starting point. Conf-A is represented in light blue, Conf-B in pink, and the TMD structure in silver. (c) and (d) illustrate conformations sampled along the TMD path for the apo and bound scenarios, respectively. (e) and (f) display the entirety of the conformational space explored in the combined 23 microseconds of trajectory data utilized for constructing the MSM. The initial conformations are denoted by black dots.

computed free energy landscapes affirm that EGF-A binding significantly modifies the conformational landscape/population of distal loop 211–222. In the "bound" system, while Conf-B represents the global minimum as anticipated, there are multiple metastable minima corresponding to different macrostates: IM-1, IM-2, and Conf-A [Figure 6.13 (b)]. These metastable states are accessible to loops 211–222 even in the presence of the EGF-A domain, as depicted in Figure 6.14, indicating kinetically resolved conformational diversity. The transitions between these conformations occur on a microsecond time scale, elucidating why the loop persists in the Conf-B configuration even after the deletion of the EGF-A domain of LDLR. Moreover, Figure 6.15 illustrates that the center of mass (COM) distance between the 211–222 loop and the EGF-A domain of LDLR exceeds 2 nm, suggesting significant spatial separation. Nonetheless, interactions between the loop and LDLR are facilitated by side chains in both Conf-A and Conf-B. In Conf-B, the R218 residue interacts with the S13 residue

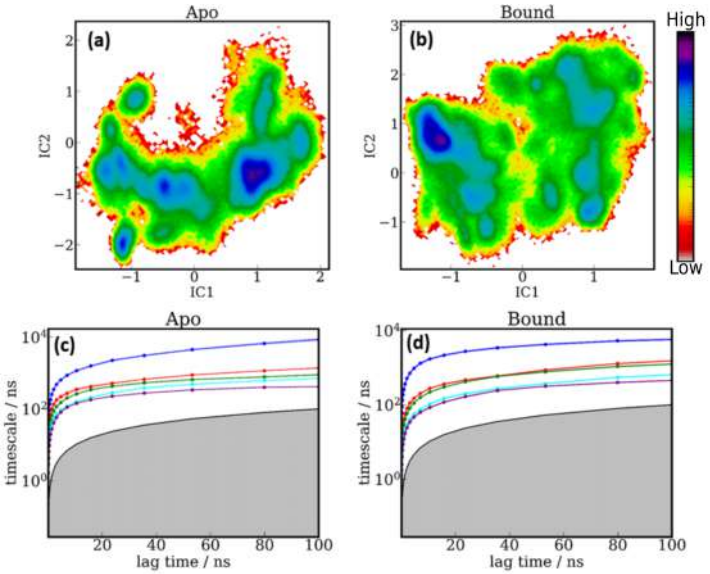

Figure 6.9: (a) and (b) display the 2D-probability density plots of the independent components IC1 and IC2 for the apo and bound systems, respectively. (c) and (d) illustrate the top five slowest implied timescales plotted against the lag time. The Markov State Model was constructed using a lag time of 25 ns.

of the EGF-A domain of LDLR, while in Conf-A, the Q219 residue interacts with the L26 residue of the EGF-A domain of LDLR (Figure 6.16). Specific interactions (hydrogen bonds and salt bridges) stabilizing these intermediate conformational states (Conf-A, IM-1, IM-2, and Conf-B) are given in Table 6.6. This suggests that these metastable conformational states could serve as potential targets in the design of PCSK9 inhibitors.

6.4.3 The Two-Way Influence of Allostery Connects the Binding Affinity to Conformational Preference at the Allosteric Site

Thus far, we have established that EGF-A binding dynamically alters the conformational free energy landscape of a remote loop, which may exist in various distinct,

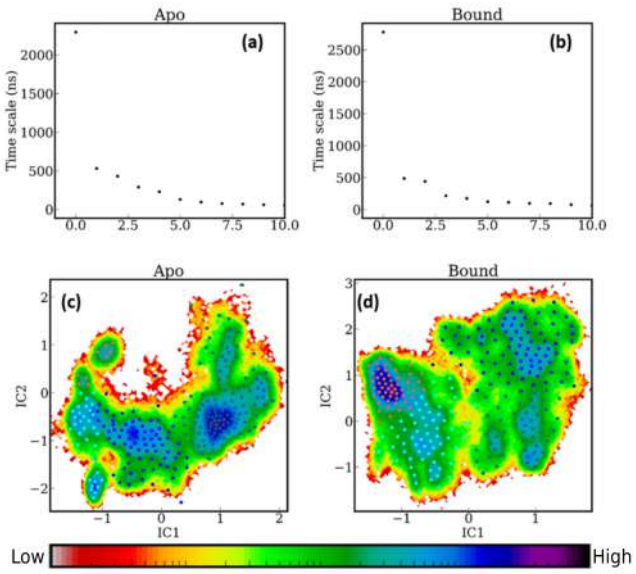

Figure 6.10: (a) and (b) display the timescales of the initial ten slow processes. We have selected the first three slow processes (comprising 4 states) for coarse-graining the Markov State Model (MSM). In (c) and (d), the four metastable states are depicted on the free energy surface for the apo and bound systems, respectively.

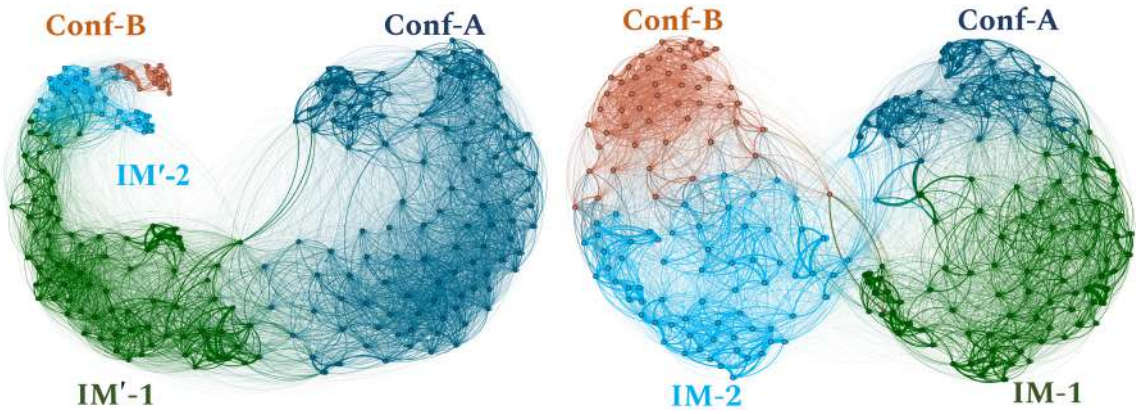

Figure 6.11: Visualization of the transition probability matrix created with Gephi software, emphasizing the conformational shifts among the diverse metastable states for the apo (left) and bound (right) systems. Conf-A and Conf-B are depicted in blue and pink, respectively, while the intermediate states are represented in green and cyan.

stable, and dynamically resolved conformational states. As illustrated in the thermodynamic cycle in Figure 6.1, the reverse scenario must also hold true. In other words,

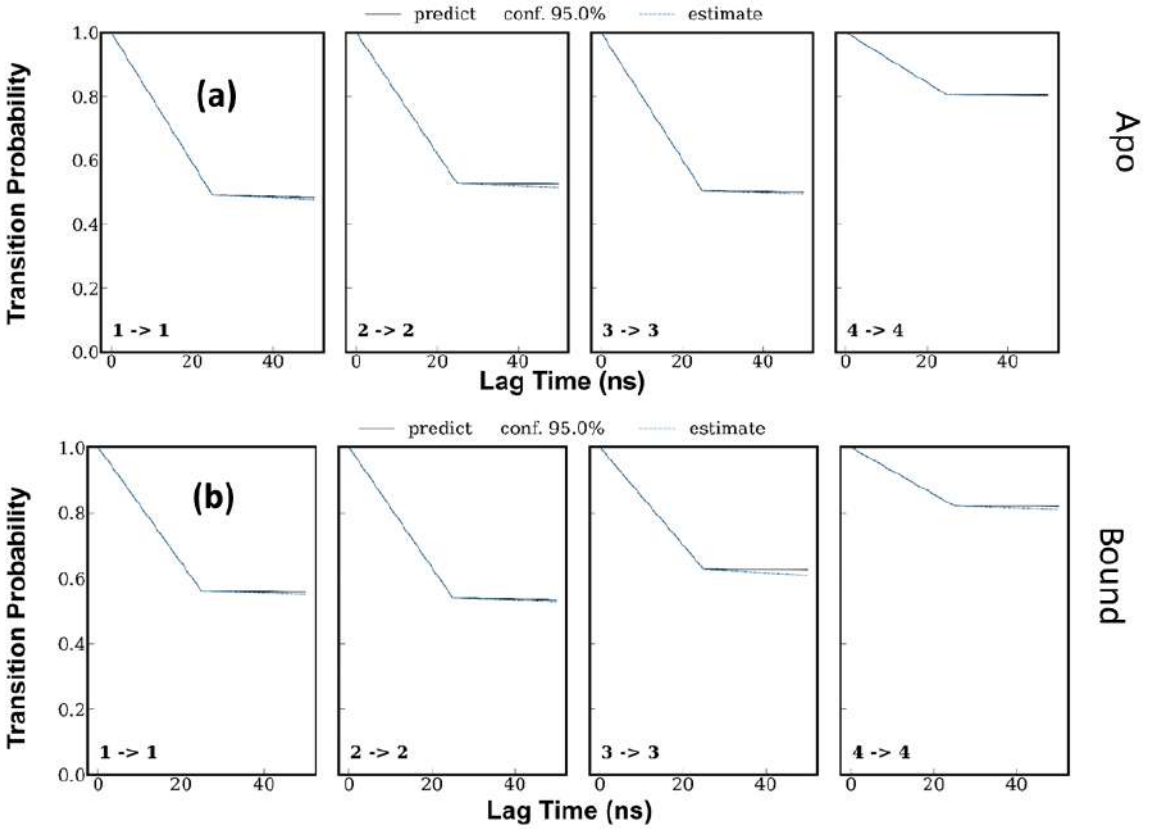

Figure 6.12: Validation of the Markov State Model (MSM) using the Chapman-Kolmogorov (CK) test. The plot illustrates the MSM data (black line) alongside the observed trajectory (blue dotted line).

if we can manipulate or adjust the conformation of this loop region, it would reciprocally modulate the binding affinity of the ligand (EGF-A) in a reversible manner, which aligns with our therapeutic objective.

Applying the concepts depicted in Figure 6.1 to this specific PCSK9 system, we can infer that since the Conf-A conformation exhibits higher stability compared to Conf-B in the apo state, this loop conformation should correspond to a lower PCSK9–LDLR binding affinity. In simpler terms, any loop conformations that are more prevalent in the apo state compared to the EGF-A bound state should exhibit reduced EGF-A binding affinity. Consequently, if we can stabilize these loop conformations using small molecules (inhibitors), our therapeutic objective would be fulfilled. To verify

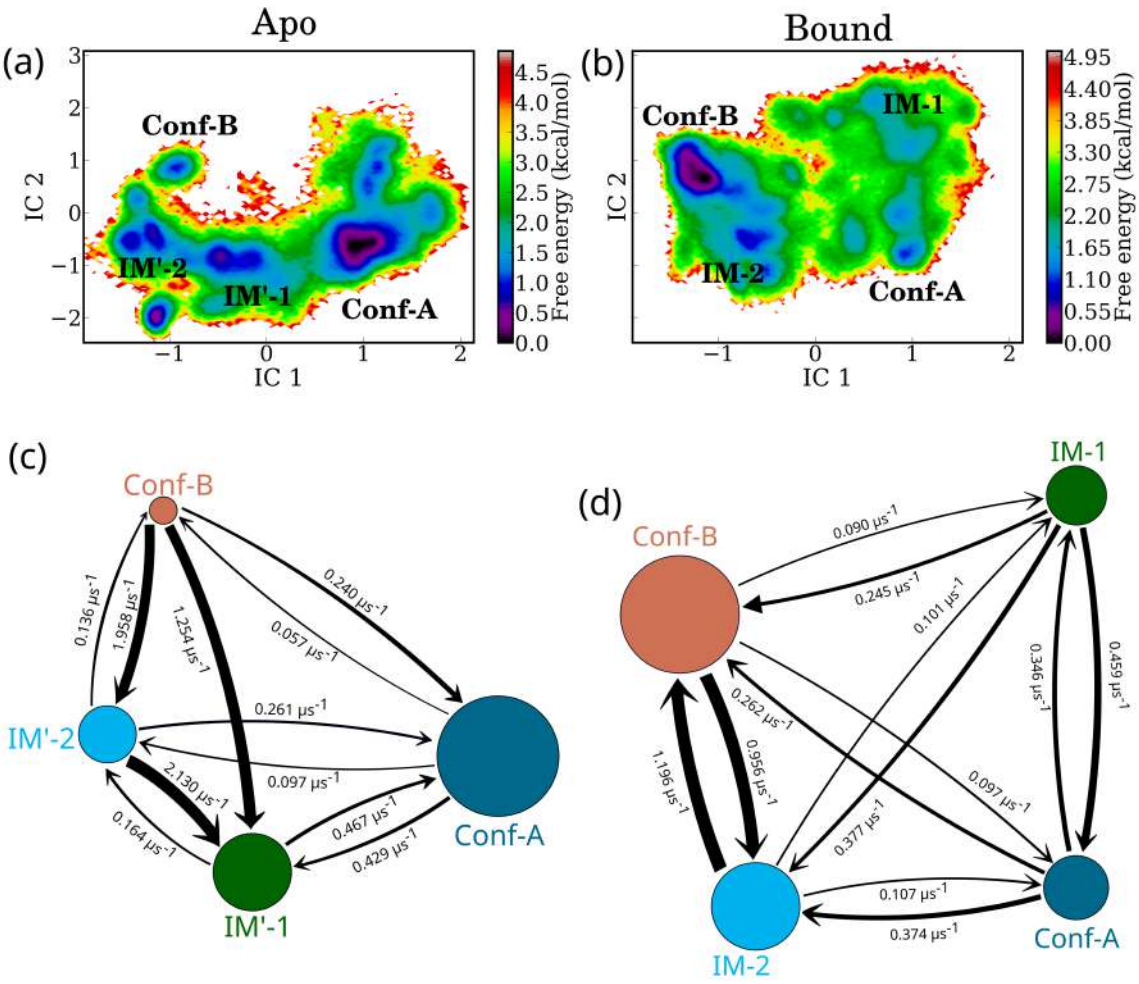

Figure 6.13: (top) Reweighted free energy landscapes derived from the Markov State Model (MSM) projected onto the slowest time-lagged independent components (TIC) IC1 and IC2 for (a) the apo system and (b) the bound system. (bottom) Transition probabilities between distinct metastable conformations for both apo and bound systems. Conf-A and Conf-B are depicted in blue and pink, respectively, while intermediate states are represented in green and cyan.

our hypothesis, we computed the binding energy between PCSK9 and EGF-A using the MM/PBSA method for various conformations of the loop 211–222. It's important to note that the MM/PBSA method is suitable for assessing the relative scoring of binding affinity, rather than providing a precise measurement of absolute binding free energy. In this context, our objective is to establish the relative ranking of protein conformations based on their binding affinity rather than obtaining a quantitative

State	Population (%)	Free Energy (G/KT)
Conf-A	4	2.73
IM′-1	14	1.41
IM′-2	24	0.91
Conf-B	58	0.00

Table 6.4: Population distribution and relative free energy for each macrostate derived from the Markov State Model (MSM) for the apo system.

State	Population (%)	Free Energy (G/KT)
Conf-A	13	1.32
IM-1	14	1.22
IM-2	26	0.58
Conf-B	47	0.00

Table 6.5: Population distribution and relative free energy for each macrostate derived from the Markov State Model (MSM) for the bound system.

estimate of absolute binding free energy. For this purpose, we gathered a hundred structures for the macrostate of the MSM for the bound system (Figure 6.17) and analyzed the distribution of binding energy for these macrostates individually [Figure 6.18 (a)].

Interestingly, the distribution of binding energy for different conformational states demonstrates a systematic trend with the loop conformation following the sequence: Conf-A < IM-1 < IM-2 < Conf-B, mirroring the stability order of these conformations in the apo form. It's worth noting that while the apo system doesn't precisely feature IM-1 and IM-2 metastable states, it does include IM′-1 and IM′-2 states, which closely resemble the IM-1 and IM-2 states observed in the bound system. This consistency aligns with the predictions of the thermodynamic cycle presented in Figure 6.1, suggesting that conformations more predominantly populated in the "apo" ensemble compared to the "bound" ensemble should exhibit lower binding affinity for the substrate (LDLR). To validate the MM/PBSA binding energy values, we assessed the

Residue Pairs	Conf-A (%)	IM-1 (%)	IM-2 (%)	Conf-B (%)
E211-R251	91	97	7	0
E211-C255	42	66	19	5
E206-R215	54	28	0	0
D224-R251	27	21	68	2
E210-K258	60	41	6	1
D192-R218	56	0.0	0	0
R218-S373	49	0.0	0	0
E211-R218	20	0	0	0
S188-D224	0	74	25	25
E211-R215 (SideChain-SideChain)	11	22	69	95
E211-Q256	0.0	2	24	51
D175-S178	0.0	0	5	32
E211-G213	10.0	0	0	33
F216-Q219	20.0	1	7	63
G1-S153	20.0	1	7	67
T214-R215	1.0	12	3	49
K222-H226	0.0	8	0	50
H226-Q382	4.0	5	10	62
D212-Q256	3.0	1	47	89

Table 6.6: Hydrogen bonds essential for stabilizing each metastable conformational macrostate as identified by the Markov state model (MSM) in the bound system. Analysis performed using 100 snapshots from each metastable minimum, as depicted in Figure 6.17.

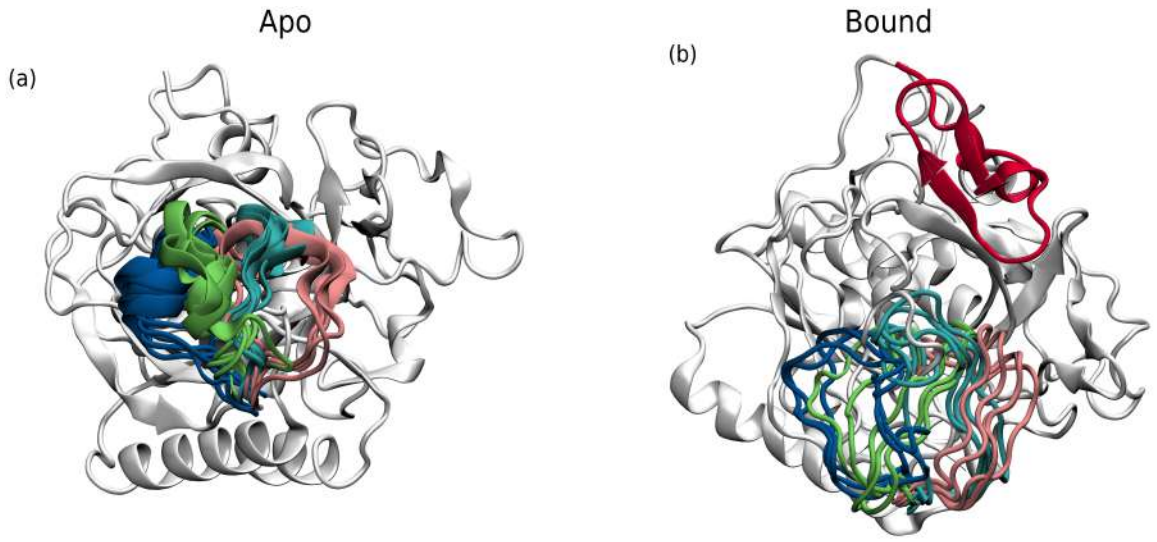

Figure 6.14: Configurations of the loop corresponding to each metastable state derived from the Markov State Model (MSM) for (a) Apo and (b) Bound systems. Conf-A and Conf-B are represented in blue and pink, respectively. The intermediate states are depicted in green (IM-1 and IM′-1) and cyan (IM-2 and IM′-2).

PCSK9–LDLR binding energy for two known point mutations: A443T and D374Y. The A443T mutation, where the A443 residue is distant from the EGF-A binding site, and the D374Y mutation, known to enhance LDLR binding affinity, serve as reference points. Our MM/PBSA results corroborate these experimental findings, with the A443T mutation showing nearly consistent binding energy (–67.9 ± 2.2 kcal/mol compared to –69.8 ± 2.5 kcal/mol for Conf-B), while the D374Y mutation exhibits a significant increase (–100.9 ± 2.9 kcal/mol) in binding affinity between PCSK9 and the EGF-A domain of LDLR. Another gain-of-function mutation, F216L, also shows minimal impact on binding energy, reducing Furin/PC5/6A dependent cleavage and enhancing the functional PCSK9 level[58]. The average binding energy values are summarized in Table 6.7. Additionally, we employed steered molecular dynamics to compute the binding energy between PCSK9 and the EGF-A domain of LDLR. The average work profile for each of the metastable conformations obtained from the bound state MSM is depicted in Figure 6.19, and the corresponding binding energy values are listed in Table 6.8. The work required to detach LDLR from the PCSK9

System	Binding Energy (kcal/mol)
Conf-A	-21.1 ± 3.2
IM-1	-22.7 ± 2.0
IM-2	-43.8 ± 3.4
Conf-B	-69.8 ± 2.5
A443T (Loss of Function)	-67.9 ± 2.2
D374Y (Gain of Function)	-100.9 ± 2.9

Table 6.7: Average binding energies between PCSK9 and the EGF-A domain of LDLR calculated using MM/PBSA methodology, including binding energies for two specific known mutations.

System	Binding Energy (kcal/mol)
Conf-A	-28.1 ± 1.1
IM-1	-35.4 ± 1.0
IM-2	-41.1 ± 1.0
Conf-B	-47.1 ± 1.0

Table 6.8: Binding energy between PCSK9 and LDLR computed using Jarzynski equality.

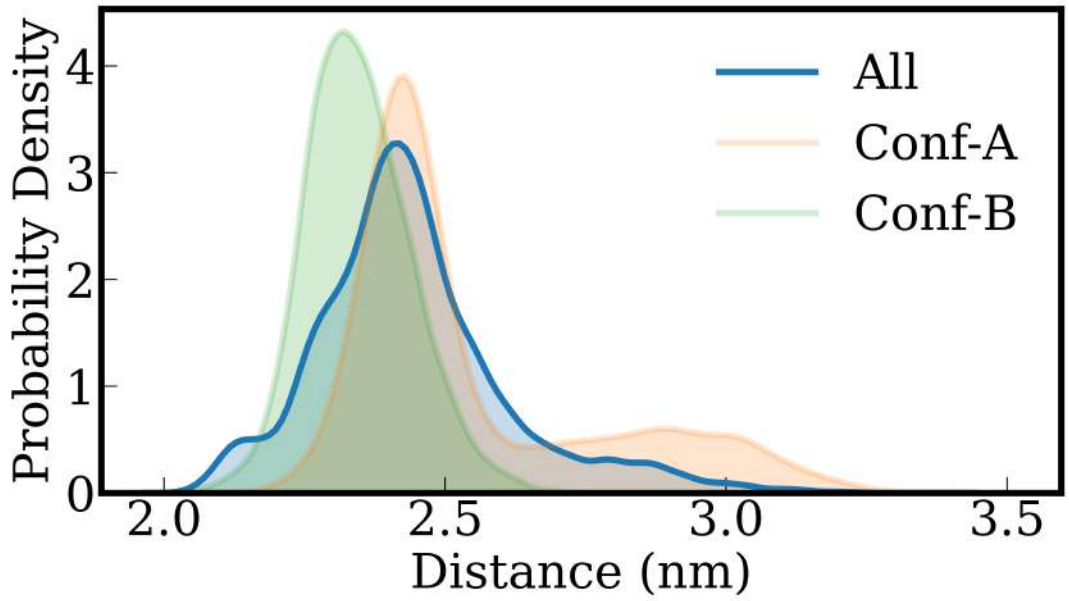

Figure 6.15: Distribution of center-of-mass (COM) to center-of-mass distance between the 211-222 loop and the EGF-A domain of LDLR. Additionally, the distance distributions for the Conf-A and Conf-B sub-ensembles are depicted.

binding site follows the order Conf-B > IM-2 > IM-1 > Conf-A Figure 6.20. This suggests that the most energy is needed to detach LDLR from the Conf-B configuration, while the least effort is required to detach LDLR from the Conf-A. The binding free energy estimated using Jarzynski equality (Table 6.8) aligns precisely with the trend obtained from MM/PBSA calculation.

In this study, our aim is to provide a qualitative proof-of-concept demonstration that validates the thermodynamic cycle proposed in Figure 6.1. Given the thermodynamic principles outlined earlier, we can confidently assert that the loop conformations predominant in the apo form correlate with lower LDLR binding energy and thus can potentially serve as promising targets for inhibitor design. Furthermore, our study underscores the intriguing possibility that even seemingly "structure-less" or disordered loops can exhibit well-defined and distinct conformational states with varying degrees of allosteric influence[59]. Moreover, these specific loop conformations

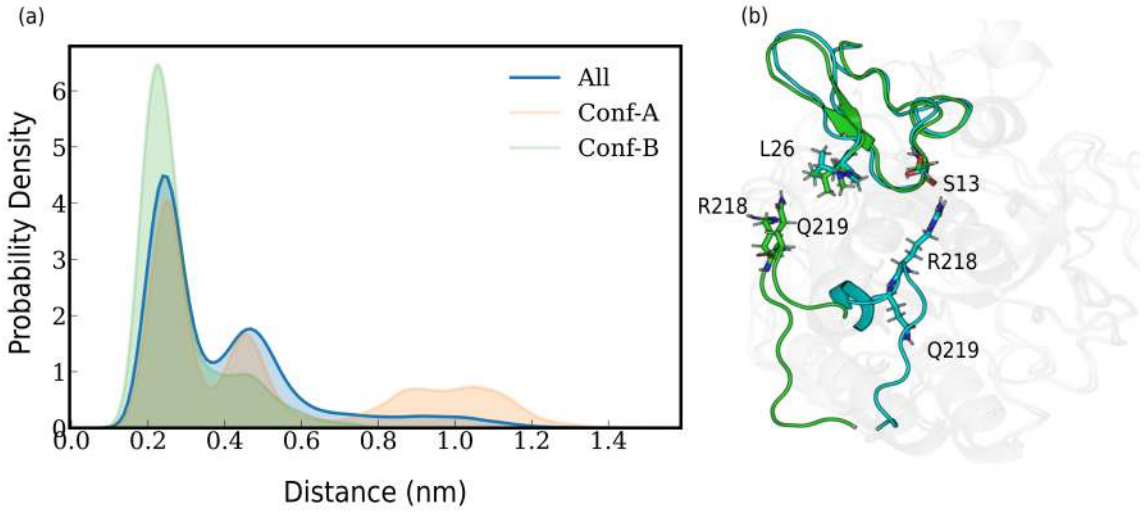

Figure 6.16: Distribution of the minimum distance between the loop 211-222 and the EGF-A domain of LDLR. The minimum distance distributions for the Conf-A and Conf-B sub-ensembles are presented. This illustrates the loop's interactions with the EGF-A domain of LDLR. In Conf-B, residue R218 interacts with residue S13 of LDLR, while in Conf-A, residue Q219 interacts with residue L26 of LDLR.

display unique thermodynamic properties concerning their structure and interactions with the remainder of the protein. Each of these conformational states is stabilized by a specific set of interactions that vary across the conformational states. These interactions play a crucial role in establishing the kinetic barrier to interconversion between them. The stability or prolonged lifespan of these states renders them suitable as targets for therapeutic intervention.

6.4.4 Allosteric Influence Evidenced by Redistribution of Specific Interactions

To delve deeper into the molecular mechanisms underlying the allosteric interplay between loop 211–222 and the EGF-A binding region of PCSK9, we revisited our previous approach where we demonstrated the sensitivity of nonbonded interaction energy to subtle conformational changes[59,60]. Building upon this concept, we constructed a network to depict the perturbation in average residue-pairwise nonbonded

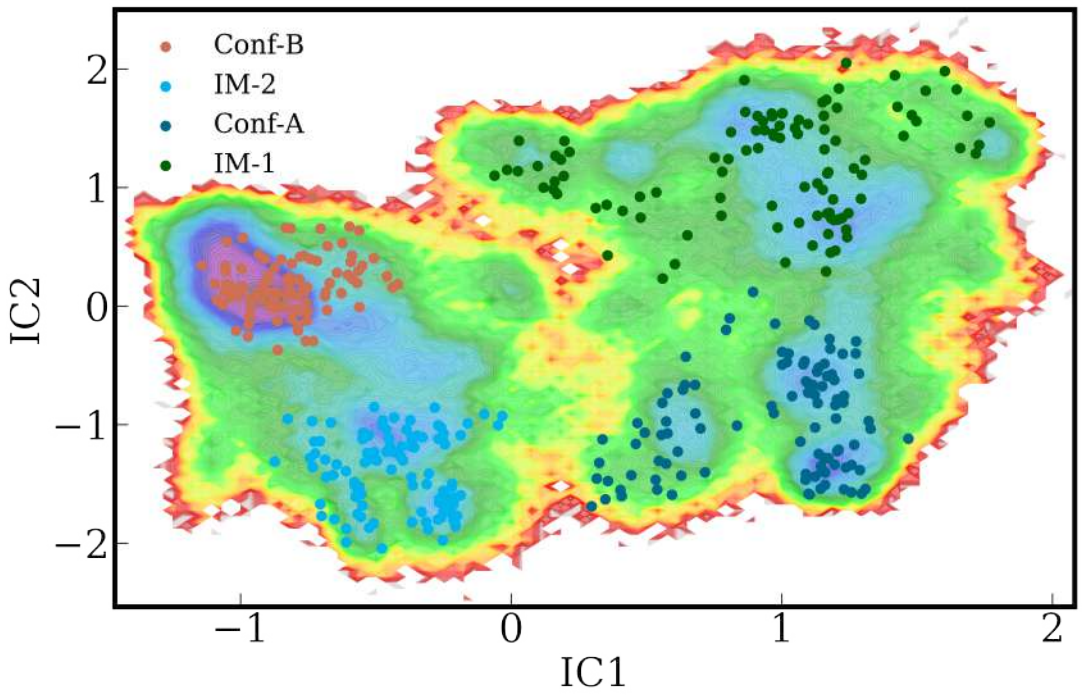

Figure 6.17: Frames representative of each minimum used for the calculation of binding energy using MM/PBSA. The structures corresponding to Conf-A, IM-1, IM-2, and Conf-B are depicted in cadet blue, green, blue, and pink colors, respectively.

interaction energy, expressed as $\Delta E_{ij} = \langle E_{ij} \rangle_{\text{Conf-B}} - \langle E_{ij} \rangle_{\text{Conf-A}}$, where $\langle E_{ij} \rangle_{\text{Conf-B}}$ and $\langle E_{ij} \rangle_{\text{Conf-A}}$ represent the average nonbonded interaction energy between the i-th and j-th residues of Conf-B and Conf-A configurations respectively. The connectivity pattern between residue pairs with substantial absolute values of ΔE_{ij} is visualized in Figure 6.18(b), with detailed values provided in Table 6.9. A detailed examination of the ΔE_{ij} values reveals a localized reconfiguration of electrostatic interactions resulting from the reorganization of pairwise specific interactions. This redistribution spans from the LDLR binding site to the allosteric site of loop 211–222. Specific structural changes between the apo and bound conformations are highlighted in Figure 6.21. For instance, in Conf-A, residue S153 forms a favorable interaction with D169, absent in Conf-B, while the interaction between S153 and D7 (from the

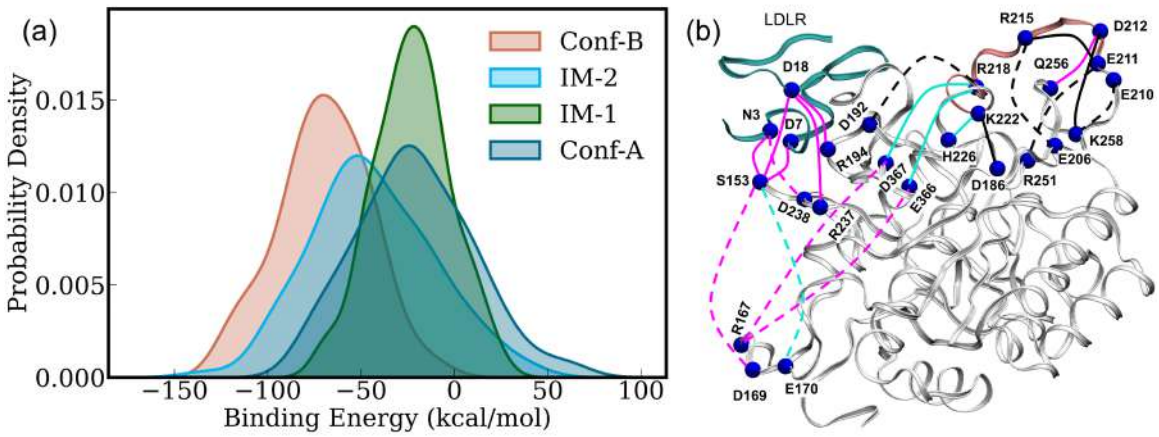

Figure 6.18: (a) Distributions of MM/PBSA binding energies for the PCSK9-LDLR complex structures extracted from each macrostate of MSM (representing minima in the bound system's free energy surface). The average binding energy sequence is as follows Conf-A < IM-1 < IM-2 < Conf-B.(b) Visualization of energy redistribution between Conf-A and Conf-B indicated by the alteration in average residue pairwise nonbonded interaction energy: $\langle \Delta E_{ij} \rangle = \langle E_{ij} \rangle_{\text{Conf-B}} - \langle E_{ij} \rangle_{\text{Conf-A}}$. Solid and dashed lines denote negative and positive $\langle \Delta E_{ij} \rangle$ values, respectively. Line colors indicate $\langle \Delta E_{ij} \rangle$ magnitude: (i) black: $|\langle \Delta E_{ij} \rangle| > 20$, (ii) cyan: $20 > |\langle \Delta E_{ij} \rangle| > 10$, (iii) magenta: $10 > |\langle \Delta E_{ij} \rangle| > 1$. Residues engaged in this network are represented as blue spheres.

EGF-A domain of LDLR) strengthens significantly in Conf-B, correlating with the enhanced PCSK9–LDLR binding affinity in Conf-B. Likewise, residue R218 in the loop forms a robust salt bridge with E366 in Conf-B (absent in Conf-A), whereas it interacts with D192 in Conf-A. Figure 6.22 illustrates this phenomenon through the probability distribution of the minimum distance between these residue pairs. Our findings further substantiate the notion of conformational rearrangement and alteration of local electrostatic interactions leading to long-range allosteric effects. Notably, the disruption of one set of favorable interactions in Conf-A is compensated by the formation of a different set of interactions in Conf-B. Consequently, multiple conformational states accessible within a similar energy range emerge due to this local cancellation or redistribution of interaction energies. This substantial conformational entropy associated with loop conformations appears pivotal for the conformational selection mode of allostery, as proposed earlier[61,62]. The significance of conforma-

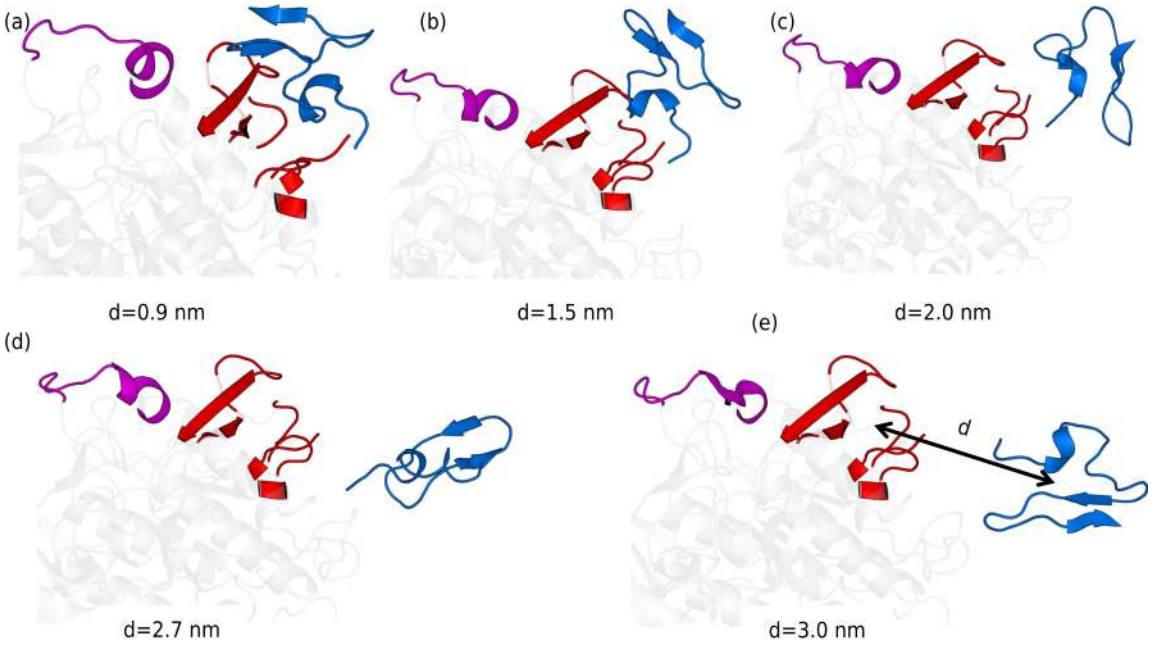

Figure 6.19: Constant velocity pulling is conducted along the distance (d) between the center of mass (COM) of LDLR (depicted in blue) and the COM of the interacting region of PCSK9 (depicted in red). The 211-222 loop is illustrated in magenta.

tional entropy and the ensemble nature in allosteric regulation has been underscored in previous literature[11,63–65].

6.5 Conclusion

To conclude, this article has introduced a general framework for identifying potential allosteric sites capable of modulating protein-protein interactions (PPIs), including binding and cryptic sites. The method involves comparing the conformational profiles of the target protein in its ligand-bound and apo states to identify significant conformational changes. If such distinctive conformations are identified and their thermodynamic properties are confirmed, they can serve as targets for virtual screening or structure-based drug discovery.

Figure 6.23 expands upon the thermodynamic cycle outlined in Figure 6.1 by incorporating the binding affinity of the allosteric inhibitor. The inhibitor is expected to ex-

Residue Pairs	$\langle E \rangle_{Conf-A}$	$\langle E \rangle_{Conf-A}$	$\Delta E = \langle E \rangle_{Conf-B} - \langle E \rangle_{Conf-A}$
E211-R215	-45.7	-89.9	-44.2
K222-D186	-11.2	-43.7	-32.5
D212-K258	-14.1	-40.6	-26.5
R218-D367	-4.1	-23.2	-19.1
R218-E366	-2.3	-14.9	-12.6
K222-H226	0.7	-11.5	-12.2
D212-Q256	-1.7	-9.3	-7.6
S153-N3	-5.7	-8.5	-2.8
D18-R194	-79.0	-81.7	-2.7
S153-D18	-17.5	-20.0	-2.5
D18-R237	-23.0	-25.0	-2.0
S153-D7	-39.6	-40.6	-1.0
R167-D367	-7.6	-5.6	2.0
R167-E366	-7.5	-5.3	2.2
N3-D238	-11.0	-7.3	3.7
S153-D169	-20.1	-10.9	9.2
S153-E170	-20.0	-8.6	11.4
E210-K258	-76.7	-28.1	48.6
E206-R215	-63.6	-14.9	48.7
R218-D192	-59.3	-3.8	55.5
E211-R251	-87.7	-26.9	60.8

Table 6.9: Change in the average pairwise interaction energy between Conf-A and Conf-B states.

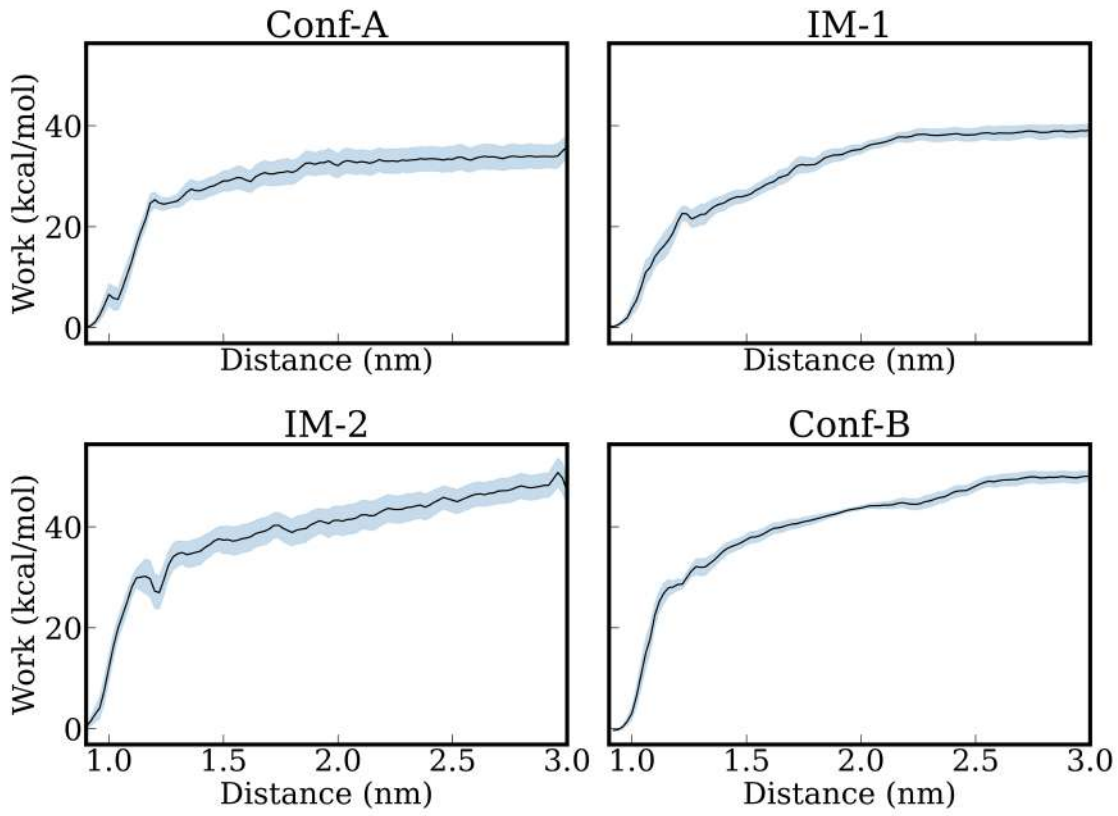

Figure 6.20: The work needed to separate LDLR from the complex. The black line indicates the average work, while the blue shaded area represents the standard deviation. Both the mean and standard deviation are computed across five independent steered molecular dynamics (SMD) simulations. The work required to detach LDLR from the complex follows the sequence Conf-A < IM-1 < IM-2 < Conf-B

hibit higher binding affinity for apo-specific conformations compared to bound-specific conformations. According to this cycle, the equation $\Delta G^{I}_{b,S} - \Delta G^{0}_{b,S} = \Delta G^{holo}_{b,I} - \Delta G^{apo}_{b,I}$ is established, implying that $\Delta G^{holo}_{b,I} > \Delta G^{apo}_{b,I}$ would lead to $\Delta G^{I}_{b,S} > \Delta G^{0}_{b}$ and S. In other words, if the inhibitor (I) binds with higher affinity to the apo conformation than the bound conformation, it suggests lower substrate (S) binding affinity in the presence of the inhibitor. Our future investigations will involve identifying small molecule inhibitors to stabilize apo-specific conformations. It's worth noting that this bidirectional aspect of allostery has been discussed previously in various contexts.

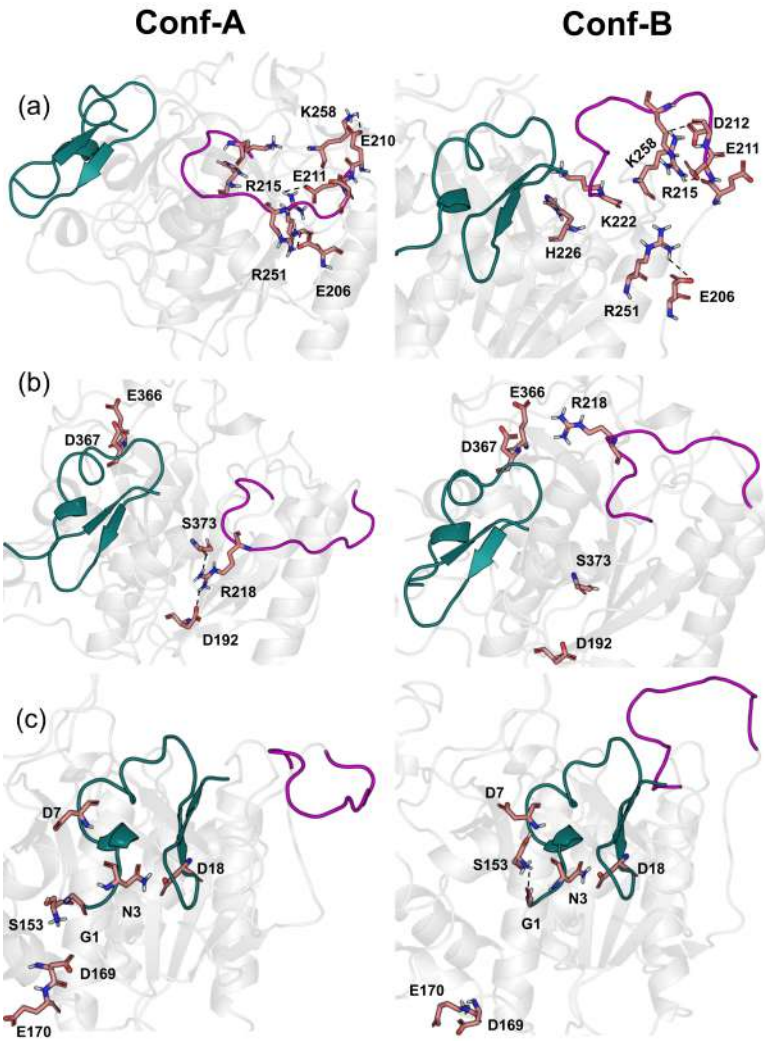

Figure 6.21: Significant structural interaction variations are depicted between Conf-A (left panel) and Conf-B (right panel) across distinct regions: (a) in proximity to the 211-222 loop, (b) in the intermediate region bridging the 211-222 loop and the LDLR binding site, and (c) in the vicinity of the LDLR binding site.

In this study, we have employed and confirmed this strategy by pinpointing a distant loop segment (residues 211–222) within PCSK9 capable of influencing the PCSK9–LDLR interaction via reorganization and adjustment of the hydrogen-bonded connections. Our upcoming research endeavors will concentrate on identifying appropriate inhibitors designed to stabilize specific loop configurations exhibiting reduced

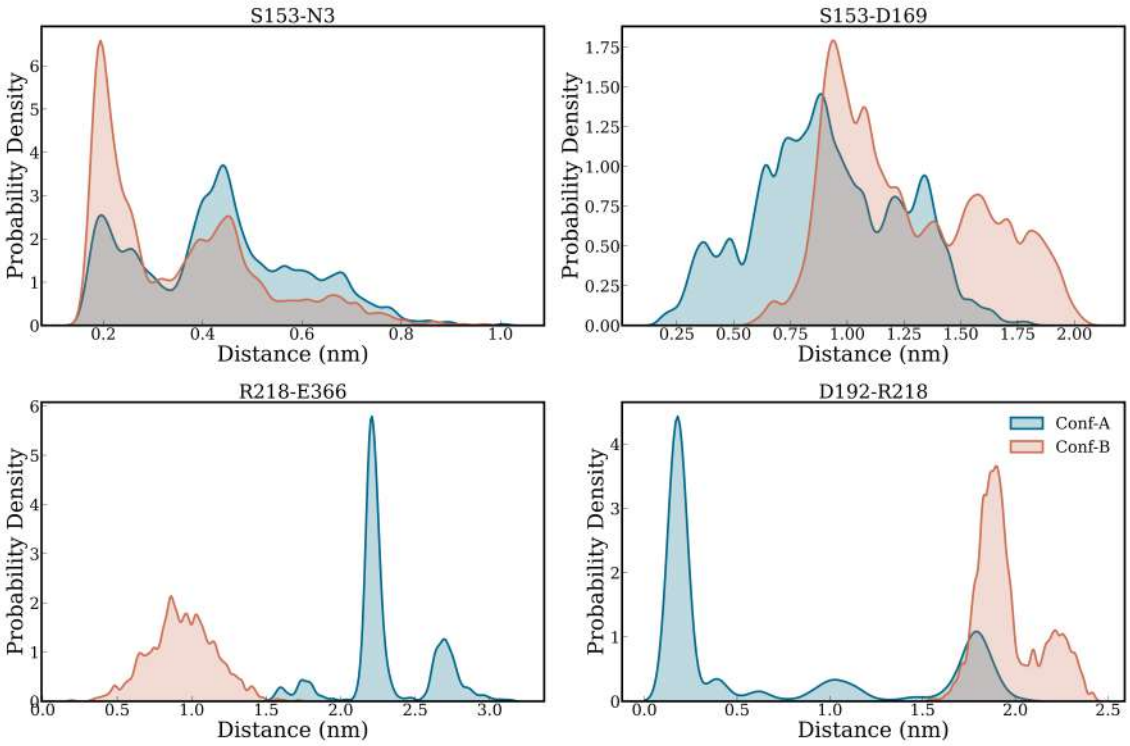

Figure 6.22: Distribution of population regarding the minimum distance between significant residue pairs: (a) S153-N3 (where N3 represents the N-terminus of the EGF-A domain of LDLR), (b) S153-D169, (c) R218-E366, and (d) D192-R218.

binding affinity with LDLR, such as Conf-A, IM-1, among others. Furthermore, we plan to extend this methodology to investigate other clinically significant protein-protein interactions for potential allosteric drug development. Notably, our findings challenge conventional assumptions by revealing that even seemingly disordered loops can exhibit identifiable structures and undergo transitions to stable conformational states, underscoring the need for exploring such phenomena in future drug discovery initiatives.

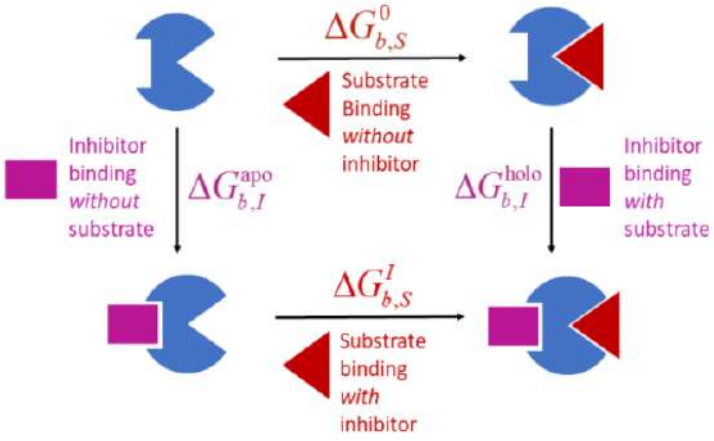

Figure 6.23: Representation of the thermodynamic cycle elucidating the operation of an allosteric inhibitor. The inhibitor's binding affinity must vary contingent upon whether the receptor is in its "apo" ($\Delta G^{b,I}_{apo}$) or "holo" ($\Delta G^{b,I}_{holo}$) state. $\Delta G^{I}_{b,S}$ and $\Delta G^{0}_{b,S}$ represent the substrate binding affinity in the absence and presence of the allosteric inhibitor.

Bibliography

[1] Jin, L, Wang, W, & Fang, G. (2014) Targeting protein-protein interaction by small molecules. *Annual Review of Pharmacology and Toxicology* **54**, 435–456.

[2] Jubb, H, Higueruelo, A. P, Winter, A, & Blundell, T. L. (2012) Structural biology and drug discovery for protein–protein interactions. *Trends in Pharmacological Sciences* **33**, 241–248.

[3] Ivanov, A. A, Khuri, F. R, & Fu, H. (2013) Targeting protein–protein interactions as an anticancer strategy. *Trends in Pharmacological Sciences* **34**, 393–400.

[4] Watkins, A. M & Arora, P. S. (2015) Structure-based inhibition of protein–protein interactions. *European Journal of Medicinal Chemistry* **94**, 480–488.

[5] Bai, F, Morcos, F, Cheng, R. R, Jiang, H, & Onuchic, J. N. (2016) Elucidating the druggable interface of protein–protein interactions using fragment docking and coevolutionary analysis. *Proceedings of the National Academy of Sciences* **113**.

[6] Vajda, S, Beglov, D, Wakefield, A. E, Egbert, M, & Whitty, A. (2018) Cryptic binding sites on proteins: definition, detection, and druggability. *Current Opinion in Chemical Biology* **44**, 1–8.

[7] Pan, T, Wang, Y, Xue, X, & Zhang, C. (2022) Rational design of allosteric switchable catalysts. *Exploration* **2**.

[8] Ni, D, Wei, J, He, X, Rehman, A. U, Li, X, Qiu, Y, Pu, J, Lu, S, & Zhang, J. (2021) Discovery of cryptic allosteric sites using reversed allosteric communication by a combined computational and experimental strategy. *Chemical Science* **12**, 464–476.

[9] Leroux, A. E & Biondi, R. M. (2020) Renaissance of allostery to disrupt protein kinase interactions. *Trends in Biochemical Sciences* **45**, 27–41.

[10] Berezovsky, I. N & Nussinov, R. (2022) Multiscale allostery: Basic mechanisms and versatility in diagnostics and drug design. *Journal of Molecular Biology* **434**, 167751.

[11] Toogood, P. L. (2002) Inhibition of protein–protein association by small molecules: Approaches and progress. *Journal of Medicinal Chemistry* **45**, 1543–1558.

[12] Nussinov, R & Tsai, C.-J. (2013) Allostery in disease and in drug discovery. *Cell* **153**, 293–305.

[13] Durrant, J. D & McCammon, J. A. (2011) Molecular dynamics simulations and drug discovery. *BMC Biology* **9**.

[14] Kuzmanic, A, Bowman, G. R, Juarez-Jimenez, J, Michel, J, & Gervasio, F. L. (2020) Investigating cryptic binding sites by molecular dynamics simulations. *Accounts of Chemical Research* **53**, 654–661.

[15] Meller, A, Bhakat, S, Solieva, S, & Bowman, G. R. (2023) Accelerating cryptic pocket discovery using alphafold. *Journal of Chemical Theory and Computation* **19**, 4355–4363.

[16] Meller, A, Ward, M, Borowsky, J, Kshirsagar, M, Lotthammer, J. M, Oviedo, F, Ferres, J. L, & Bowman, G. R. (2023) Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications* **14**.

[17] Smith, R. D & Carlson, H. A. (2021) Identification of cryptic binding sites using mixmd with standard and accelerated molecular dynamics. *Journal of Chemical Information and Modeling* **61**, 1287–1299.

[18] Hertig, S, Latorraca, N. R, & Dror, R. O. (2016) Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. *PLOS Computational Biology* **12**, e1004746.

[19] Tee, W.-V, Guarnera, E, & Berezovsky, I. N. (2018) Reversing allosteric communication: From detecting allosteric sites to inducing and tuning targeted allosteric response. *PLOS Computational Biology* **14**, e1006228.

[20] Poirier, S, Mayer, G, Benjannet, S, Bergeron, E, Marcinkiewicz, J, Nassoury, N, Mayer, H, Nimpf, J, Prat, A, & Seidah, N. G. (2008) The proprotein convertase pcsk9 induces the degradation of low density lipoprotein receptor (ldlr) and its closest family members vldlr and apoer2. *Journal of Biological Chemistry* **283**, 2363–2372.

[21] Cunningham, D, Danley, D. E, Geoghegan, K. F, Griffor, M. C, Hawkins, J. L, Subashi, T. A, Varghese, A. H, Ammirati, M. J, Culp, J. S, Hoth, L. R, Mansour, M. N, McGrath, K. M, Seddon, A. P, Shenolikar, S, Stutzman-Engwall, K. J, Warren, L. C, Xia, D, & Qiu, X. (2007) Structural and biophysical studies of pcsk9 and its mutants linked to familial hypercholesterolemia. *Nature Structural & Molecular Biology* **14**, 413–419.

[22] HORTON, J, COHEN, J, & HOBBS, H. (2007) Molecular biology of pcsk9: its role in ldl metabolism. *Trends in Biochemical Sciences* **32**, 71–77.

[23] Kwon, H. J, Lagace, T. A, McNutt, M. C, Horton, J. D, & Deisenhofer, J. (2008) Molecular basis for ldl receptor recognition by pcsk9. *Proceedings of the National Academy of Sciences* **105**, 1820–1825.

[24] Piper, D. E, Jackson, S, Liu, Q, Romanow, W. G, Shetterly, S, Thibault, S. T, Shan, B, & Walker, N. P. (2007) The crystal structure of pcsk9: A regulator of plasma ldl-cholesterol. *Structure* **15**, 545–552.

[25] Costet, P, Krempf, M, & Cariou, B. (2008) Pcsk9 and ldl cholesterol: unravelling the target to design the bullet. *Trends in Biochemical Sciences* **33**, 426–434.

[26] Steinberg, D & Witztum, J. L. (2009) Inhibition of pcsk9: A powerful weapon for achieving ideal ldl cholesterol levels. *Proceedings of the National Academy of Sciences* **106**, 9546–9547.

[27] Seidah, N. G & Prat, A. (2012) The biology and therapeutic targeting of the proprotein convertases. *Nature Reviews Drug Discovery* **11**, 367–383.

[28] Taechalertpaisarn, J, Zhao, B, Liang, X, & Burgess, K. (2018) Small molecule inhibitors of the pcsk9·ldlr interaction. *Journal of the American Chemical Society* **140**, 3242–3249.

[29] Tucker, T. J, Embrey, M. W, Alleyne, C, Amin, R. P, Bass, A, Bhatt, B, Bianchi, E, Branca, D, Bueters, T, Buist, N, Ha, S. N, Hafey, M, He, H, Higgins, J, Johns, D. G, Kerekes, A. D, Koeplinger, K. A, Kuethe, J. T, Li, N, Murphy, B, Orth, P, Salowe, S, Shahripour, A, Tracy, R, Wang, W, Wu, C, Xiong, Y, Zokian, H. J, Wood, H. B, & Walji, A. (2021) A series of novel, highly potent, and orally bioavailable next-generation tricyclic peptide pcsk9 inhibitors. *Journal of Medicinal Chemistry* **64**, 16770–16800.

[30] Sun, H, Wang, J, Liu, S, Zhou, X, Dai, L, Chen, C, Xu, Q, Wen, X, Cheng, K, Sun, H, & Yuan, H. (2021) Discovery of novel small molecule inhibitors disrupting the pcsk9-ldlr interaction. *Journal of Chemical Information and Modeling* **61**, 5269–5279.

[31] Petrilli, W. L, Adam, G. C, Erdmann, R. S, Abeywickrema, P, Agnani, V, Ai, X, Baysarowich, J, Byrne, N, Caldwell, J. P, Chang, W, DiNunzio, E, Feng, Z, Ford, R, Ha, S, Huang, Y, Hubbard, B, Johnston, J. M, Kavana, M, Lisnock, J.-M, Liang, R, Lu, J, Lu, Z, Meng, J, Orth, P, Palyha, O, Parthasarathy, G, Salowe, S. P, Sharma, S, Shipman, J, Soisson, S. M, Strack, A. M, Youm, H, Zhao, K, Zink, D. L, Zokian, H, Addona, G. H, Akinsanya, K, Tata, J. R, Xiong, Y, & Imbriglio, J. E. (2020) From screening to targeted degradation: Strategies for the discovery and optimization of small molecule ligands for pcsk9. *Cell Chemical Biology* **27**, 32–40.e3.

[32] Kuzmich, N, Andresyuk, E, Porozov, Y, Tarasov, V, Samsonov, M, Preferanskaya, N, Veselov, V, & Alyautdin, R. (2022) Pcsk9 as a target for development of a new generation of hypolipidemic drugs. *Molecules* **27**, 434.

[33] Raal, F. J, Stein, E. A, Dufour, R, Turner, T, Civeira, F, Burgess, L, Langslet, G, Scott, R, Olsson, A. G, Sullivan, D, Hovingh, G. K, Cariou, B, Gouni-Berthold, I, Somaratne, R, Bridges, I, Scott, R, Wasserman, S. M, & Gaudet, D. (2015) Pcsk9 inhibition with evolocumab (amg 145) in heterozygous familial hypercholesterolaemia (rutherford-2): a randomised, double-blind, placebo-controlled trial. *The Lancet* **385**, 331–340.

[34] Evison, B. J, Palmer, J. T, Lambert, G, Treutlein, H, Zeng, J, Nativel, B, Chemello, K, Zhu, Q, Wang, J, Teng, Y, Tang, W, Xu, Y, Rathi, A. K, Kumar, S, Suchowerska, A. K, Parmar, J, Dixon, I, Kelly, G. E, & Bonnar, J. (2020) A small molecule inhibitor of pcsk9 that antagonizes ldl receptor binding via interaction with a cryptic pcsk9 binding groove. *Bioorganic & Medicinal Chemistry* **28**, 115344.

[35] Brousseau, M. E, Clairmont, K. B, Spraggon, G, Flyer, A. N, Golosov, A. A, Grosche, P, Amin, J, Andre, J, Burdick, D, Caplan, S, Chen, G, Chopra, R, Ames, L, Dubiel, D, Fan, L, Gattlen, R, Kelly-Sullivan, D, Koch, A. W, Lewis, I, Li, J, Liu, E, Lubicka, D, Marzinzik, A, Nakajima, K, Nettleton, D, Ottl, J, Pan, M, Patel, T, Perry, L, Pickett, S, Poirier, J, Reid, P. C, Pelle, X, Seepersaud, M, Subramanian, V, Vera, V, Xu, M, Yang, L, Yang, Q, Yu, J, Zhu, G, & Monovich, L. G. (2022) Identification of a pcsk9-ldlr disruptor peptide with in vivo function. *Cell Chemical Biology* **29**, 249–258.e5.

[36] Šali, A & Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **234**, 779–815.

[37] Colovos, C & Yeates, T. O. (1993) Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science* **2**, 1511–1519.

[38] Messaoudi, A, Belguith, H, & Ben Hamida, J. (2013) Homology modeling and virtual screening approaches to identify potent inhibitors of veb-1 β-lactamase. *Theoretical Biology and Medical Modelling* **10**.

[39] Laskowski, R. A, MacArthur, M. W, Moss, D. S, & Thornton, J. M. (1993) Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**, 283–291.

[40] Abraham, M. J, Murtola, T, Schulz, R, Páll, S, Smith, J. C, Hess, B, & Lindahl, E. (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25.

[41] Lindorff-Larsen, K, Piana, S, Palmo, K, Maragakis, P, Klepeis, J. L, Dror, R. O, & Shaw, D. E. (2010) Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics* **78**, 1950–1958.

[42] Jorgensen, W. L. (1981) Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *Journal of the American Chemical Society* **103**, 335–340.

[43] Bussi, G, Donadio, D, & Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**.

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

[45] Essmann, U, Perera, L, Berkowitz, M. L, Darden, T, Lee, H, & Pedersen, L. G. (1995) A smooth particle mesh ewald method. *The Journal of Chemical Physics* **103**, 8577–8593.

[46] Hess, B, Bekker, H, Berendsen, H. J. C, & Fraaije, J. G. E. M. (1997) Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463–1472.

[47] Jarzynski, C. (1997) Nonequilibrium equality for free energy differences. *Physical Review Letters* **78**, 2690–2693.

[48] Park, S, Khalili-Araghi, F, Tajkhorshid, E, & Schulten, K. (2003) Free energy calculation from steered molecular dynamics simulations using jarzynski's equality. *The Journal of Chemical Physics* **119**, 3559–3566.

[49] Bonomi, M, Branduardi, D, Bussi, G, Camilloni, C, Provasi, D, Raiteri, P, Donadio, D, Marinelli, F, Pietrucci, F, Broglia, R. A, & Parrinello, M. (2009) Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **180**, 1961–1972.

[50] Scherer, M. K, Trendelkamp-Schroer, B, Paul, F, Pérez-Hernández, G, Hoffmann, M, Plattner, N, Wehmeyer, C, Prinz, J.-H, & Noé, F. (2015) Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of Chemical Theory and Computation* **11**, 5525–5542.

[51] Molgedey, L & Schuster, H. G. (1994) Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **72**, 3634–3637.

[52] Pérez-Hernández, G, Paul, F, Giorgino, T, De Fabritiis, G, & Noé, F. (2013) Identification of slow molecular order parameters for markov model construction. *The Journal of Chemical Physics* **139**.

[53] Schwantes, C. R & Pande, V. S. (2013) Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. *Journal of Chemical Theory and Computation* **9**, 2000–2009.

[54] Baker, N. A, Sept, D, Joseph, S, Holst, M. J, & McCammon, J. A. (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences* **98**, 10037–10041.

[55] Paissoni, C, Spiliotopoulos, D, Musco, G, & Spitaleri, A. (2015) Gmxpbsa 2.1: A gromacs tool to perform mm/pbsa and computational alanine scanning. *Computer Physics Communications* **186**, 105–107.

[56] Thayer, K. M, Lakhani, B, & Beveridge, D. L. (2017) Molecular dynamics–markov state model of protein ligand binding and allostery in crib-pdz: Conformational selection and induced fit. *The Journal of Physical Chemistry B* **121**, 5509–5514.

[57] del Sol, A, Tsai, C.-J, Ma, B, & Nussinov, R. (2009) The origin of allosteric functional modulation: Multiple pre-existing pathways. *Structure* **17**, 1042–1050.

[58] Benjannet, S, Rhainds, D, Hamelin, J, Nassoury, N, & Seidah, N. G. (2006) The proprotein convertase (pc) pcsk9 is inactivated by furin and/or pc5/6a. *Journal of Biological Chemistry* **281**, 30561–30572.

[59] Kumawat, A, Chakrabarty, S, & Kulkarni, K. (2017) Nucleotide dependent switching in rho gtpase: Conformational heterogeneity and competing molecular interactions. *Scientific Reports* **7**.

[60] Kumawat, A & Chakrabarty, S. (2017) Hidden electrostatic basis of dynamic allostery in a pdz domain. *Proceedings of the National Academy of Sciences* **114**.

[61] Ma, B, Tsai, C.-J, Haliloğlu, T, & Nussinov, R. (2011) Dynamic allostery: Linkers are not merely flexible. *Structure* **19**, 907–917.

[62] Papaleo, E, Saladino, G, Lambrughi, M, Lindorff-Larsen, K, Gervasio, F. L, & Nussinov, R. (2016) The role of protein loops and linkers in conformational dynamics and allostery. *Chemical Reviews* **116**, 6391–6423.

[63] Motlagh, H. N, Wrabl, J. O, Li, J, & Hilser, V. J. (2014) The ensemble nature of allostery. *Nature* **508**, 331–339.

[64] Tzeng, S.-R & Kalodimos, C. G. (2011) Protein dynamics and allostery: an nmr view. *Current Opinion in Structural Biology* **21**, 62–67.

[65] Nussinov, R & Tsai, C.-J. (2015) Allostery without a conformational change? revisiting the paradigm. *Current Opinion in Structural Biology* **30**, 17–24.

7

Mechanism of Substrate Binding to Human Acetylcholinesterase Protein

> "If I have seen further it is by standing on the shoulders of Giants."
>
> Isaac Newton

7.1 Introduction

Alzheimer's disease (AD) is a progressive neurological disorder characterized by cognitive decline, memory loss, and behavioral changes. It is the most common cause of dementia in individuals aged 65 and older, imposing a significant burden on patients, families, and healthcare systems worldwide. AD is marked by the accumulation of amyloid-beta (Aβ) plaques and neurofibrillary tangles composed of hyperphosphorylated tau protein, leading to neuronal death and brain atrophy[1]. Multiple factors, including genetics, environment, and lifestyle choices, contribute to the development

of AD. Despite extensive research, there is currently no cure, and treatments mainly focus on symptom relief and slowing disease progression.

The cholinergic hypothesis of AD, formulated in the 1970s[2], suggests that cognitive impairment in AD results from a deficiency of the neurotransmitter acetylcholine (ACh) in the brain. Declining levels of ACh are associated with the cognitive abnormalities seen in AD patients, as ACh is essential for learning and memory[3,4]. Acetylcholinesterase (AChE) is the enzyme responsible for hydrolyzing ACh into choline and acetate, thus terminating cholinergic neurotransmission between neurons[5]. Consequently, AChE is a key pharmacological target for AD; inhibiting its activity increases synaptic ACh levels, thereby enhancing cholinergic transmission and alleviating symptoms[6].

Structural studies of AChE have greatly enhanced our understanding of its function.

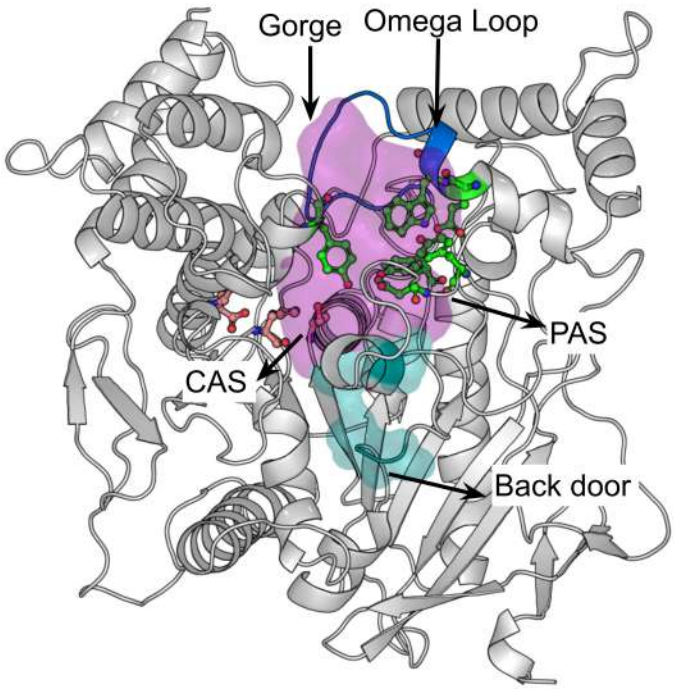

Figure 7.1: Crystal structure of AChE enzyme (PDB ID: 6O4W): The residues of the catalytic triad (S203, E334, and H447) and selected residues of the peripheral anionic site (PAS) (Y72, Y124, and W286) are highlighted in stick representation. The omega loop is shown in blue color and backdoor region is shown as cyan surface.

AChE contains a deep and narrow gorge, approximately 20 Å in depth, lined with aromatic residues that create a hydrophobic environment conducive to ACh binding[7]. The enzyme's active site, or catalytic active site (CAS), includes a glutamic acid residue (E334) and a serine residue (S203) that are crucial for ACh hydrolysis. Additionally, AChE has peripheral anionic sites (PAS), which are important for substrate binding and enzyme inhibition.[8]

AChE inhibitors work by targeting the enzyme's catalytic triad, resulting in the inactivation of AChE[9]. The enzyme's ability to catalyze the breaking of ACh molecules to acetate and choline is significantly influenced by these conformational changes of AChE enzyme, which control substrate entry and product release. Product removal from AChE's active site is challenging due to the narrow, deep gorge. Studies indicate that products are released via specific exit paths without hindering incoming substrates[10]. The preferred route, known as the "back door," is located approximately 14Å from CAS and 180° from the gorge mouth[11]. Alternative routes include the "side door," involving residues Asp74, Leu76, Met85, and Asn87 from the omega loop, and the exit site formed by sidechains in the 286-298 loop, opposite the catalytic site from the omega loop.

The narrowest part of the gorge cross-section is narrower than the diameter of choline, allowing only a single water molecule to pass through[12]. However, AChE undergoes "breathing motions" to expand the gorge width, facilitating catalytic activity[13]. Previous research highlights the importance of the "bottleneck" area in regulating the gorge access[13] of AChE. In human AChE, the open-closed transition is associated with changes in the distance between Tyr124 and Tyr341[8]. Recent studies have shown that residues Tyr133, Tyr337, and Tyr124 in the "bottleneck" area significantly influence substrate flow into the catalytic site[14]. Additionally, research on AChE inhibition by soman has revealed that residues Tyr72, Tyr124, Phe295, and Tyr341 obstruct the soman adduct from entering the gorge, even in the "open" state[15] where the distance between Tyr124 and Tyr338 is also correlated with the gorge radius.

In this study, we examined two previously described markers of gorge opening for the human AChE (hAChE) protein: the distance between Tyr124 and Tyr341, and the

distance between Tyr124 and Phe338. Our findings suggest that these distances do not completely represent gorge opening motion. Instead, we identified that a group of residues inside the gorge collectively regulates the gorge opening motion, and specific residue paris are not enough to describe the gorge opening process. Our results also indicate that the tendency for gorge opening increases in the presence of the ACh substrate, suggesting a conformational selection mechanism. This study provides new insights into the dynamic regulation of AChE's active site and proposes a novel mechanism underlying its catalytic efficiency.

7.2 Computational Details

The crystal structure of human acetylcholinesterase (AChE) was obtained from the Protein Data Bank (PDB ID: 6O4W[16]). The HIS447 residue was kept in its deprotonated state. For simulations involving AChE and acetylcholine (ACh), five ACh substrate molecules were randomly inserted into the protein solution. Parameters for the ACh substrate were generated using the CGenFF webserver[17].

We conducted five independent molecular dynamics (MD) simulations for the AChE-ACh system: four simulations of 1μs each and one simulation of 500 ns. For comparison, four independent simulations were performed for the apo system each of which is 200 ns long. All simulations were executed using GROMACS 2019.6[18] software with the CHARMM all-atom force field (version 36m) and the TIP3P water model[19,20]. The solvated systems were neutralized with counterions to achieve a salt concentration of 0.15 M, reflecting physiological conditions. Energy minimization was carried out using the steepest descent algorithm.

During the production runs, a modified Berendsen (V-rescale) thermostat[21] maintained the temperature at 300 K, and a Parrinello-Rahman barostat[22] maintained the pressure at 1 bar in the NPT ensemble. Periodic boundary conditions were applied throughout all simulations. Long-range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method[23] with a grid spacing of 0.16 nm and fourth-order cubic interpolation. Short-range electrostatic and van der Waals (vdW)

interactions were computed with a cutoff distance of 1 nm. The LINCS algorithm[24] was used to constrain all covalent bonds, and the integration time step was set to 2 fs to accurately capture the system dynamics.

7.3 Results and Discussions

7.3.1 Analysis of Specific Distances as Descriptors of Gorge Opening

Earlier work by Silva et al.[8] has suggested that the distance between Tyr124 and Tyr341 is useful for distinguishing between the "open" and "closed" states of the gorge, while Bennion et al[15]. have indicated that the distance between Tyr124 and Phe338 serves the same purpose. We intend to test this hypothesisin our work in systematic manner. First, we calculated the distances between (i) Y124-Y341 and (ii) Y124-F338, and plotted the distributions in Figure 7.2(a) and (b). The resulting plots exhibit distinct bimodal distributions, with two prominent peaks indicating the presence of both "open" and "closed" conformations. To visually confirm these findings, representative snapshots of the hAChE enzyme in both the "open" and "closed" states are shown in Figure 7.2(c) and (d). These snapshots illustrate the structural differences between the two states. However, an important question remains: do these observed "open" and "closed" states genuinely represent the opening and closing motion of the gorge? While the distance distributions clearly distinguish between two conformations.

The analysis of the Y124-Y341 minimum distance across five independent molecular dynamics simulations revealed a two-state behavior. However, the critical question was whether these states accurately represented the "open" and "closed" states of the acetylcholinesterase (AChE) gorge. To investigate this, we calculated the Y124-Y341 minimum distance for five independent runs, represented by the red solid line. The magenta line indicates the reference value, above which the gorge is considered "open" and below which it is "closed". We also plotted the minimum

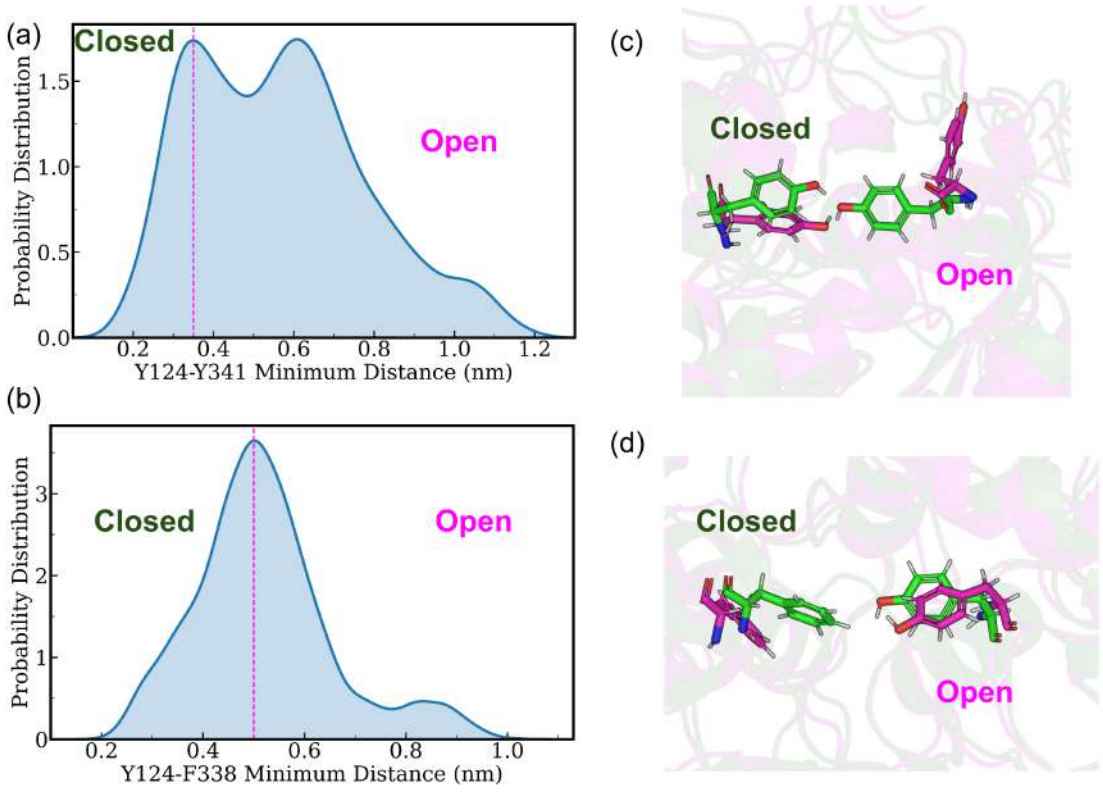

Figure 7.2: (a) and (b) Distribution plots for the distances between residues Y124-Y341 and Y124-F338, which describe the gorge opening motion of the hAChE enzyme. The plots show two distinct peaks, indicating the presence of both "open" and "closed" conformations. (c) and (d) Representative snapshots of the hAChE enzyme in its "open" and "closed" states, illustrating the structural differences corresponding to the two peaks observed in the distribution plots.

distance between the acetylcholine (ACh) molecule and the catalytic active site (CAS) of AChE, shown as the blue solid line. The green dotted line represents the threshold for ACh binding to the CAS; distances above this line indicate an "unbound" state, while those below indicate a "bound" state.

For run 1, the Y124-Y341 distance indicated an "open" state at around 180 ns and 620 ns. At approximately 180 ns, although the cavity was open, the ligand did not bind to the CAS despite being in close proximity. This suggests that the Y124-Y341 distance alone cannot adequately describe the cavity opening event.

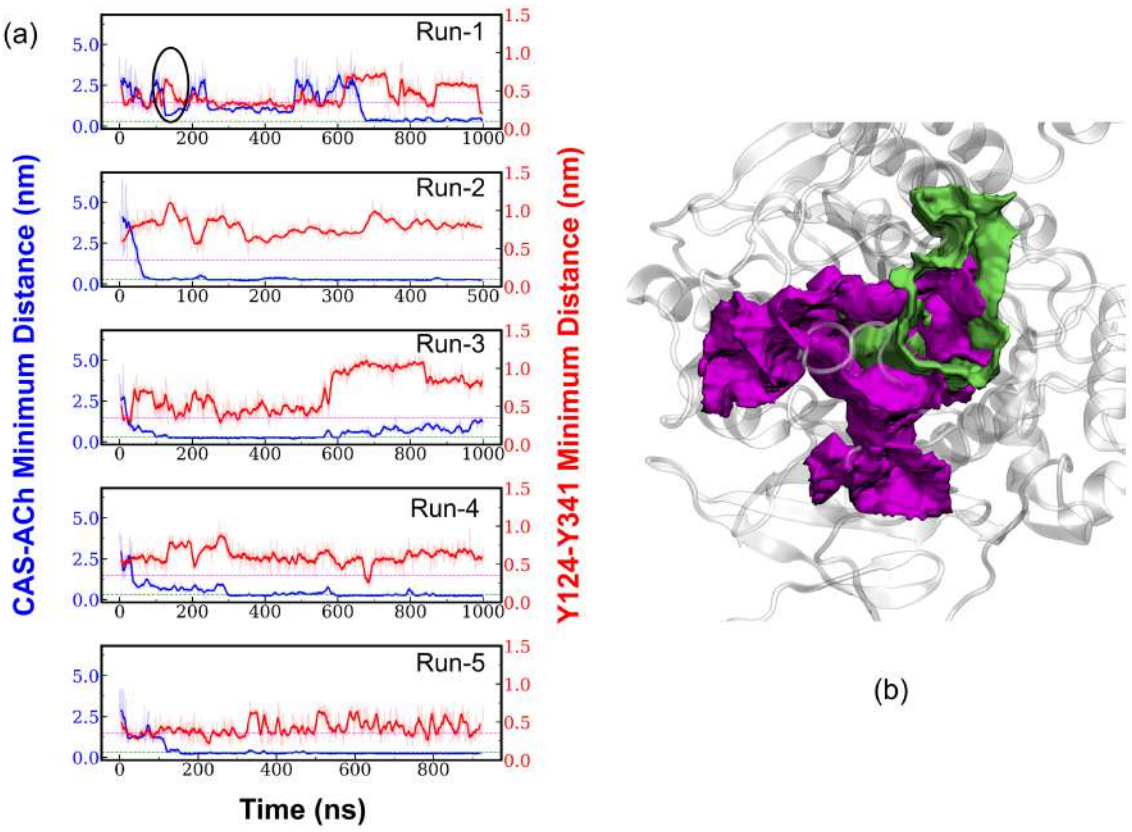

Figure 7.3: Analysis of Y124-Y341 minimum distance across 5 runs (red solid line) with the reference value (magenta line) distinguishing "open" and "closed" states. Blue solid line shows the minimum distance between ACh and the CAS site of AChE, with green dotted line indicating bound/unbound state. At 180 ns and 620 ns (run 1), Y124-Y341 was "open", but ligand binding only occurred at 620 ns. Snapshots at 180 ns (green) and 620 ns (magenta) reveal different extents of cavity opening, indicating Y124-Y341 distance alone does not represent gorge opening.

To further elucidate this, we took snapshots from 180 ns (green) and 620 ns (magenta), both having the same Y124-Y341 minimum distance. The snapshots revealed that the cavity opening at 180 ns was significantly smaller than at 620 ns. This observation indicates that the Y124-Y341 distance does not sufficiently represent the gorge opening.

These findings underscore the complexity of the gorge opening mechanism in AChE and suggest that additional structural parameters need to be considered to

accurately describe this process. Here we focus on identifying other critical residues and conformational changes that contribute to the gorge dynamics and ligand binding in AChE.

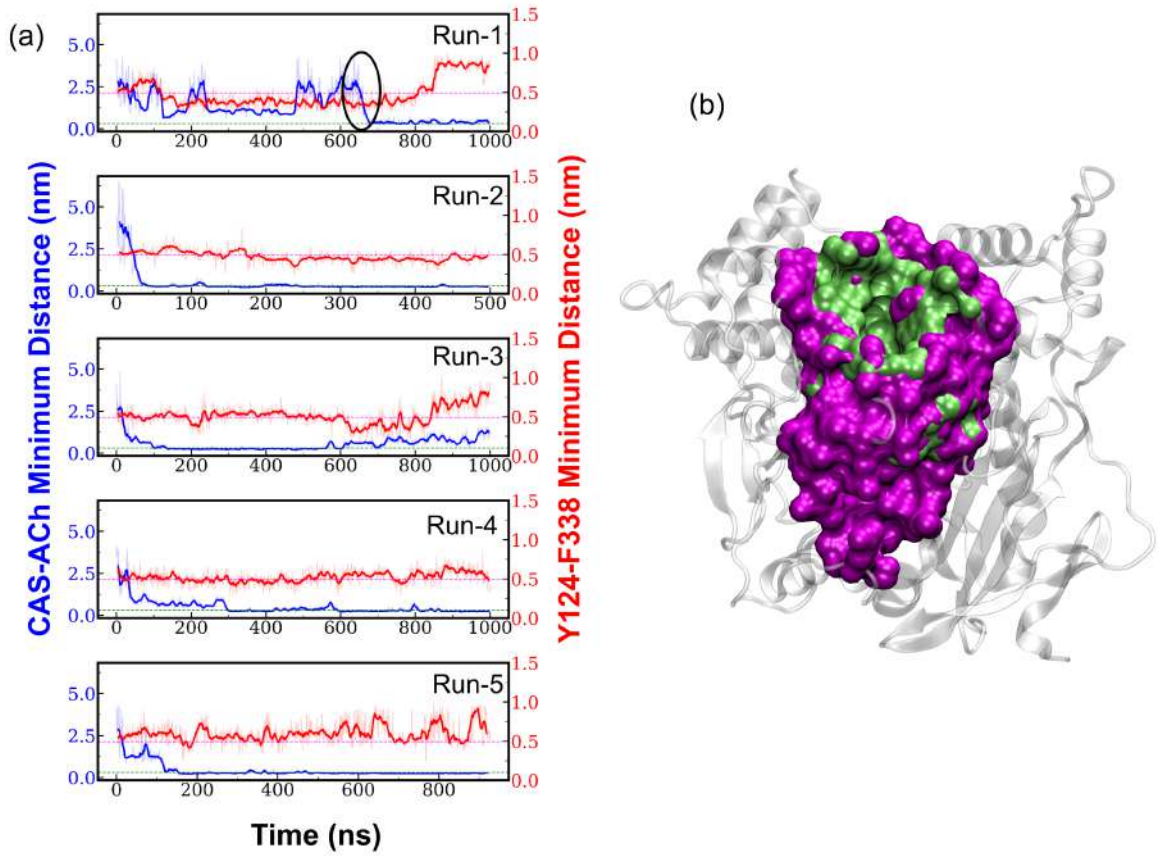

Figure 7.4: (a) Analysis of Y124-F338 minimum distance across five simulation runs (red solid line) with the reference value (magenta line) distinguishing "open" and "closed" states. The blue solid line shows the minimum distance between ACh and the CAS site of AChE, with the green dotted line indicating the bound/unbound state. (b) Comparison of cavity size from two structures with same Tyr124-F338 minimum distance value, it shows that the cavity size is much smaller in the one structure (green) than other (magenta), indicating that the Tyr124-Phe338 distance alone does not represent the gorge opening.

We analyzed the Y124-F338 minimum distance across five simulation runs, as shown by the red solid lines in Figure 7.4(a). The reference value, indicated

by the magenta line, helps distinguish between the "open" and "closed" states of the enzyme. Additionally, the blue solid line represents the minimum distance between ACh and the CAS site of AChE, with the green dotted line marking the bound/unbound state.

Notably, in Run-2, the Y124-F338 distance remains relatively constant throughout the simulation. To further investigate this, we compared the cavity size of two structures from Run-2: one from the initial part of the simulation and one from the later part, where the Y124-F338 distance is the same. As shown in Figure 7.4(b), the initial structure (green) has a much smaller cavity size compared to the later structure (magenta), which is in an open conformation.

This observation indicates that the Y124-F338 distance alone does not fully represent the gorge opening of the enzyme. While the distance measurements provide insights into the conformational states of the enzyme, they may not capture the complete picture of the dynamic opening and closing motions of the gorge.

Cavity Residues
THR83, TRP86, TYR121 LEU227, TYR337, PHE338, TYR448, TYR449

Table 7.1: List of Cavity Residues

7.3.2 Indentification of New Structural Metric of Gorge Opening

As previously discussed, the Y124-F338 and Y124-Y341 distances alone do not fully explain the opening of the gorge. To identify which residues play crucial roles in the gorge opening, we analyzed five independent simulations to determine which residues form contacts with the ACh molecule when it binds to the CAS site. These residues are referred to as "cavity residues."

Figure 7.5(a) shows the cavity residues (highlighted in blue and list of residues are shown in Table 7.1), situated at the bottleneck region of the gorge. This positioning

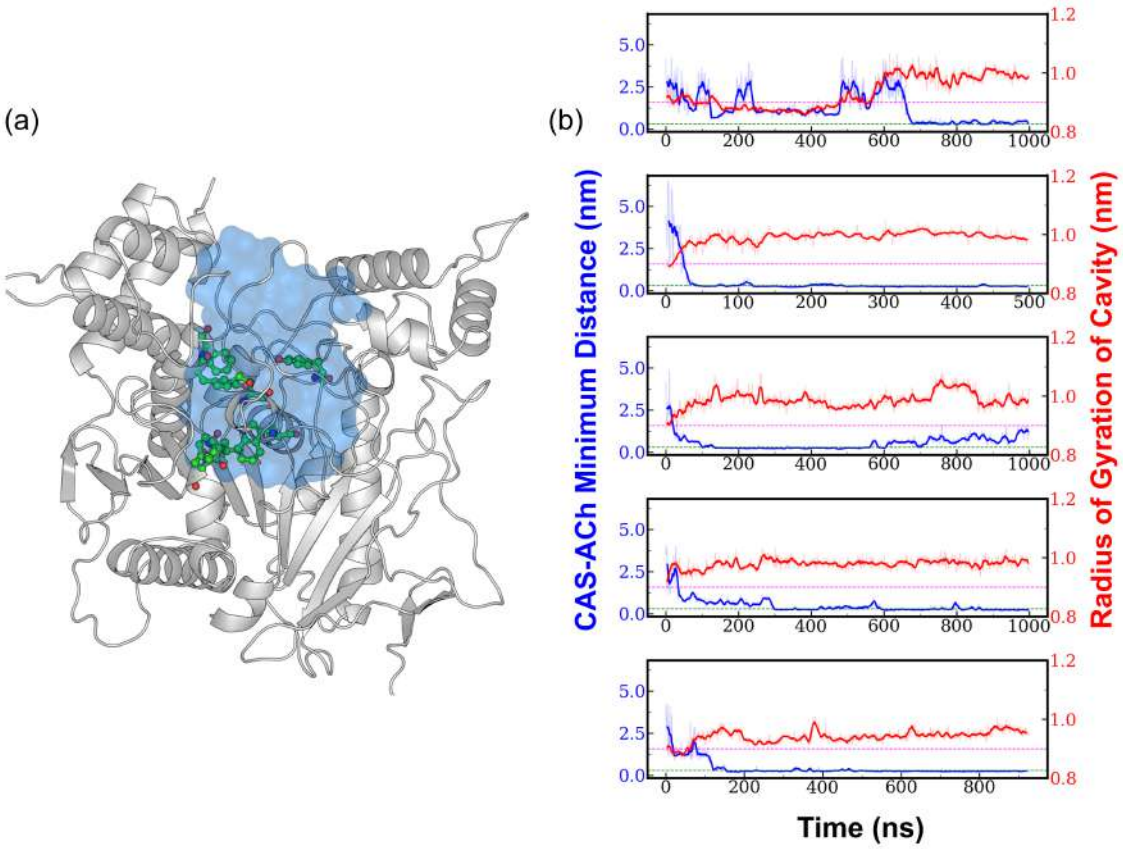

Figure 7.5: (a) Representation of "cavity residues" (shown in sticks representations) situated at the bottleneck region of the gorge (shown as blue surface) in the hAChE enzyme, which play a critical role in gorge opening. (b) Analysis of the radius of gyration (R_g) of cavity residues (red solid line) across five simulation runs with the reference value (magenta line) distinguishing "open" and "closed" states. The blue solid line indicates the minimum distance between ACh and the CAS site of AChE, with the green dotted line marking the bound/unbound state.

suggests that these residues may significantly influence the gorge opening motion. To test this hypothesis, we calculated the radius of gyration (Rg) of these cavity residues across the five simulation runs. Figure 7.5(b) illustrates the Rg of cavity residues (red solid line) compared with the reference value (magenta line) distinguishing the "open" and "closed" states. The blue solid line represents the minimum distance between ACh and the CAS site of AChE, with the green dotted line indicating the bound/unbound state.

The results show that in each independent simulation, ACh only enters the CAS when the cavity residues are in the "open" state. For instance, in Run-1, around 180 ns, ACh approaches the CAS, but it cannot enter the gorge as the cavity residues are in the "closed" state. Later, when the cavity residues transition to the "open" state, ACh successfully enters the gorge.

Thus, it can be inferred that the identified cavity residues provide a more accurate description of the gorge opening motion compared to the previously analyzed distances between Y124-F338 and Y124-Y341. Our analysis has established

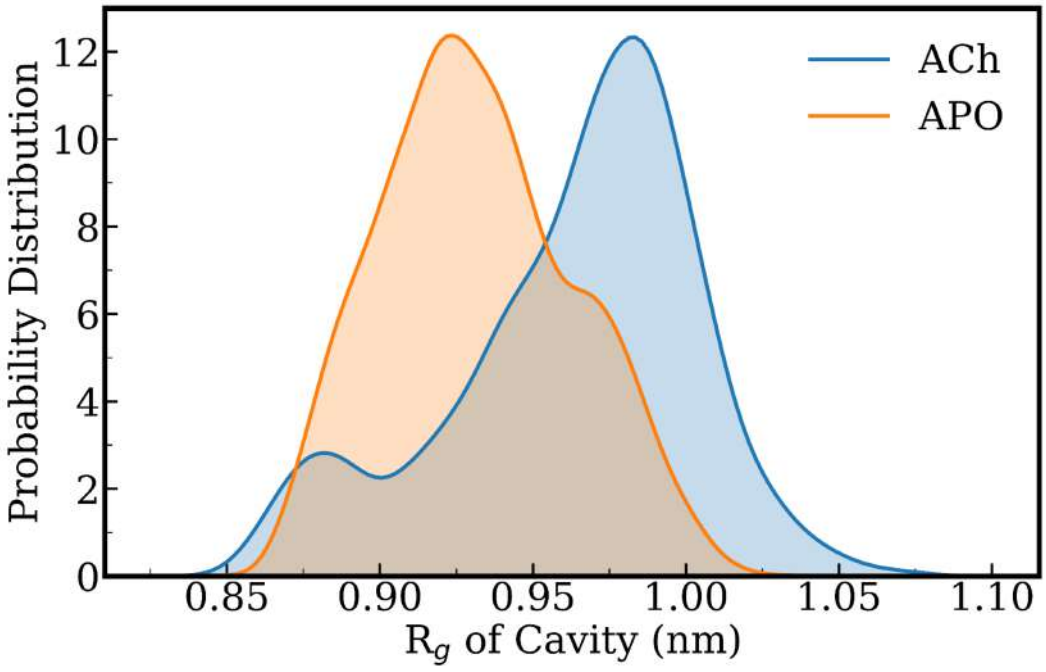

Figure 7.6: Comparison of the distribution of Rg of cavity residues for the Apo system (without substrate) and the system with ACh. The distributions indicate that both "open" and "closed" states exist in the presence and absence of ACh. However, the Apo system predominantly populates the "closed" conformation, while the presence of ACh significantly increases the population of the "open" conformation, suggesting a conformational selection mechanism for ligand binding.

that the hAChE enzyme's gorge can exist in two distinct conformations: "open" and "closed." The postulated cavity residues accurately describe these conformations.

7.3.3 Conformational Selection Model of Substrate Entry in AChE

An important question arises: does the "open" conformation occur only in the presence of a substrate or ligand, or does the substrate binding affect the protein's overall conformation? To address this, we compared the distribution of the radius of gyration of the cavity residues in both the Apo system (without substrate) and the system with the ACh molecule.The results indicate that both "open" and "closed" conformations are present in the absence (Apo system) and presence of ACh. However, there is a notable difference in the population distribution of these states. In the Apo system, the "closed" conformation is predominantly populated, suggesting that the enzyme tends to remain in a closed state when the substrate is absent. Conversely, in the presence of the ACh molecule, there is a significant increase in the population of the "open" conformation. The increased population of the "open" conformation in the presence of ACh suggests a conformational selection mechanism for ligand binding. In this model, the enzyme exists in equilibrium between "open" and "closed" states, but the ligand preferentially binds to and stabilizes the "open" conformation. This stabilization shifts the equilibrium towards the "open" state, increasing its population.

We performed Principal Component Analysis (PCA) using the coordinates of the cavity residues as input features. The analysis revealed that the first principal component (PC1) motion was highly correlated with the radius of gyration of the cavity residues, with a correlation coefficient(ρ) of approximately 0.6. To interpret the PC1 motion, we projected the PC1-eigenvectors onto the protein coordinates (Figure 7.7 a). This projection clearly indicates that the PC1 motion represents the opening of the gorge.

To further illustrate this, we examined two snapshots: one from the "open" state (Figure 7.7 b) and one from the "closed" state (Figure 7.7). In the "open" state snapshot, the substrate can directly enter the gorge and interact with the catalytic active site (CAS). Conversely, in the "closed" conformation, the substrate is obstructed by some

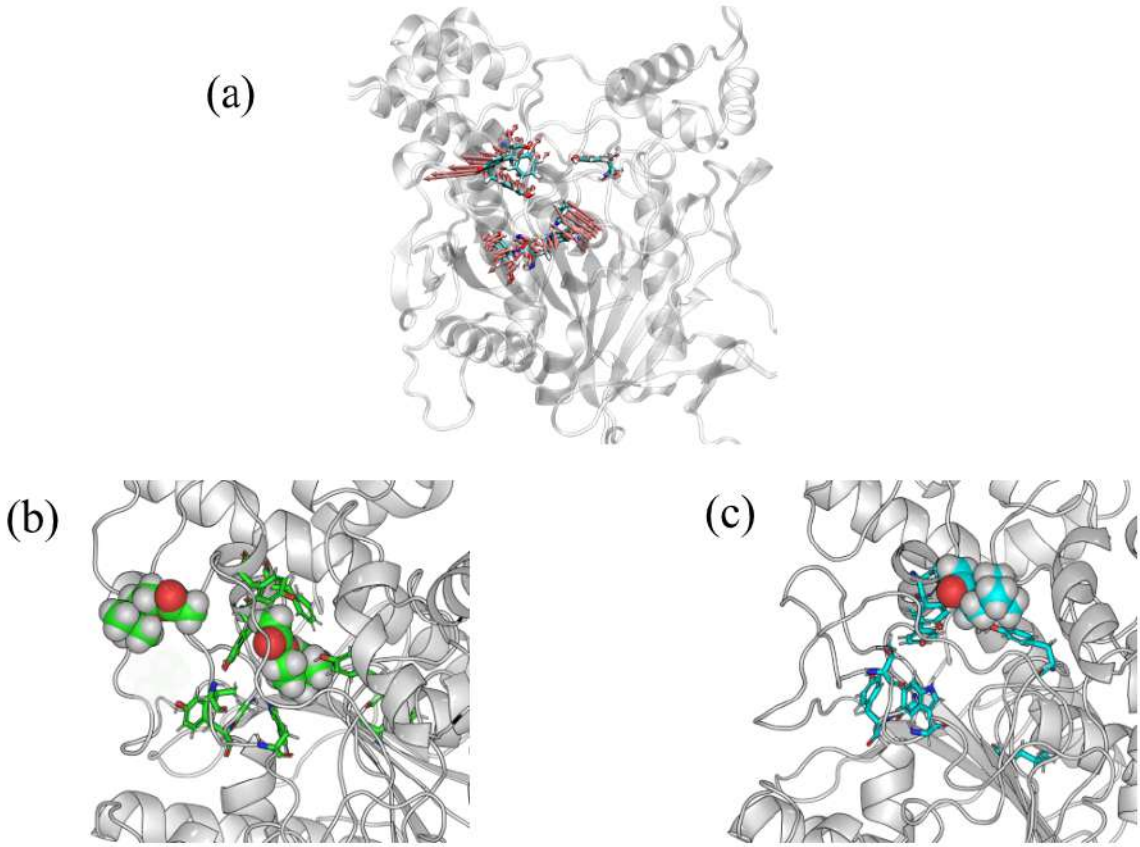

Figure 7.7: a) Projection of the PC1-eigenvectors onto the protein coordinates, illustrating the PC1 motion representing the opening of the gorge. b) Snapshot of the enzyme in the "open" state, where the substrate can directly enter the gorge and interact with the catalytic active site (CAS). (c) Snapshot of the enzyme in the "closed" state, where the substrate is blocked by some cavity residues and cannot enter the gorge to bind with the CAS.

cavity residues, preventing it from entering the gorge and binding with the CAS.

7.4 Conclusions

In this study, we investigated the dynamic mechanisms regulating the gorge opening motion of the human acetylcholinesterase (hAChE) enzyme. Our findings challenge the traditional markers used to describe gorge opening, specifically the distances between residues Y124-Y341 and Y124-F338. Through extensive molecular dynamics

simulations and principal component analysis, we identified a set of "cavity residues" that provide a more accurate description of the enzyme's conformational states.
Our analysis revealed that these cavity residues significantly influence the gorge opening motion. The radius of gyration (R_g) of these residues showed a clear distinction between the "open" and "closed" conformations of the enzyme. Notably, the presence of the acetylcholine (ACh) substrate increased the population of the "open" state, suggesting a conformational selection mechanism where the substrate preferentially binds to and stabilizes the open conformation of the enzyme.
The principal component analysis further supported our findings, indicating that the primary mode of motion (PC1) corresponds to the opening of the gorge. This motion was highly correlated with the radius of gyration of the cavity residues. Snapshots from our simulations illustrated that in the open state, the substrate can enter the gorge and interact with the catalytic active site (CAS), while in the closed state, the substrate is obstructed from binding to the CAS.
Our study enhances the understanding of the structural dynamics of hAChE and proposes a novel perspective on the enzyme's catalytic efficiency. By identifying critical residues involved in gorge opening, our findings provide valuable insights for the development of more effective acetylcholinesterase inhibitors, which are essential for the treatment of Alzheimer's disease.

Bibliography

[1] Knopman, D. S, Amieva, H, Petersen, R. C, Chételat, G, Holtzman, D. M, Hyman, B. T, Nixon, R. A, & Jones, D. T. (2021) Alzheimer disease. *Nature Reviews Disease Primers* **7**.

[2] BARTUS, R. T, DEAN, R. L, PONTECORVO, M. J, & FLICKER, C. (1985) The cholinergic hypothesis: A historical overview, current perspective, and future directions. *Annals of the New York Academy of Sciences* **444**, 332–358.

[3] Craig, L. A, Hong, N. S, & McDonald, R. J. (2011) Revisiting the cholinergic hypothesis in the development of alzheimer's disease. *Neuroscience & Biobehavioral Reviews* **35**, 1397–1409.

[4] Martyn, J. A. J, Fagerlund, M. J, & Eriksson, L. I. (2009) Basic principles of neuromuscular transmission. *Anaesthesia* **64**, 1–9.

[5] Rosenberry, T. L. (1975) Acetylcholinesterase. *Advances in enzymology and related areas of molecular biology* **43**, 103–218.

[6] Colovic, M. B, Krstic, D. Z, Lazarevic-Pasti, T. D, Bondzic, A. M, & Vasic, V. M. (2013) Acetylcholinesterase inhibitors: pharmacology and toxicology. *Current neuropharmacology* **11**, 315–335.

[7] Martis, E. A, Chandarana, R. C, Shaikh, M. S, Ambre, P. K, D'Souza, J. S, Iyer, K. R, Coutinho, E. C, Nandan, S. R, & Pissurlenkar, R. R. (2014) Quantifying ligand–receptor interactions for gorge-spanning acetylcholinesterase inhibitors for the treatment of alzheimer's disease. *Journal of Biomolecular Structure and Dynamics* **33**, 1107–1125.

[8] Silva, M. A, Kiametis, A. S, & Treptow, W. (2020) Donepezil inhibits acetylcholinesterase via multiple binding modes at room temperature. *Journal of Chemical Information and Modeling* **60**, 3463–3471.

[9] Colovic, M. B, Krstic, D. Z, Lazarevic-Pasti, T. D, Bondzic, A. M, & Vasic, V. M. (2013) Acetylcholinesterase inhibitors: Pharmacology and toxicology. *Current Neuropharmacology* **11**, 315–335.

[10] Colletier, J.-P, Fournier, D, Greenblatt, H. M, Stojan, J, Sussman, J. L, Zaccai, G, Silman, I, & Weik, M. (2006) Structural insights into substrate traffic and inhibition in acetylcholinesterase. *The EMBO Journal* **25**, 2746–2756.

[11] Van Belle, D, De Maria, L, Iurcu, G, & Wodak, S. J. (2000) Pathways of ligand clearance in acetylcholinesterase by multiple copy sampling. *Journal of Molecular Biology* **298**, 705–726.

[12] Xu, Y, Colletier, J.-P, Weik, M, Qin, G, Jiang, H, Silman, I, & Sussman, J. L. (2010) Long route or shortcut? a molecular dynamics study of traffic of thiocholine within the active-site gorge of acetylcholinesterase. *Biophysical Journal* **99**, 4003–4011.

[13] Cheng, S, Song, W, Yuan, X, & Xu, Y. (2017) Gorge motions of acetylcholinesterase revealed by microsecond molecular dynamics simulations. *Scientific Reports* **7**.

[14] ARIEL, N, ORDENTLICH, A, BARAK, D, BINO, T, VELAN, B, & SHAFFERMAN, A. (1998) The 'aromatic patch' of three proximal residues in the human acetylcholinesterase active centre allows for versatile interaction modes with inhibitors. *Biochemical Journal* **335**, 95–102.

[15] Bennion, B. J, Essiz, S. G, Lau, E. Y, Fattebert, J.-L, Emigh, A, & Lightstone, F. C. (2015) A wrench in the works of human acetylcholinesterase: Soman induced conformational changes revealed by molecular dynamics simulations. *PLOS ONE* **10**, e0121092.

[16] Gerlits, O, Ho, K.-Y, Cheng, X, Blumenthal, D, Taylor, P, Kovalevsky, A, & Radić, Z. (2019) A new crystal form of human acetylcholinesterase for exploratory room-temperature crystallography studies. *Chemico-Biological Interactions* **309**, 108698.

[17] Vanommeslaeghe, K & MacKerell, A. D. (2012) Automation of the charmm general force field (cgenff) i: Bond perception and atom typing. *Journal of Chemical Information and Modeling* **52**, 3144–3154.

[18] Abraham, M. J, Murtola, T, Schulz, R, Páll, S, Smith, J. C, Hess, B, & Lindahl, E. (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25.

[19] Klauda, J. B, Venable, R. M, Freites, J. A, O'Connor, J. W, Tobias, D. J, Mondragon-Ramirez, C, Vorobyov, I, MacKerell, A. D, & Pastor, R. W. (2010) Update of the charmm all-atom additive force field for lipids: Validation on six lipid types. *The Journal of Physical Chemistry B* **114**, 7830–7843.

[20] Huang, J, Rauscher, S, Nawrocki, G, Ran, T, Feig, M, de Groot, B. L, Grubmüller, H, & MacKerell, A. D. (2016) Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73.

[21] Bussi, G, Donadio, D, & Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**.

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

[23] Essmann, U, Perera, L, Berkowitz, M. L, Darden, T, Lee, H, & Pedersen, L. G. (1995) A smooth particle mesh ewald method. *The Journal of Chemical Physics* **103**, 8577–8593.

[24] Hess, B, Bekker, H, Berendsen, H. J. C, & Fraaije, J. G. E. M. (1997) Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463–1472.

8

Summary and Future Outlook of the Thesis

"You cannot teach a man anything; you can only help him discover it in himself"

Galileo

This thesis provides a comprehensive exploration of the molecular mechanisms underpinning protein-protein interactions and allosteric regulation, with significant implications for drug design. The research is segmented into four primary studies, each contributing unique insights into the field. In chapter 3 we focus on the phosphorylation mechanism of the Escherichia coli toxin HipA, specifically the autophosphorylation of Ser150. Through extensive molecular dynamics simulations and detailed structural analysis, the study reveals a complex multi-minima pathway for the autophosphorylation process. This pathway elucidates how buried residues can transiently become exposed, allowing phosphorylation to occur. This study challenges the traditional view of proteins having a single functionally important structure by highlighting the role of metastable and excited conformations in the

functional regulation of HipA. Understanding this intricate process provides a new perspective on how bacterial toxins are regulated, which could inform the development of novel antibacterial therapies targeting similar phosphorylation-dependent processes in pathogenic bacteria. In chapter 4 we explore the molecular mechanisms underlying the regulation of Rho GTPase interactions by phosphorylation, focusing on the dissociation processes of RhoA and Rac1 from RhoGDI complexes.

For RhoA, phosphorylation at Ser34 on RhoGDI induces significant structural changes that weaken the binding between RhoA and RhoGDI. These changes include alterations in the N-terminal region of RhoGDI and the polybasic region of RhoA, leading to the destabilization of the complex. The findings support the concept of a "phosphorylation code," where specific phosphorylation sites cause distinct structural and energetic responses, ultimately regulating the selective release of GTPases. This enhanced understanding of the regulatory mechanisms controlling RhoA-RhoGDI interactions provides valuable insights for potential therapeutic strategies targeting diseases involving dysregulated Rho GTPase signaling.

For Rac1, the research highlights the role of histidine protonation states in the phosphorylation-induced dissociation of the Rac1-RhoGDI complex. Phosphorylation at Ser101 and Ser174 on RhoGDI significantly shifts the protonation states of key histidine residues, leading to conformational changes and increased flexibility that reduce the binding affinity between Rac1 and RhoGDI. These changes disrupt critical interactions and the local electrostatic environment, contributing to complex destabilization. This study elucidates the intricate regulatory mechanisms by which phosphorylation and protonation states govern protein-protein interactions in Rac1 signaling, offering insights into therapeutic strategies for diseases characterized by aberrant Rac1 activity.

The research advances our understanding of the "phosphorylation code" and its implications for cellular signaling and disease, providing a foundation for the development of targeted therapeutic interventions.

In chapter 5 we propose a conceptual thermodynamic framework to identify al-

losteric and cryptic sites. The framework is based on the bidirectional nature of allostery, where binding at one site induces conformational changes at a distant site. We use this frame work to inhibit the PCSK9-LDLR interaction which is a promising therapeutic approach for hypercholesterolemia. The results indicate that binding of LDLR induces significant structural changes in the distal loop of PCSK9, with distinct conformational states (Conf-A and Conf-B) representing the apo and bound states, respectively. The study demonstrates that stabilizing specific loop conformations can effectively reduce the binding affinity of LDLR, supporting the potential of allosteric inhibition. The binding energy calculations using the MM-PBSA method and steered molecular dynamics (SMD) simulations confirm the reduced binding affinity in "apo" specific conformation validating our thermodynamic framework.

The proof-of-concept results for PCSK9-LDLR interactions demonstrate the feasibility of this approach in drug discovery. Future research will focus on identifying small molecule inhibitors to stabilize specific loop conformations and extending this methodology to other clinically significant PPIs.

In Chapter 6, we have conducted a comprehensive study on the substrate (acetylcholine, ACh) binding process of human acetylcholinesterase (hAChE). hAChE features two structurally significant regions: the catalytic active site (CAS) and the peripheral anionic site (PAS). This chapter elucidates the roles of these sites in the substrate binding process. Our investigation reveals that previously utilized distances (Y124-Y341 and Y124-F338) to describe the gorge opening motion are inadequate in capturing the true dynamic "breathing" motion of the gorge.

In Chapter 6, we have conducted a comprehensive study on the substrate (acetylcholine, ACh) binding process of human acetylcholinesterase (hAChE). hAChE features two structurally significant regions: the catalytic active site (CAS) and the peripheral anionic site (PAS). This chapter elucidates the roles of these sites in the substrate binding process. Our investigation reveals that previously utilized distances (Y124-Y341 and Y124-F338) to describe the gorge opening motion are inadequate in capturing the true dynamic "breathing" motion of the gorge.

These insights provide a refined understanding of the structural dynamics and func-

tional mechanisms of hAChE, offering potential avenues for the development of more effective therapeutic interventions targeting this enzyme.

A significant challenge we encountered throughout these studies was identifying an appropriate reaction coordinate capable of capturing biologically significant phenomena in complex biological systems. Traditional reaction coordinates, while effective for simpler systems, often fail to account for the intricacies inherent in more complex biological processes. As the complexity of biological systems increases, these traditional methods may not adequately represent the dynamic changes and intricate behaviors of molecules.

To address these challenges, researchers have increasingly turned to dimensionality reduction techniques such as Principal Component Analysis (PCA) and Time-Lagged Independent Component Analysis (TICA). These methods help identify proper reaction coordinates by reducing the dimensionality of the data, highlighting the most relevant features. However, these techniques are inherently linear transformations, which may not effectively capture the non-linear dynamics present in many biological systems.

In recent years, machine learning, particularly autoencoder-based approaches, has emerged as a promising tool for identifying suitable reaction coordinates in complex biological processes.

Autoencoders are particularly beneficial in molecular simulations because they can capture non-linear relationships and complex patterns within the data. By training an autoencoder on simulation data, the latent space can be interpreted as a non-linear reaction coordinate that captures the essential dynamics of the system. This approach can provide a more accurate and comprehensive understanding of molecular processes, especially those not well-described by linear methods.

In future research, we aim to leverage autoencoder-based techniques to identify proper reaction coordinates for understanding complex biological processes. By integrating these advanced machine learning methods, we hope to gain deeper insights into the dynamic behaviors of biomolecules, paving the way for more accurate modeling and potentially novel therapeutic strategies. This approach represents a significant step

forward in the study of complex biological systems, offering the potential to overcome the limitations of traditional and linear dimensionality reduction techniques.