Modeling Protein Dynamics and the Topological Assembly Pathway of the Multi-subunit Macromolecular Complexes

Jules Berlin Nde Kengne

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Margaret S. Cheung, Chair

Marcel Den Nijs

Jens Gundlach

Program Authorized to Offer Degree:

Physics

University of Washington

Abstract

Modeling Protein Dynamics and the Topological Assembly Pathway of the Multi-subunit Macromolecular Complexes

Jules Berlin Nde Kengne

Chair of the Supervisory Committee:
Margaret S. Cheung
Physics

Proteins are dynamic entities that constantly undergo conformational changes, influenced by factors such as temperature and ligand binding. Therefore, understanding protein dynamics, protein-protein interactions, and protein complex formation is essential for unraveling the intricate mechanisms behind biological functions. This thesis delves into the dynamic realm of the calcium-binding protein calmodulin (CaM), investigating the nuanced interplay between the structural flexibility and functional specificity of calmodulin as it interacts with the CaM-dependent protein kinase II (CaMKII) peptide. Employing an experimentally guided computational technique, the thesis aims to unravel the dynamic behaviors of CaM, the reciprocal relationship between calcium, CaM, and CaMKII peptide, which is a finely tuned signaling cascade essential for cellular function such as synaptic plasticity, learning and memory. Furthermore, protein-protein interactions are key determinants of cellular processes, enabling the formation of functional complexes and signaling pathways, and advancements in experimental and computational methods have facilitated the study of these interactions on a large scale. However, due to the transient nature of the interactions

and the high flexibility of the protein complexes and their components, using a single technique to investigate the three-dimensional structure of the protein complexes, which is paramount for deciphering the function of the complexes, becomes an enormous challenge. The thesis also describes an integrative modeling technique, which uses a combined experiment, computational, and machine learning approaches to determine the three-dimensional structure of macromolecular complexes, while using the INO80 chromatin remodeling complex as an example.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Almighty God, whose grace and blessings have guided me through every step in obtaining a Ph.D. in Physics.

It has been a long and arduous journey, with ups and downs and I would not have accomplished it without the countless individuals who have influenced and inspired me along the way. I am profoundly indebted to my advisor, Dr Margaret S. Cheung, for the exceptional mentorship, unwavering support, and invaluable guidance. Her wisdom, encouragement, and scholarly insights have been instrumental in shaping this thesis and my growth as a researcher. Heartfelt appreciation is extended to the members of my doctoral committee, Drs Ivana Bozic, Marcel Den Nijs, Jens Gundlach, Steve Sharpe, for their scholarly contributions, constructive feedback, and unwavering commitment to excellence. I am grateful to the current Cheung's group members, Chengxuan Li and Jiayi Wang, and the past members for their support and help. Special thanks to Catherine Provost for her encouragement and support. I am deeply grateful to the Department of Physics at the University of Washington, the Department of Physics at the University of Houston, and the Center for Theoretical Biological Physics (CTBP) at Rice University for providing the resources, facilities, and academic environment conducive to research and scholarly inquiry. The opportunities and support extended by the faculty, staff, and fellow students have been instrumental in facilitating this academic pursuit. Special thanks are extended to my collaborators Drs Peter G. Wolynes, Neal Waxham, Pengzhi Zhang, Mihaela Sardiu, Bernard Fongang whose

willingness to share their experiences and insights enriched these research findings. Their contributions are deeply valued and have significantly enhanced the quality of this thesis.

To my family, Mom Marguerite, Madelaine, Bernadette, Bernard, Simplice, Martiale, Josiane, Fany, as well as my extended family, I owe an immeasurable debt of gratitude. To my beloved wife Gaelle, whose unwavering love, endless patience, and steadfast support have been my rock throughout this journey. Her sacrifices and understanding have made all the difference. To my wonderful children, Aiden and Klervia, your boundless energy, laughter, and love have brought light and joy to my life, even during the most challenging times of this endeavor. I am also thankful to my friends for their understanding, encouragement, and companionship during the trials and triumphs of a Ph.D. in Physics. Their support and camaraderie have made this journey memorable and enjoyable.

To everyone not directly mentioned above who has played a role in this journey, either directly or indirectly, I extend my sincerest gratitude.

DEDICATION

To the loving memory of my dad papa Michel KENGNE.

Chapter 1. INTRODUCTION

Proteins are essential macromolecules in living organisms, playing a crucial role in various biological processes. They are involved in structural support, enzymatic catalysis, transport, signaling, and many other functions. Understanding protein dynamics, protein-protein interactions, and the formation of protein complexes is vital for unraveling the complexities of cellular processes.

1.1 GENERALITIES ON PROTEIN FOLDING, DYNAMICS, AND PROTEIN-PROTEIN INTERACTIONS

The study of proteins is a crucial aspect of biochemistry, molecular biology, and biophysics, as proteins play a fundamental role in the structure and function of living organisms [1, 2]. Proteins are large, complex molecules made up of amino acid chains, and they participate in virtually every biological process [2]. Protein synthesis constitutes two main steps: transcription and translation. During the transcription, the Deoxyribonucleic Acid (DNA) code is transcribed into messenger RNA (mRNA) in the nucleus; the mRNA is used in the translation process as a template to assemble a sequence of amino acids (Figure 1.1), forming a polypeptide chain (Figure 1.2). The polypeptide chain constitutes a linear sequence of amino acids, which is also known as the primary structure; the unique sequence determines the protein's identity and function [2-4]. The primary sequence folds into recurring patterns, such as alpha helices or beta sheets (Figure 1.3), driven by hydrogen bonding between amino acids; these patterns represent the secondary structure of a protein. The tertiary structure is further represented by the overall three-dimensional shape of a protein, resulting from interactions between amino acids that may be far apart in the primary

sequence (Figure 1.4). Some proteins are composed of multiple polypeptide chains (subunits) that assemble to form a functional protein, which is known as the quaternary structure (Figure 1.4**,** each color represents a single subunit**)**. A misfolding of a protein can lead to diseases such as cancer, Alzheimer's, and Parkinson's [5-7].

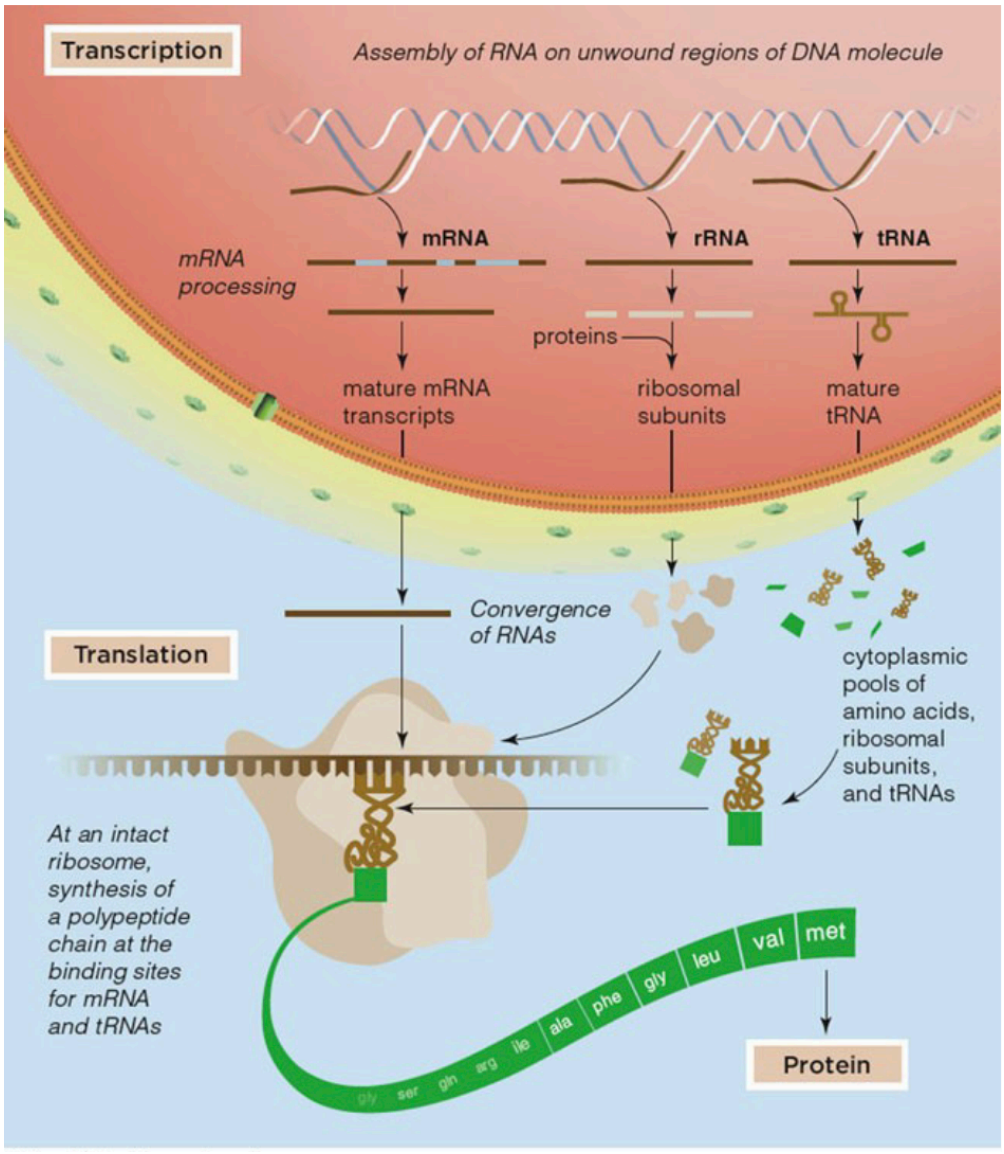

Figure 1.1. From DNA to protein: protein synthesis machinery. Adapted from [4].

Numerous investigations have demonstrated the importance of determining the structure of proteins [1, 8]. These structures dictate the functions, which are characterized as follows: enzymes, proteins that catalyze biochemical reactions, speeding up the conversion of substrates into products; structural proteins, which provide support and shape to cells and tissues (e.g., collagen in connective tissues); transport proteins, which facilitate the movement of molecules across cell membranes; antibodies, which play a key role in the immune system by recognizing and neutralizing foreign substances; hormones, which serve as signaling molecules that regulate various physiological processes; receptors such as calmodulin (CaM), which are proteins on cell surfaces that bind to signaling molecules and initiate cellular responses. [4, 9, 10] CaM binding to the signaling molecule calcium (Ca^{2+}) initiates a wide range of cellular processes as a result of the Ca^{2+}-induced conformational changes on CaM [11].

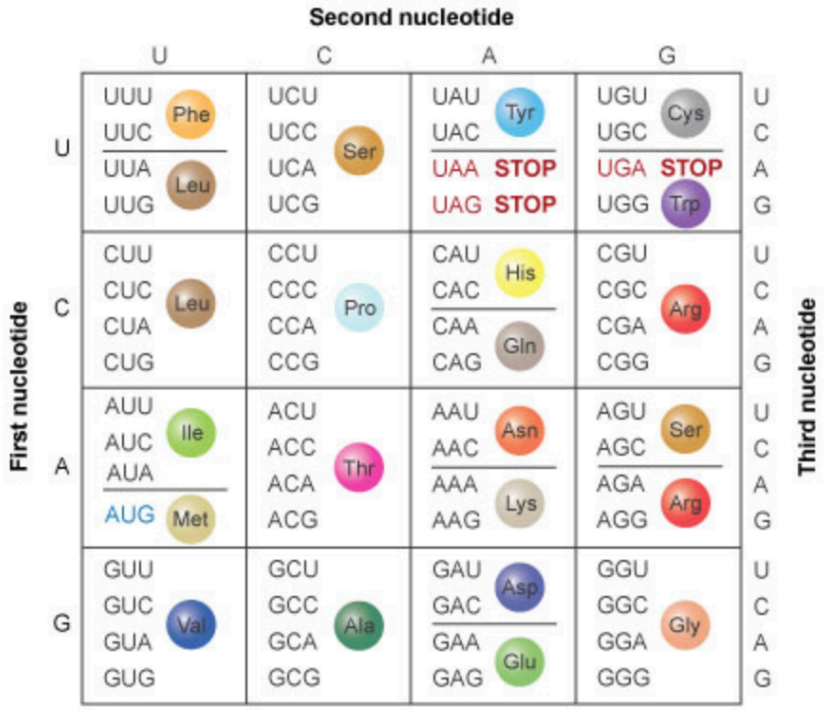

Figure 1.2. List of amino acids that constitute every protein from genetic codons [3].

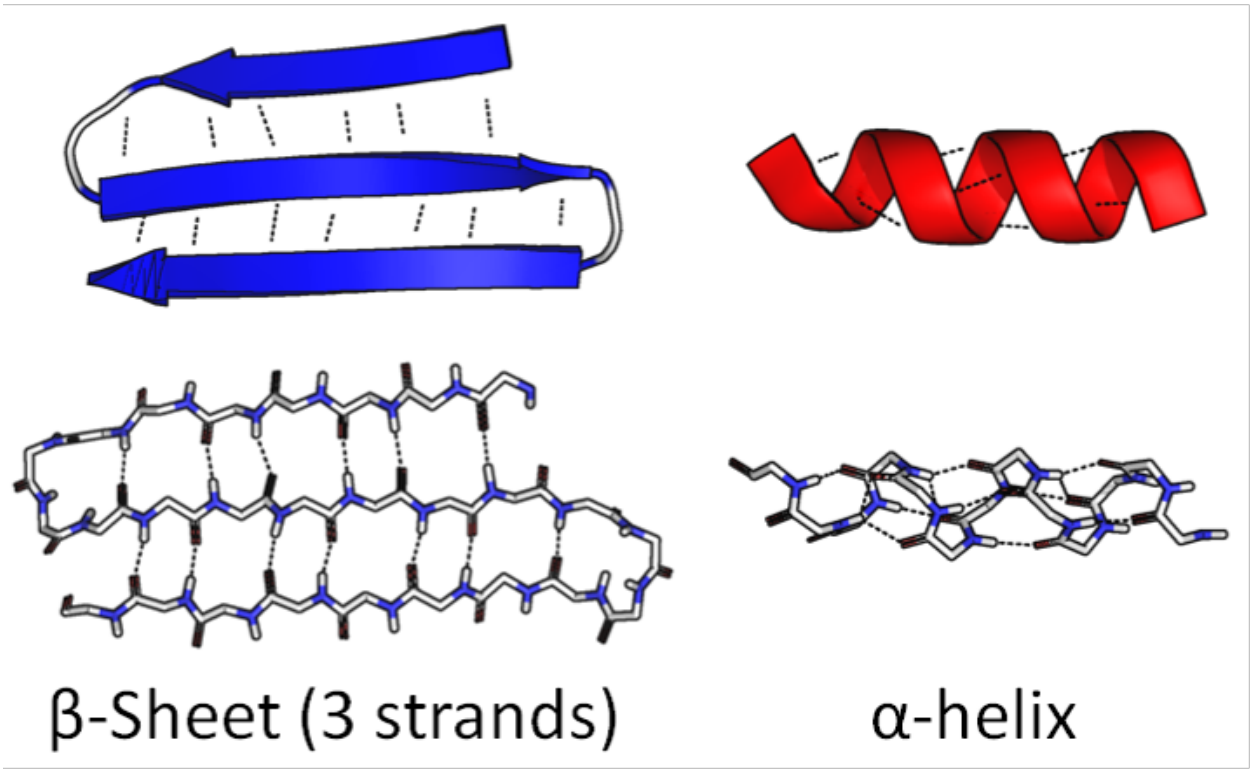

Figure 1.3. Secondary structure of the beta-sheet (left) and alpha-helix (right) [12] of a protein.

The versatility of CaM in binding to different targets highlights the dynamic and adaptable nature of protein-protein interactions and its interactions with various target proteins lead to the formation of functional protein complexes [13]. These complexes are involved in diverse biological functions, such as the regulation of muscle contraction, neurotransmitter release, and gene transcription [14]. Therefore, the study of CaM and its interactions serves as an illuminating example of how a single protein can exhibit dynamic structural changes, engage in diverse protein-protein interactions, and participate in the formation of functionally important protein complexes [15, 16]. Studying such cases, which are presented in Chapter 2 and Chapter 3, provides valuable insights into the molecular mechanisms underlying cellular processes and contributes to our broader understanding of protein biology.

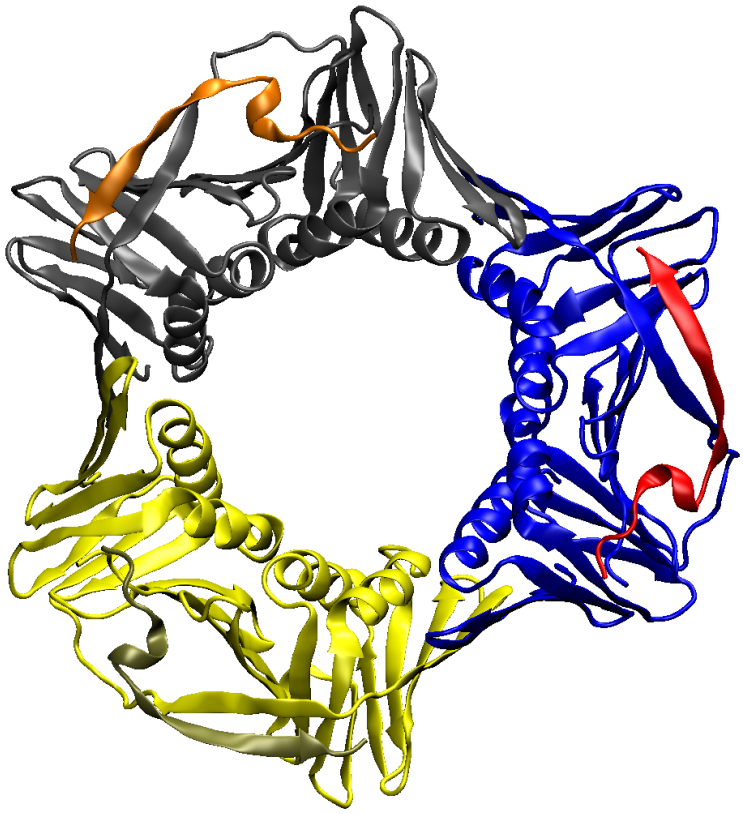

Figure 1.4. Example of the three-dimensional representation of a protein (quaternary structure) PDB ID 1AXC. Each color represents one subunit. [12]

1.2 GENERALITIES ON THE DYNAMICS OF CALMODULIN AND THE INTERACTIONS BETWEEN CALMODULIN AND CALMODULIN'S BINDING TARGET PROTEINS

Proteins are not static structures, they can adopt different conformations and undergo conformational changes in response to environmental cues, ligand binding, or interactions with other molecules [17, 18]. Examples of such highly dynamic proteins include CaM [19]. CaM is a highly conserved and multifunctional calcium-binding protein found in almost all eukaryotic cells [20]. It serves as a versatile regulator, responding to changes in intracellular calcium levels and plays a central role in calcium signal transduction by binding calcium ions and transducing the

calcium's effect to the downstream proteins [14]. The targets of CaM include enzymes, ion channels, and structural proteins [21]. The binding of CaM to its targets often induces conformational changes in both CaM and the target protein, thereby modulating the target's activity as well as the CaM's properties for calcium [22]. For example, during the regulation of the activities of kinases, CaM binding activates or inhibits their enzymatic activity. On the other hand, CaM's interactions with ion channels modulate their opening and closing, influencing ion flux across membranes [23, 24]. Understanding the intricacies of CaM's interactions with its protein targets requires studying the specific structural details of these interactions, including the regions on CaM and its targets responsible for binding. The dynamic nature of these interactions highlights the adaptability and responsiveness of biological systems to external signals and environmental changes.

One of the key targets of CaM is CaM-dependent protein kinase II (CaMKII), which is a serine/threonine kinase protein [15]. The CaM/CaMKII interaction forms a sophisticated signaling cascade that orchestrates cellular responses to calcium dynamics, which plays a pivotal role in neuronal signaling and synaptic plasticity, especially in long-term potentiation (LTP) [25]. The involvement in the modulation of synaptic strength underscores the importance of LTP in memory and learning. An overview of the interaction between CaM and CaMKII is as follows: calcium ions bind to the four calcium-binding sites on CaM when intracellular calcium levels increase; this calcium binding induces a conformational change in CaM, exposing its hydrophobic pockets. The conformationally changed CaM interacts with and binds to the regulatory domain of CaMKII, indeed, inducing a conformational change in CaMKII, which leads to its activation. This activation involves autophosphorylation of specific threonine residues within the kinase domain. As a result, the activated CaMKII can phosphorylate a variety of downstream target proteins, thereby

regulating their activity and participating in a myriad of biological functions [15, 22, 25]. Interestingly, the activated CaMKII reciprocally modulates the properties of CaM for calcium, which leads to feedback loops between calcium, CaM, and CaMKII. This reciprocal relationship, which is further discussed in Chapter 3, represents a finely tuned and adaptable signaling network crucial for the synaptic plasticity [15]. As calcium levels decrease, CaM undergoes another conformational change, leading to the non-spontaneous dissociation of CaM from CaMKII, which is aided by a third protein, neurogranin (Ng), that preferentially binds to CaM-free calcium [26-28]. This terminates the signaling process, and CaMKII returns to its inactive state.

1.3 GENERALITIES ON PROTEIN COMPLEXES

From the simplicity of a single protein to the elegance of a complex molecular ensemble, the journey within the cellular landscape is a testament to the beauty and sophistication of life at the molecular level. This journey from individual proteins to protein complexes is a fascinating exploration of the molecular choreography within living cells. Proteins, the workhorses of cellular function, rarely operate in isolation. Instead, they intricately collaborate, forming dynamic and highly organized structures known as protein complexes [29-31]. The transition from individual proteins to complexes involves a series of orchestrated steps, each playing a crucial role in cellular processes and functions.

A protein complex refers to a group of two or more associated proteins that interact physically and functionally to perform a specific cellular function (see Figure 1.5) [29, 31]. The protein complexes are often dynamic structures, meaning their composition and conformation can change in response to cellular signals or environmental conditions [29]. Studying these complexes provides insights

into how various cellular processes, including DNA replication, transcription, translation, signal transduction, and metabolism are regulated and executed. Dysregulation of protein complexes is implicated in various diseases, including cancer, neurodegenerative disorders, and metabolic diseases [32, 33].

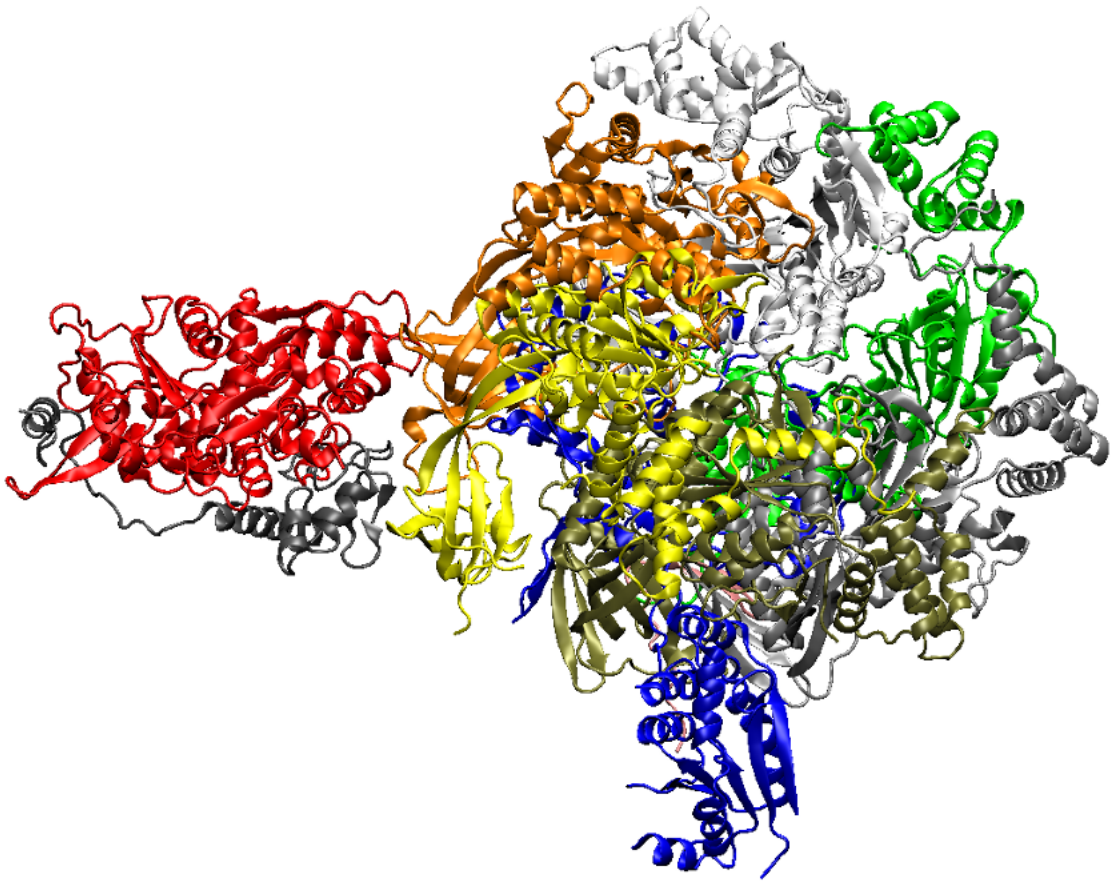

Figure 1.5. Example of protein complex with 8 different subunits. Each color represents one subunit. PDB ID 8ETS [34]

Protein complexes are formed through the protein-protein interactions mechanisms [32], which are the molecular handshakes that underpin cellular activities. The interactions can be transient, with proteins briefly coming together to fulfill a specific function, or more enduring, leading to the formation of stable protein complexes [29]. The concept of oligomerization adds another layer, as proteins may exist as monomers or assemble into complexes with multiple subunits. This assembly is not haphazard as it is often guided by the complementary shapes and charges on the interacting proteins. Protein complexes are not just conglomerates of proteins; they are molecular machines with specific functions [35-37]. The arrangement and composition of proteins within a complex dictate its function, and the dynamic nature of these assemblies allows for adaptability in response to cellular needs [29].

Cells contain a wide variety of protein complexes, each with a unique function. However, studying these protein complexes in more physiologically relevant conditions is challenging due to their dynamic nature and the technical difficulties associated with their isolation and characterization [31]. The stability also represents a key factor in the existence of the protein complexes, where some complexes are stable and long-lasting, while others are transient and form only when needed. Non-covalent forces, such as hydrogen bonds and hydrophobic interactions, contribute to their structural integrity. Additionally, post-translational modifications act as molecular switches, fine-tuning the stability and activity of these complexes in response to cellular signals. Novel protein complexes and their roles in cellular processes are continually being discovered, expanding the scope of research in the field. The investigation of intrinsically disordered proteins and their involvement in dynamic protein complexes is an evolving area of interest.

1.4 EXPERIMENTAL TECHNIQUES TO STUDY PROTEIN FOLDING, PROTEIN DYNAMICS, AND PROTEIN-PROTEIN INTERACTIONS

The study of proteins involves a variety of experimental techniques that allow researchers to investigate their structure, function, interactions, and expression levels [38, 39]. Table A.1 provides a comprehensive list of key experimental techniques used in protein science. These experimental techniques collectively provide a comprehensive toolkit to explore various aspects of proteins, advancing our understanding of their structures, functions, and roles in biological systems. While experimental techniques for studying proteins have advanced significantly, they come with certain limitations and challenges [39]. These limitations can impact the accuracy, resolution, and scope of the obtained data. Table A.2 provides a recapitulative of some known limitations of the experimental approaches to study protein as well as their implications. Researchers often address these limitations by employing a combination of complementary experimental techniques and computational methods, aiming to triangulate and validate findings. Despite these challenges, experimental techniques have significantly advanced our understanding of proteins, contributing to breakthroughs in various fields, including medicine and biotechnology.

1.5 COMPUTATIONAL TECHNIQUES TO STUDY PROTEIN FOLDING, PROTEIN DYNAMICS, AND PROTEIN-PROTEIN INTERACTIONS

Computational approaches play a crucial role in the study of proteins, offering valuable insights into their structure, function, interactions, and dynamics [40-43]. These methods leverage the power of physics principles, computers, and mathematical algorithms to analyze large datasets, simulate biological processes, and predict molecular behaviors [42, 44]. Table A.3 shows a list of key computational approaches used in the study of proteins. These computational approaches are

powerful tools that complement experimental techniques, providing valuable insights into the complex world of proteins. Their integration with experimental data contributes to a holistic understanding of protein structure, function, and interactions in various biological contexts. Computational techniques for studying proteins have made significant strides, but they also come with certain limitations and challenges. These limitations can affect the accuracy, reliability, and applicability of the results obtained through computational approaches. A list of known limitations and the associated implications is given in Table A.4. Despite these limitations, computational approaches are invaluable for generating hypotheses, guiding experimental design, and offering insights that complement experimental findings. Researchers in the field often integrate multiple computational approaches and experimental techniques to overcome the challenges and enhance the robustness of their conclusions.

1.6 BASICS OF MOLECULAR DYNAMICS SIMULATIONS OF PROTEINS

Molecular dynamics (MD) simulations represent a cornerstone in computational chemistry and biophysics, offering a unique window into the dynamic behavior of molecular systems [45-47]. MD simulations allow researchers to observe the movements and interactions of individual atoms and molecules over time, providing invaluable insights into the structure, dynamics, and thermodynamics of diverse systems [47, 48]. MD simulations of proteins were initially developed in the early 1980 [49] and have become increasingly useful in studying biological systems of biomedical interest, and not just in the study of model or toy systems [50]. The growing availability of experimentally determined protein structures and the emerging power of computer with the

availability of graphics processing units (GPUs) have increased its applications over years (Figure 1.6).

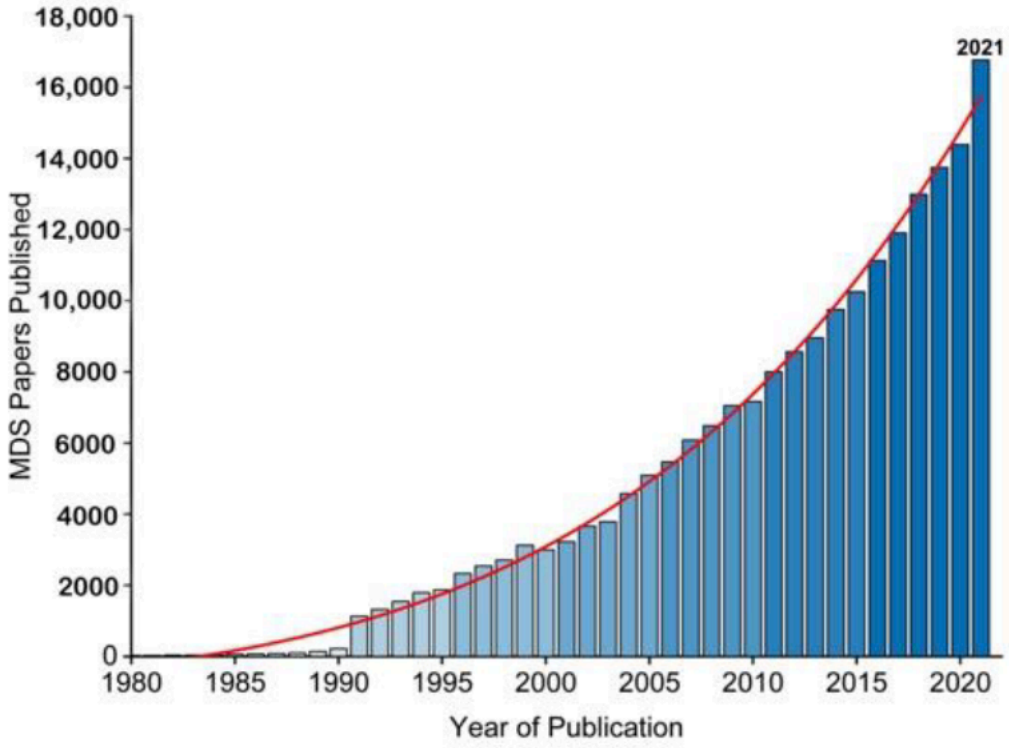

Figure 1.6.The growing use of MD simulations investigations over the years as reflected by publication (1980 – 2021). Adapted from [51].

At its core, MD simulations rely on Newton's laws of motion to predict the trajectories of atoms and molecules within a defined system. By numerically integrating these equations over time, one can simulate the behavior of complex molecular systems under varying conditions, such as changes in temperature, pressure, or the presence of solvent molecules. For a simple atomic system, the classical equations of motion are written as:

$$m_i \ddot{r}_i = f_i = -\frac{\partial u}{\partial r_i} \tag{1.1}$$

One wants to calculate the forces f_i acting on each atom, and these are usually from the potential energy $u(r^N)$, where $r^N = (r_1, r_2, \ldots, r_N)$ represents the complete set of 3N atomic coordinate. Determining $u(r^N)$ represents a major challenge as it involves the many body interactions. The set of N second order differential equations (Equation 1.1) are solved numerically at discrete time steps to determine the trajectory of each atom in the form,

$$r_i(t + \Delta t) = 2r_i(t) - r_i(t - \Delta t) + \frac{\Delta t^2}{m_i} F_i(t) \tag{1.2}$$

The Verlet algorithm, which shows that a new position depends only on the position and acceleration at time t and the position at time $t - \delta t$, is commonly utilized to determine the trajectory of each atom, starting from the basic kinematic equations.

$$r(t + \delta t) \approx r(t) + v(t)\delta t + \frac{1}{2} a(t)\delta t^2 \tag{1.3}$$

$$r(t - \delta t) \approx r(t) - v(t)\delta t + \frac{1}{2} a(t)\delta t^2 \tag{1.4}$$

By summing Equations 1 and 2, one gets

$$r(t + \delta t) \approx 2r(t) - r(t - \delta t) + a(t)\delta t^2 \tag{1.5}$$

In Equation 1.5, one can see that $a(t)\delta t^2 = -\nabla u(r_N)$ where $\nabla u(r_N)$ represents the force field. This term has been a subject of intense investigations, since the quality of the simulated results heavily depends on the force field used [50, 51]. This force field is a collection of equations and associated constants designed to reproduce molecular geometry and selected properties of tested structures (Figure 1.7). Since every protein system is described as a series of charged points

(atoms) linked by springs (bonds) in MD simulations (Figure 1.8), one uses a force field to determine the time evolution of bond lengths, bond angles and torsion, as well as non-bonding van der Waals and electrostatics interactions between atoms. Its expression is commonly given as,

$$u(r_N) = \sum_{bonds} k_i^{bond}(r_i - r_0)^2 + \sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2 + \sum_{dihedrals} k_i^{dihed}[1 + cos(n_i\phi_i + \delta_i)]\cdot + \sum_i \cdot \sum_{j \neq i} 4\epsilon_{ij}[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}] + \sum_i \cdot \sum_{i \neq j} \frac{q_i q_j}{\varepsilon r_{ij}}$$

This Equation 1.6 has four main parts: U_{bond}, which defines the oscillation about the equilibrium bond length; U_{angle}, which defines the oscillation of three atoms about an equilibrium angle; $U_{dihedral}$, which defines the torsional rotation of four atoms about a central bond; $U_{nonbond}$, which defines the non-bonded energy terms (electrostatics and Lenard-Jones). This equation and the corresponding bond or non-bonded are summarized in Figure 1.8 and Figure 1.9. The basic algorithm commonly used in most MD simulations is shown in Figure 1.10.

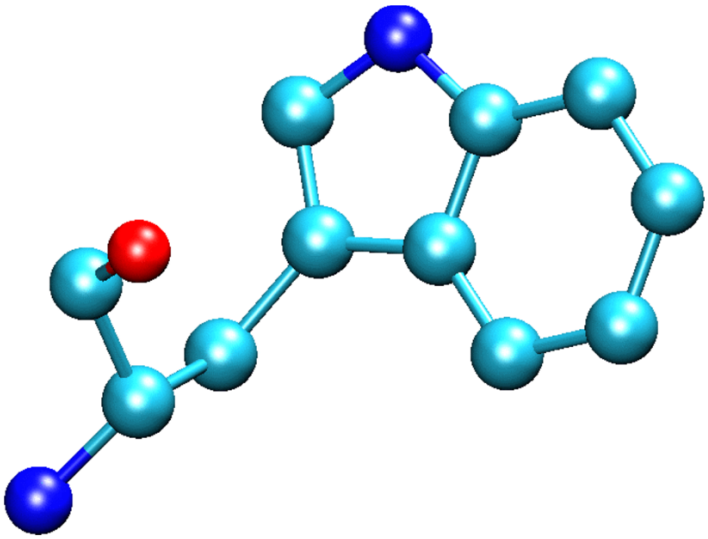

Figure 1.7. Geometry representation of the atoms that constitute an amino acid in protein [52].

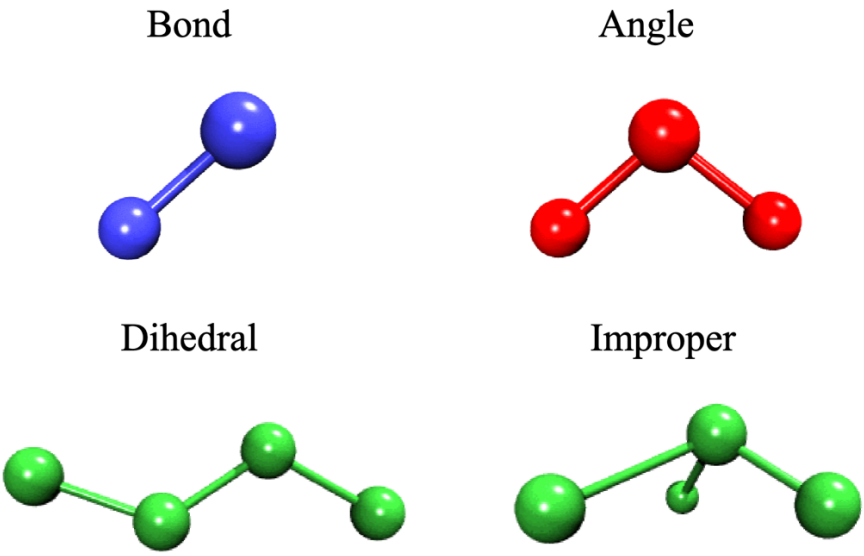

Figure 1.8. Schematic representation of the different types of interactions among atoms of the residues in protein [52].

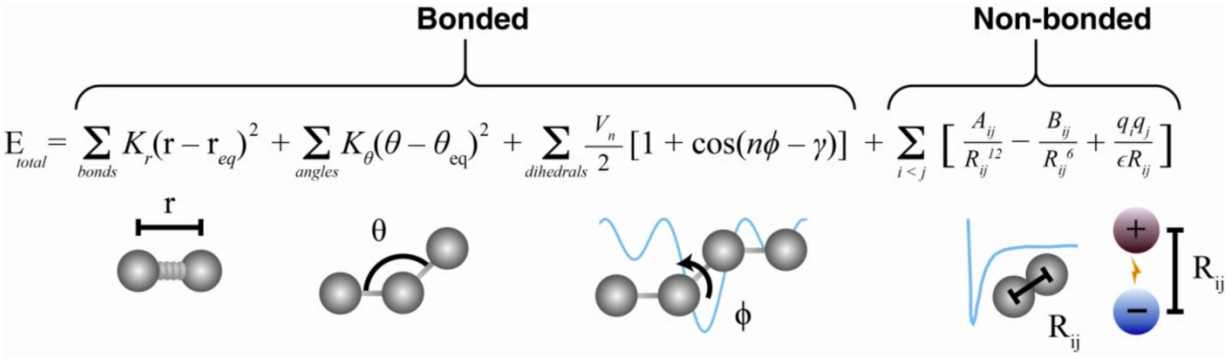

Figure 1.9. Summary of the interactions and the associated schematic representation [44].

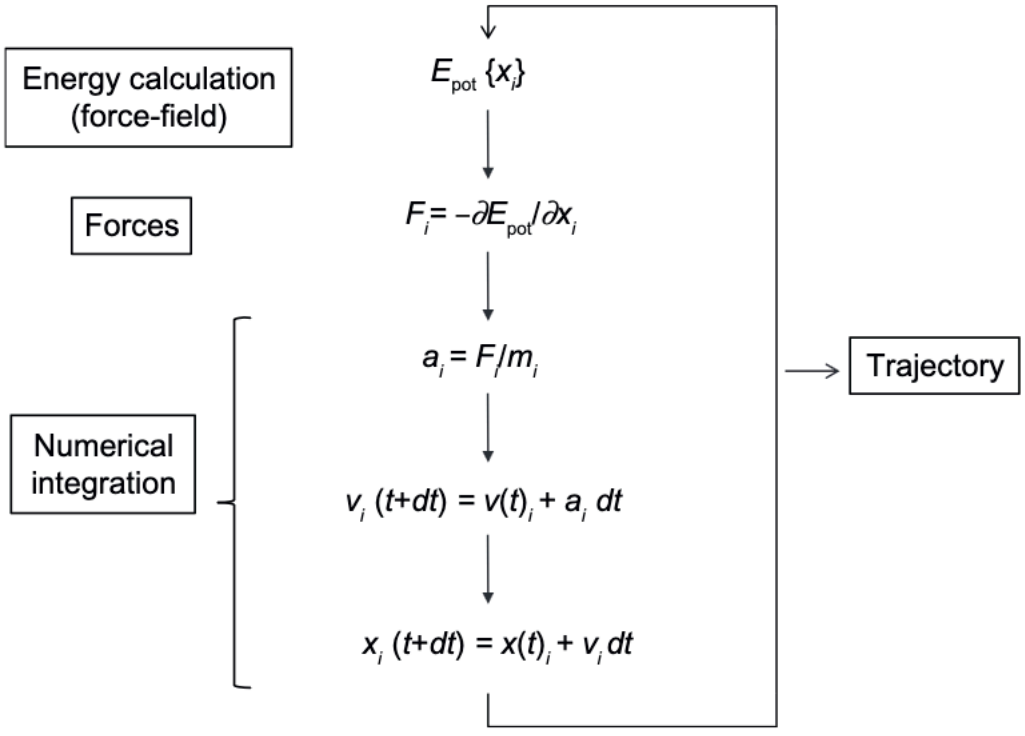

Figure 1.10. Basic algorithm commonly used in molecular dynamics simulations. [53]

The versatility of MD simulations spans across various scientific disciplines, including chemistry, biology, and materials science [53]. In chemistry, MD simulations elucidate reaction mechanisms,

solvent effects, and molecular properties, aiding in the design of new catalysts and understanding chemical processes. In biology, MD simulations shed light on protein folding, ligand binding, membrane dynamics, and other biologically relevant phenomena, offering insights that complement experimental findings [43]. Despite their computational demands and inherent limitations, MD simulations continue to push the boundaries of our understanding of molecular systems, offering testable hypotheses and guiding experimental investigations. Through interdisciplinary collaborations and integration with experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryogenic electron microscopy (cryo-EM), and single-molecule experiments, MD simulations contribute to a holistic understanding of molecular phenomena and drive innovation in fields ranging from fundamental science to applied research and drug discovery. Continued advancements in force field development, coupled with improvements in computational algorithms and hardware, have expanded the scope and accuracy of MD simulations, enabling the study of increasingly complex systems over longer timescales [54]. The various temporal and spatial scales of the MD simulations allows researchers to investigate a wide range of phenomena [54]. Figure 1.11 shows some of the temporal and spatial scales commonly encountered in MD simulations. These temporal and spatial scales provide a framework for understanding the capabilities and limitations of MD simulations in exploring various physical, chemical, and biological phenomena. The choice of timescale and spatial scale depends on the specific system under investigation and the scientific questions being addressed.

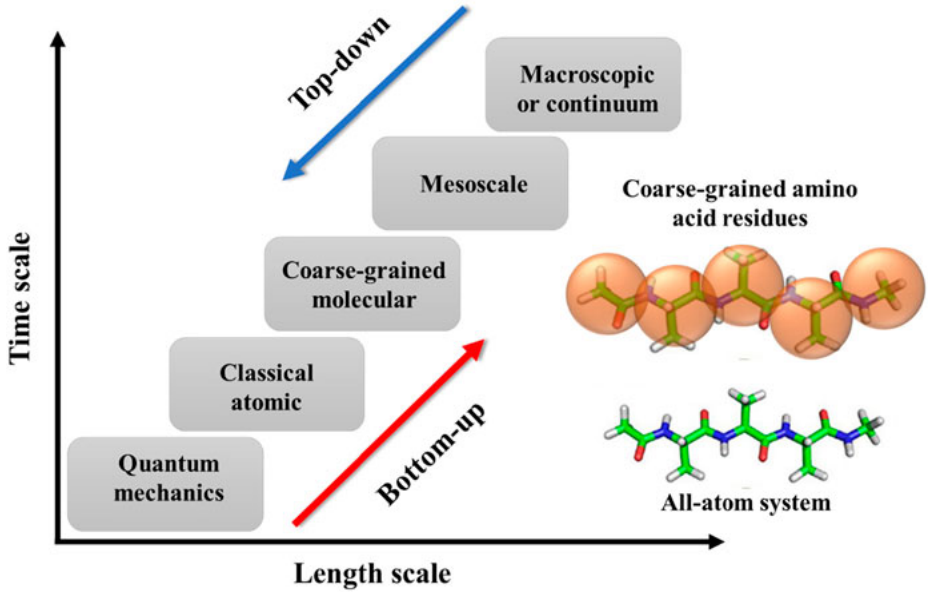

Figure 1.11. Schematic diagram to show different spatial and temporal scale in molecular dynamic simulations [54].

1.7 CONCLUSION

Proteins are dynamic molecules with diverse functions, and their interactions with each other (protein-protein interactions) lead to the formation of intricate structures known as protein complexes. Understanding protein dynamics and the intricacies of protein-protein interactions is crucial for unraveling the complexity of cellular processes. Their study is interdisciplinary, involving aspects of biology, chemistry, physics, and computational science and advances in protein research have profound implications for medicine, biotechnology, and our understanding of life processes at the molecular level.

The rest of the thesis is organized as follows: Chapter 2 delves into the dynamic behavior of proteins, focusing on calmodulin as a prime example. It explores the fundamental principles of protein dynamics, including conformational changes and flexibility, while highlighting the

significance of these dynamics in biological processes. Through a review of studies, experiments, and computations, it aims to provide insights into the dynamic nature of calmodulin and its implications in cellular signaling and regulation using a coarse-grained molecular simulations approach. Chapter 3 centers on the interplay between calmodulin and CaMKII. It underscores the importance of protein-protein interactions in cellular signaling and regulation, with a specific focus on the partnership among calcium, calmodulin and CaMKII. By reviewing relevant studies and experiments, it aims to elucidate the complex reciprocal relationships among calcium, calmodulin, and CaMKII, shedding light on their functional consequences in biological processes such as synaptic plasticity. Chapter 4 introduces integrative modeling as a powerful approach to studying protein complexes. It outlines the methodologies involved in integrative modeling, spanning computational techniques, structural biology, and biochemical assays. Through the INO80 chromatin remodeling complex case study, it demonstrates the application of integrative modeling in deciphering the three-dimensional structure, dynamics, and function of protein complexes, particularly those containing transient interactions with unknown PDB structures. It also discusses the potential of integrative modeling in advancing our understanding of protein biology and proposes future developments in the field. The final chapter synthesizes the main findings from each preceding chapter and their contributions to the broader field of protein biology. It reflects on the implications of the research presented in this thesis, emphasizing its significance in advancing our understanding of protein dynamics, interactions, and integrative modeling. Furthermore, it discusses potential avenues for future research, highlighting unresolved questions, methodological advancements, and new directions for investigation.

Chapter 2. COMPUTATIONAL APPROACH TO STUDY CALCIUM-INDUCED CONFORMATIONAL CHANGES ON CALMODULIN

2.1 INTRODUCTION

Calcium ion (Ca^{2+}) is the most abundant metal ion [55] that is involved in a myriad of cellular processes [11, 56, 57]. These processes include cellular motility, neurogenesis, memory formation, muscle contraction, and neuronal transmission [57-59]. Ca^{2+} binds to effector proteins, which are responsible for stimulating many Ca^{2+}-dependent biological functions [57, 59-62]. One such effector protein is calmodulin (CaM). CaM is composed of two globular domains separated by a central linker. The two domains (N- and C-domains) are constituted of two calcium binding motifs each, named helix-loop-helix (EF-hand) motif [63, 64]. Upon sufficient increase in Ca^{2+} concentration, Ca^{2+}-free CaM (Figure 2.1 (A)) [65] is activated and transitions to the Ca^{2+}-loaded CaM (Figure 2.1 (B)) [63]. This transition is accompanied by large conformational changes, which expose the target binding surfaces of CaM to the solvent (Figure 2.2) [8, 18, 20, 64, 66, 67] to regulate the activity of a large array of different target proteins [62, 68].

CaM is a Ca^{2+}-binding protein that is present in all eukaryotic cells [21, 62, 69]. The crystal structure of the Ca^{2+}-CaM shows a dumbbell-like structure with a central α-helical linker (Figure 2.1 (B)) [63]. However, in solution, the central linker shows higher flexibility [16, 70-73], which enables the two globular domains of CaM to orient independently from each other [73-75]. Moreover, this flexibility of the central linker could justify the compact structure of the Ca^{2+}-CaM (Figure 2.1 (D)) upon target binding [67, 74]. Additionally, Ca^{2+}-CaM has been shown to crystallize in the collapsed conformation in the absence of the target (Figure 2.1 (C)), indeed providing more evidence of the backbone plasticity of the Ca^{2+}-CaM [76] and domain-domain

interactions. A large and growing body of literature has investigated the dynamics of CaM upon calcium binding [16, 72, 73, 77]. While a few studies have shown that the two domains of the Ca^{2+}-CaM free target interact with one another in solution with a percent ratio of 5-20 % [16, 72], these observations remain debatable [18, 74].

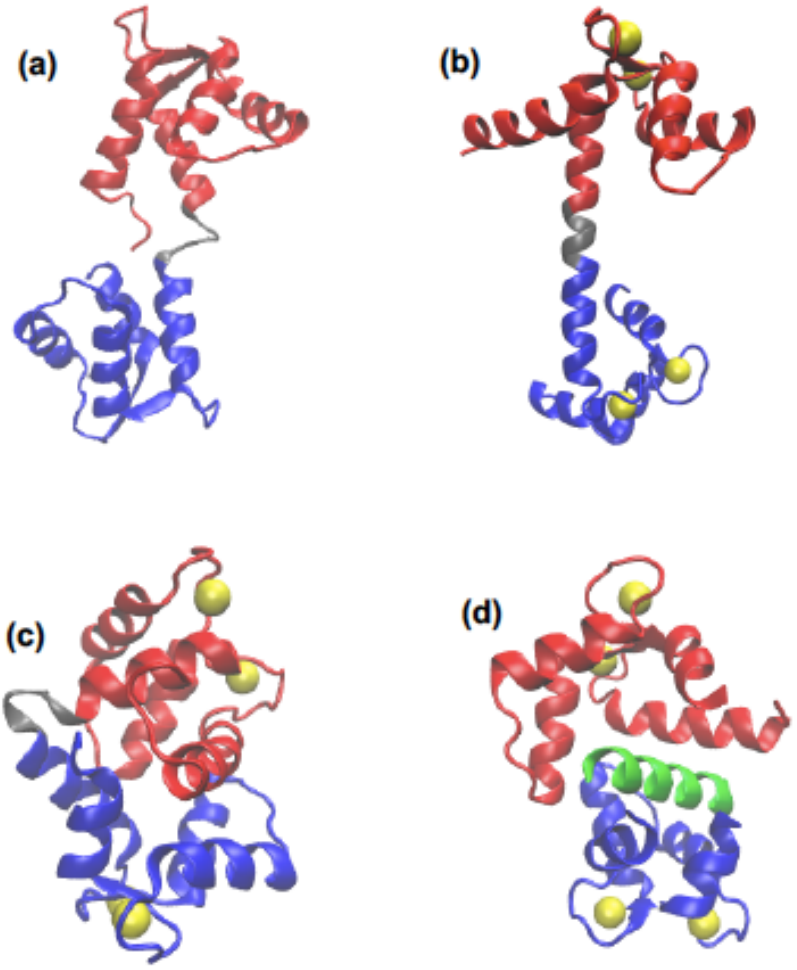

Figure 2.1. Representative conformations of calmodulin (CaM). (a) Ca^{2+}-free CaM, or apoCaM (PDB ID: 1CFD); (b) Ca^{2+}-bound CaM, or holoCaM, in the extended conformation (PDB ID: 1CLL [63]); (c) holoCaM in the collapsed conformation without a target peptide (PDB ID: 1PRW); (d) holoCaM with a target peptide (PDB ID 1CDM). We visualized the structures using the Visual Molecular Dynamics (VMD) program [52]. Ca^{2+} ions are shown in yellow spheres, the N-terminal

and C-terminal domains of CaM are shown in red and blue respectively, the central linker is shown grey, and the target peptide is in green. [14]

A combined approach of experiments and computational simulations have shown that both Ca^{2+}-CaM and its target undergo mutually induced conformational changes upon binding for recognition [78]. This theoretical model, however, has yet to illustrate how CaM's affinity for Ca^{2+} is tuned by its target binding [62, 68, 73, 79-82]. One of the main challenges for developing an appropriate force field involving Ca^{2+} is that CaM is a structurally flexible protein about its central linker while the binding of Ca^{2+} to the two domains of CaM increases the rigidity of the local domains. A compromise between rigidity and softness must be considered for CaM to accurately select, bind, and activate its binding targets.

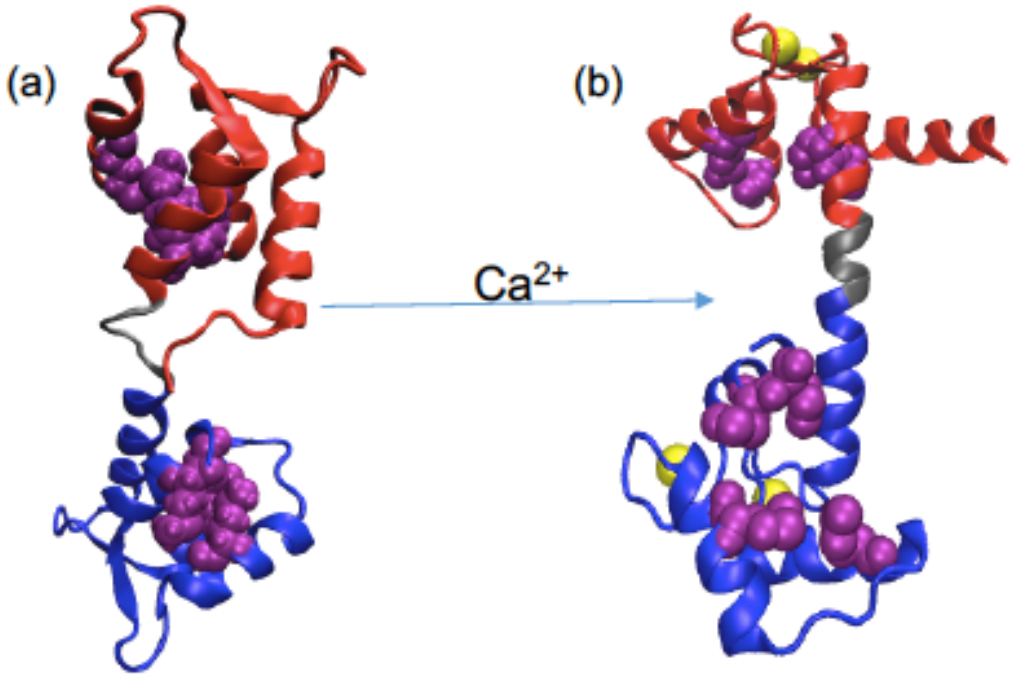

Figure 2.2. Ca^{2+} induced conformational change. (a) apoCaM (b) holoCaM. Ca^{2+} ions are shown in yellow spheres, the N-terminal and C-terminal domains of CaM are shown in red and

blue respectively, the central linker is shown grey. Upon conformational change, the four methionine residues (shown as beads in purple), initially buried inside the EF-hands, are exposed to the solvent. Adapted from. [14]

A myriad of computational approaches, ranging from all-atom to coarse-grained [45, 67, 73, 80, 83-93] have been employed to study the dynamics of CaM. In those approaches, CaM is treated as a folded protein [74, 87, 93] or a folded protein with intrinsically disordered regions [45], where in most cases Ca^{2+} effects are ignored for simplicity [45, 74, 83, 84]. A Lennard-Jones potential is used for fixing the position of a particle in the modified versions of the abovementioned approaches to account for such effects [91, 92]. When Ca^{2+} ions are implicitly included, their charges are often distributed to all the negatively charged residues present in the EF-hand loops [90] without any consideration of calcium's coordination geometry. A previous study has proposed a model to study the dynamics of CaM that accounts for both considerations, but it lacks the ability to reproduce the reciprocal relationship between calcium binding and target binding [26, 78].

How to properly include the calcium ion in the force field of a coarse-grained model is a crucial step since this would allow one to understand the relationship between calcium binding to CaM and binding of a target [78, 94]. Particularly, the coordination chemistry of calcium ions and the structural flexibility allowing CaM to flex according to its binding targets and surrounding environment is important in modeling the adaptive behavior of CaM. Motivated by this need, we expanded the capability of the coarse-grained model by upgrading our previous models [26, 95] to the Associative memory, Water mediated, Structure, and Energy Model (AWSEM), which is aided by the experimentally determined structures, and uses the energy landscapes theory to explore more biologically relevant conformations of CaM. Although in separate studies, there have been development in AWSEM force fields on structurally flexible proteins [96] or on di-valent ions

[97], both features need to be combined for modeling CaM. By considering the coordination chemistry of the Ca^{2+} ions as well as the flexible nature of the central linker of CaM, our proposed model for Ca^{2+}-CaM elucidates the many-body effects of calcium binding on the geometrical shape of the calcium binding loop and allows large conformational changes for binding various targets.

2.2 COARSE-GRAINED SIMULATIONS USING THE AWSEM FORCE FIELD TO EXPLORE THE FOLDING AND TARGET BINDING ENERGY LANDSCAPE OF CAM.

We used the AWSEM force field [43, 97] to build the model of Ca^{2+}-CaM. AWSEM is a transferable coarse-grained protein model, which uses three beads (C_α, C_β, and O atom from the peptide bond) to represent each amino-acid residue (except glycine, which lacks C_β). The positions of all the rest of other atoms along the backbone are deduced by assuming an ideal geometry of the system. The total energy function used in AWSEM is given in Equation 2.1

$$V = V_{BB} + V_{PMF} + V_{DH} + V_{FM} \tag{2.1}$$

Which is composed of two main terms: the physics-based terms, $V_{BB} + V_{PMF} + V_{DH}$, and the knowledge-based term, V_{FM}.

(a) The backbone term, V_{BB}, from Equation (2.1), maintains the backbone geometry of protein. Its expression is given in Equation 2.2,

$$V_{BB} = V_{con} + V_{chain} + V_{\chi} + V_{rama} + V_{excl} \tag{2.2}$$

V_{con} and V_{chain} represent the connectivity and chain terms, respectively. The former links neighboring residues, while the latter maintains ideal bond angles around the C_α atoms of each residue. The chirality potential, V_χ, which is applied to all residues except glycine, maintains the local residue chirality. V_{rama} represents the Ramachandran potential, which biases the protein chain conformation toward the allowed Ramachandran regions. The excluded volume potential, V_{excl}, is used to avoid chain collapse and unphysical entanglements.

(b) The many-body potential of mean force V_{PMF}, from Equation 2.1 considers the nature of the interacting residues, and its expression is given in Equation 2.3.

$$V_{\mathrm{PMF}} = V_{\mathrm{contact}} + V_{\mathrm{burial}} + V_{\mathrm{HB}} \tag{2.3}$$

The contact term, V_{contact} represents the pairwise additive and the many-body terms that consider interactions between residues far apart in sequence. The burial term, V_{burial} accounts for a particular residue's type propensity. The hydrogen bond term, V_{HB}, is composed of three other terms, which are given in Equation 2.4,

$$V_{HB} = V_\beta + V_{P-AP} + V_{helical} \tag{2.4}$$

Where V_β and V_{P-AP} are for the hydrogen bonding interactions in β-sheet conformations. The former stabilizes an already formed β hydrogen bonding, while the latter allows a protein to undergo parallel and anti-parallel β-sheet conformations before the stabilization of the hydrogen bond. The helical term, $V_{helical}$, ensures the formation of alpha-helices.

(c) The electrostatic interactions in solution with implicit solvent is described by the Debye-Hückel term [97] V_{DH} from Equation 2.1. The expression is given in Equation 2.5,

$$V_{DH} = K_{Elect} \sum_{i<j} \frac{q_i q_j}{\varepsilon_r r_{ij}} e^{-\frac{r_{ij}}{l_D}} \tag{2.5}$$

Where q_i and q_j represent the charges of beads i and j, and r_{ij} is the distance between the two beads; $K_{Elect} = (4\pi\varepsilon_0)^{-1} = 333.24\ kcal\ mol^{-1}e^{-2}$Å; ε_r represents the dielectric constant of the media; and l_D is the Debye-Hückel screening length, which is given by $l_D = \sqrt{\varepsilon_r \varepsilon_0 k_B T / 2e^2 I}$. k_B is the Boltzmann constant, *T* is the temperature, *e* represents the elementary electric charge, and *I* is the ionic strength of the implicit solvent.

(d) V_{FM} from Equation 2.1, is a bioinformatic fragment memory potential that structurally biases short fragments of the protein chain, typically composed of 3-9 residues at a time, towards conformations that are based on "memory" structures. As listed in Table 2.1, the memory structures selected from the Protein Data Bank (PDB) were used to speed-up the conformational search of the native states. Its expression is given in Equation 2.6,

$$V_{FM} = -\lambda_{FM} \sum_{m=1}^{60} \cdot \sum_{n=1}^{3} w_n^m \sum_{i,j \ni 3 \le |i-j| \le 9} \gamma_{ij} exp[-\frac{(r_{ij} - r_{ij}^m)^2}{2\sigma_{ij}^2}] \tag{2.6}$$

Where the outer sum is over the aligned memory (from m = 1 to 60), the sum in the middle with a variable *n* is over the three segments of CaM (n=1 is N-domain, n=2 is central linker, and n=3 is

C-domain), and the inner sum is over all possible pairs of C_α and C_β atoms within the memory fragment that are separated by two or more residues; |i-j| ≤ 9 gives the maximum sequence separation of interacting residues, which is either the length of the memory or the maximum cutoff, whichever is shorter. r_{ij} and r_{ij}^m represent the instantaneous distance and the corresponding distance in the memory fragment between the atoms i and j; γ_{ij} is the residue type dependence interaction strength; σ_{ij} represents the sequence separation dependent width. Its expression is given by $\sigma_{ij}=|i-j|^{0.15}$.

We used 60 non-redundant structures of CaM taken from [98] (Table 2.1), either free or in complex with the target peptides, to build the memory fragments for the coarse-grained model of CaM. We used the 60 sequence homologs of Ca^{2+}-CaM since We were interested in studying the dynamics and not the folding properties of CaM. We divided each memory into three segments: (1) N-domain of CaM, from residue 5 to residue 76; (2) the flexible central linker, from residue 77 through residue 81; (3) C-domain of CaM, from residue 82 to residue 147 (as illustrated in Figure 2.1). The weight of each segment is controlled by a memory weight parameter w_n^m and a global parameter λ_{FM} (Equation 2.6). Because the central linker has its own memories of large variations in the structures, essentially it was modelled with more structural flexibility than the two globular domains of CaM.

We summarized CaM models using various values of λ_{FM} and w_n^m parameters in Table 2.2 for the molecular simulations. All other parameters for AWSEM were set by default as defined in the original ASWEM study [43, 97]. The parameters of the default setting are also provided in appendix (B 1).

Table 2.1. The PDB IDs of the 60 non-redundant CaM memory structures for the AWSEM model and the R_g values of the CaM. In the case of CaM complexes, only the coordinates of CaM were used for memory. m is the index for the memories. [14]

m	PDB ID	R_g (Å)	m	PDB ID	R_g (Å)	m	PDB ID	R_g (Å)
1	1PRW	14.60	21	3GP2	15.96	41	2BBM	17.03
2	3EWT	15.17	22	3DVM	16.00	42	2BCX	17.25
3	2LGF	15.23	23	1CKK	16.05	43	2KNE	18.67
4	3DVJ	15.31	24	2FOT	16.05	44	4DCK	18.87
5	3BYA	15.38	25	2JZI	16.12	45	1SK6	19.30
6	2LHI	15.38	26	1CM1	16.20	46	1XFU	19.31
7	2HQW	15.41	27	3DVK	16.22	47	2L1W	19.46
8	2L7L	15.43	28	2VAY	16.26	48	1K9O	19.46
9	3EWV	15.52	29	2O60	16.28	49	1PKO	19.49
10	1IWQ	15.57	30	2WEL	16.35	50	1S26	19.50
11	1L7Q	15.59	31	1QTX	16.43	51	1LVC	19.70
12	3DVE	15.61	32	2O5G	16.47	52	1CFF	19.78
13	1CDM	15.66	33	1QS7	16.50	53	4G27	20.79
14	3SUI	15.68	34	3GOF	16.50	54	2L53	21.00
15	3BXL	15.80	35	3BXK	16.51	55	2KDU	25.65
16	1IQ5	15.90	36	1MXE	16.55	56	4EHQ	22.07
17	1ZUZ	15.93	37	2Y4V	16.62	57	4DJC	22.31
18	3HR4	15.94	38	2BE6	16.65	58	2YGG	23.66
19	2KOF	15.95	39	2F3Y	16.68	59	1G4Y	25.32

20	2LL6	15.95	40	1SY9	16.83	60	1CLL	21.80

Table 2.2. Summary of Ca^{2+}-CaM models with different values of the global and local memory weight and the average R_g ($\overline{R_g}$) and ratio between the collapsed and extended states ($\gamma_{c:e}$) from the umbrella sampling simulations. The models that best reproduce experimental results from the simulations is highlighted. The experimental measured R_g = 20.0 ~ 22.5 Å and $\gamma_{c:e}$ ~ 0.11 [14]

Models	λ_{FM}	W_n^m				$\overline{R_g}$ (Å)	$\gamma_{c:e}$
		W_n^{others}	W_1^{1CLL}	W_2^{1CLL}	W_3^{1CLL}		
I.1	1	1				22.05	0.001
I.2	0.1					20.51	0.48
I.3	0.01					19.86	0.64
II	0.1	1	2			20.34	0.59
			3			20.66	0.47
			4			20.43	0.50
			5			20.70	0.37
			6			20.52	0.38
			7			21.11	0.16
			8			21.15	0.14
			9			21.26	0.10
III	0.1	1	5	1	5	19.91	0.73
				2		19.70	0.82
				3		19.84	0.77
				4		20.03	0.58
				6		20.43	0.48
				7		21.20	0.16
				8		21.35	0.12
				9		21.59	0.07

2.3 MODELING CALCIUM IONS IN AWSEM FORCE FIELD

Calcium binding proteins in most cases contain the EF-hand calcium binding motif which is formed by the pentagonal bipyramidal geometry (Figure 2.3 & Figure 2.4) [99, 100]. The literature on modeling those proteins, in general, does not clearly explain how the effects of the Ca^{2+} ions are considered during the simulations. We developed an implicit Ca^{2+} model that evenly splits the

Ca^{2+} charges to the selected residues. Partial charges on the sidechain beads of the negatively charged residues that coordinate Ca^{2+} collectively in the Ca^{2+} binding loops (Figure 2.3) were adjusted to reflect the calcium binding effect (Table 2.3). To note, in the Ca^{2+} binding loops in CaM, the charges on the negatively charged residues that do not coordinate Ca^{2+} remain -1 e. This simple approach incorporates the coordination geometry of the Ca^{2+} and allowed us to improve the modeling of many-body effects in the local Ca^{2+} binding loops which further controls the open/close state of the helix-loop-helix EF-hand and the global conformation of calmodulin.

Table 2.3 The adjusted partial charges of the negatively charged residues that participate in calcium coordination in the four calcium-binding loops of CaM; the amino-acid sequences of the calcium binding loops are provided as a reference. The subscript indices stand for the positions of the starting/ending residues in the CaM sequence. Underscored residues are the selected amino acids with negative charges as shown in the third column. I adjusted the partial charges on the sidechain beads of those residues based on the bipyramidal pentagonal coordination geometry of the calcium ion. [14]

Calcium loop	Amino-acid sequence	Adjusted partial charges on selected amino acids
Loop 1	$_{20}$<u>D</u>K<u>D</u>G<u>D</u>GTITTK<u>E</u>$_{31}$	-0.5, -0.5, -0.5, -0.5
Loop 2	$_{56}$<u>D</u>A<u>D</u>GNGTI<u>D</u>FP<u>E</u>$_{67}$	-0.5, -0.5, -0.5, -0.5
Loop 3	$_{93}$<u>D</u>K<u>D</u>GNGYISAA<u>E</u>$_{104}$	-0.33, -0.33, -0.33
Loop 4	$_{129}$<u>D</u>I<u>D</u>G<u>D</u>GQVNYE<u>E</u>$_{140}$	-0.5, -0.5, -0.5, -0.5

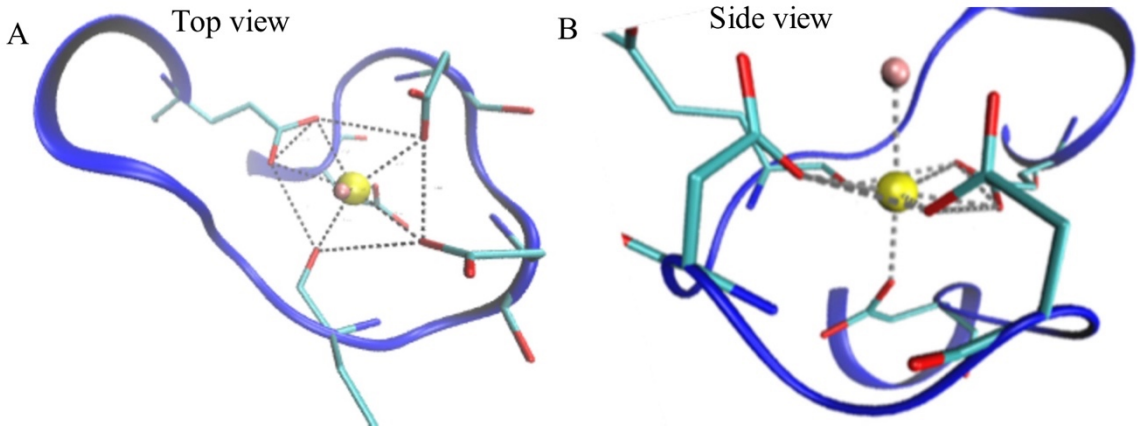

Figure 2.3. Illustration of a Ca^{2+} binding loop at the fourth EF hand of CaM (PDB ID: 1CLL). The pentagonal bipyramidal geometry of the Ca^{2+} coordination is represented with the dotted lines in top view (A) and side view (B). The red sticks linked with the dotted lines represent oxygen atoms that participate in coordination. The water molecule, shown in pink, completes the spherical coordination through the hydrogen bonding. Adapted from. [14]

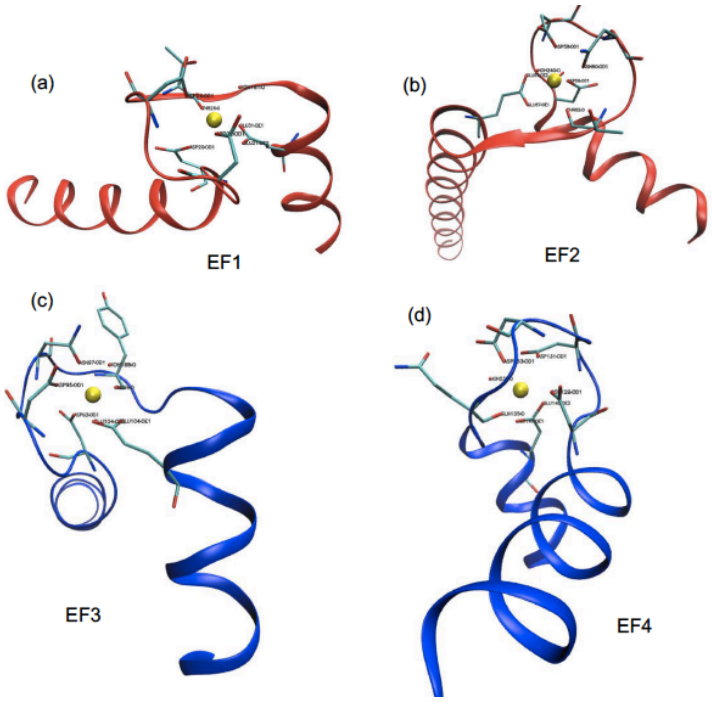

Figure 2.4. The four EF hands of CaM (PDB ID: 1CLL). Calcium ion (yellow bead) is coordinated by specific residues in the geometry. EF1 (a) and EF2 (b) are found in the N-domain of CaM while EF3 (c) and EF (d) are found in the C-domain. Adapted from. [14]

In this study, we compared the abovementioned approach (approach III) with two other calcium models:

I. Splitting the Ca^{2+} charges evenly to the neighboring negatively charged residues [97];
II. Splitting the Ca^{2+} charges evenly to the residues that participate with the calcium coordination and the radius of gyration of the sidechain beads are constrained by a stiff harmonic potential with a force constant of 30 kcal/mol/Å^2;
III. Same with approach II except that the harmonic constraint is removed.

Comparing to approach I of Ca^{2+} in AWSEM [97], in approach III calcium charges were split according to the coordination chemistry of calcium; comparing to approach II, more degrees of freedom to the Ca^{2+} coordination residues allow sampling of more different conformations of CaM.

2.4 SIMULATION DETAILS

We performed the coarse-grained simulations with the open-source MD package LAMMPS, in which AWSEM code were implemented [43]. We used the periodic boundary conditions on the cubic box of 400 Å on each side so that even unfolded CaM can fit in the box. Initial velocities were chosen randomly from a Boltzmann distribution with the average squared velocity equal to $3k_BT/m$, where k_B is the Boltzmann constant, m the mass, and the temperature T. We also used the cutoff distance of 3.5 Å and κ (inverse of the Debye length) = 0.0127 Å^{-1} in the electrostatic term.

Simulated annealing: firstly we denatured the protein at 450 K using the initial structure built from the crystal structure (PDB ID: 1CLL) [63]; then we performed simulated annealing on Ca^{2+}-

CaM by cooling down the system to a low temperature (280 K) for 2,000,000 time steps for each value of λ_{FM} (models I.1-3 in Table 2.2).

Production simulation: We employed umbrella sampling (US) method [101] to evaluate the thermodynamic properties of Ca^{2+}-CaM using different memory parameters. We conducted the production simulations in canonical ensemble (NVT) at the constant temperature of 300 K. R_g of Ca^{2+}-CaM was used as the reaction coordinate, which was restrained by a harmonic potential,

$$E_{Restr}^{i} = \frac{1}{2}\mathrm{k}\left(\mathrm{Rg^{i}} - \mathrm{Rg_0^{i}}\right)^2 \tag{2.7}$$

with a force constant k = 50 kcal / Å^2. The equilibrium positions of the harmonic potential Rg_0 range from 13.00 Å to 26.00 Å with a bin size of 0.25 Å, making up a total of 53 windows (i = 0, 1, 2, …, 52). The setup of the force constant and the equilibrium positions of the windows is justified by the fluctuations of R_g around the corresponding equilibrium positions as well as the overlap in the R_g distribution between neighboring windows, which are provided in Figure 2.5, Figure 2.6, and Figure 2.7, respectively. For each window, 2,000,000 time-steps of constrained molecular simulations were conducted using different initial conditions. To generate the initial structures at each window for the US simulations, molecular dynamics simulations were carried out for the Ca^{2+}-CaM from the crystal structure of CaM (PDB ID: 1CLL, in the coarse-grained model) at a high temperature T = 300 K. Structures with R_g closest to Rg_0^i were selected as the initial configurations. A set of 5 US simulations were carried out at each window with the same initial configuration and different random initial velocities. The integration time step is 2 fs. It is important to note that with the smoothed CG potential, the realistic time represented by a time step is longer than 2 fs. Coordinates and Rg were recorded every 1,000-time steps for analyses, making

up 10,000 frames of data for analysis at each window. We analyzed the data using multi-state Bennett acceptance ratio (MBAR) estimator with the *pymbar* program [102].

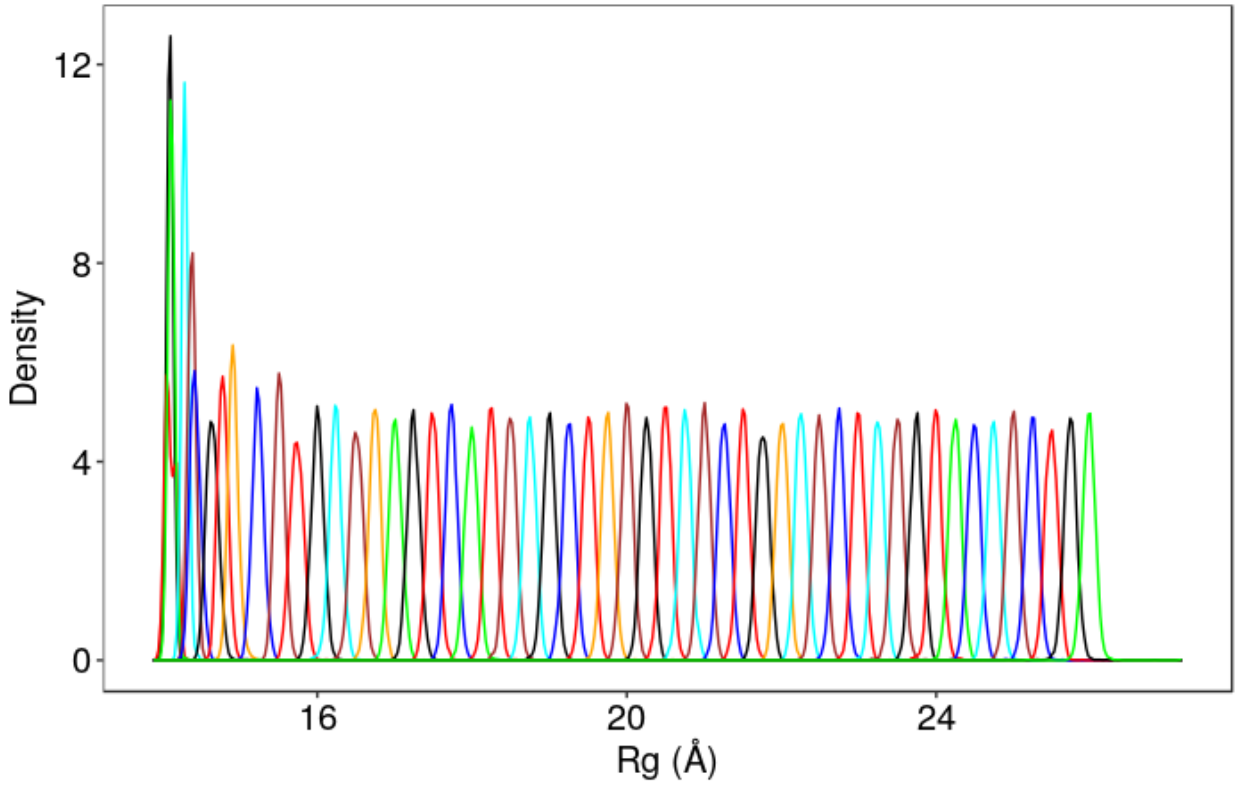

Figure 2.5. Radius of gyration (Rg) density plots of the 53 consecutive windows. Rg was varied from 13 to 26 Å in increment of 0.25. Each color represents the density plot of a single window. One can see that the plots overlap with each other. Simulations were performed for 2,000,000-time steps using λ_{FM} = 0.1, W_n^{m} =1. I calculated R_g at different biasing Rg_0 with respect to the extended crystal structure of Ca^{2+}-CaM (PDB ID 1CLL) for the C_α atoms of the backbone residues. Adapted from. [14]

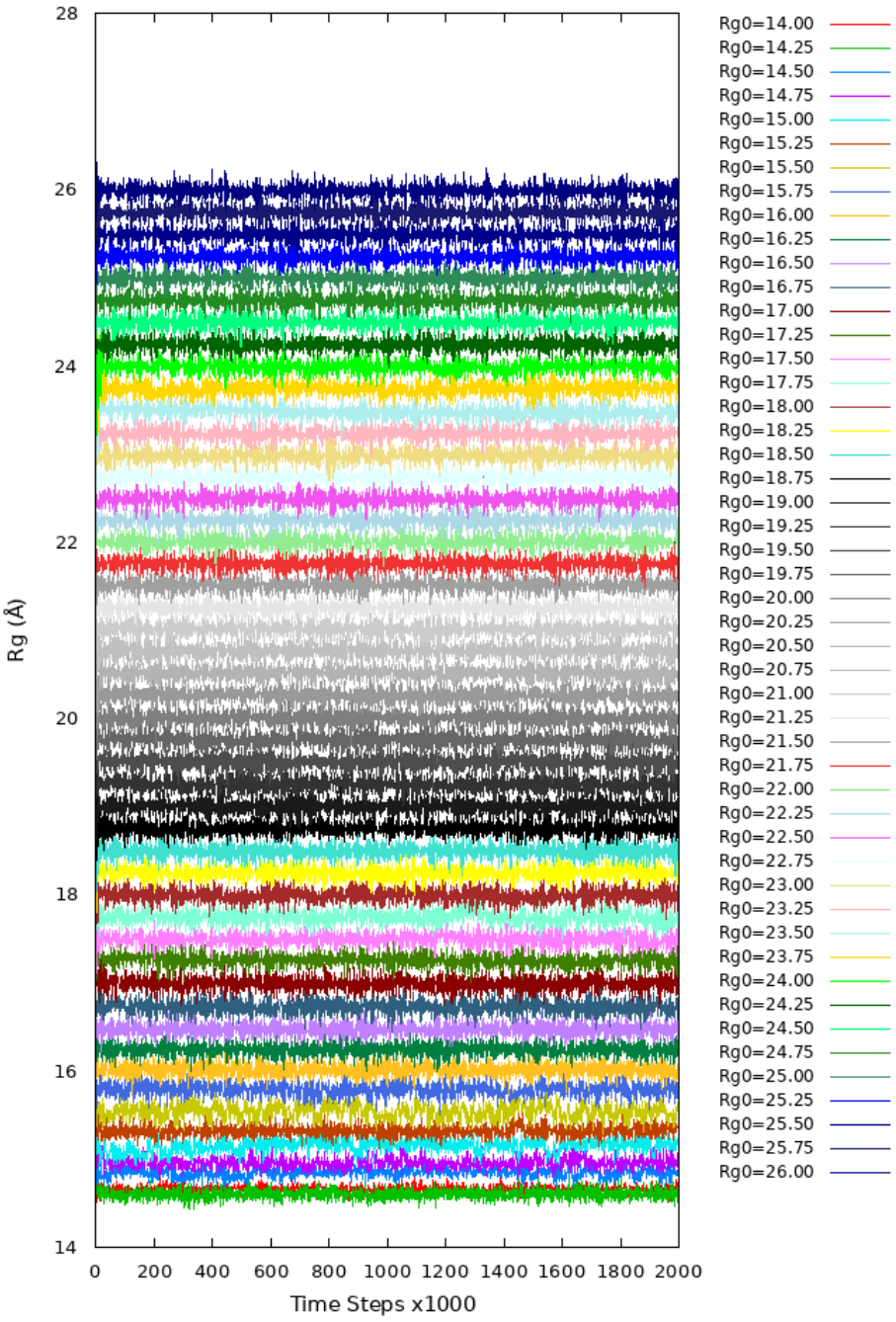

Figure 2.6. Radius of gyration (R_g) variation as a function of the simulation time. Rg was varied from 14 to 26 Å in increment of 0.25. Each color represents the time variation of the Rg of a single window with the specified biasing Rg_0. Simulations were carried out for 2,000,000-time steps using $\lambda_{FM} = 0.1$, $W_n^{\text{others}} = 1$, $W_n^{\text{1CLL}} = 5$. I calculated the R_g for each of the biasing Rg_0 with respect to the extended crystal structure of Ca^{2+}-CaM (PDB ID 1CLL) for the C_α atoms of the backbone residues. The different plots show that Rg fluctuates around the biasing Rg_0. [14]

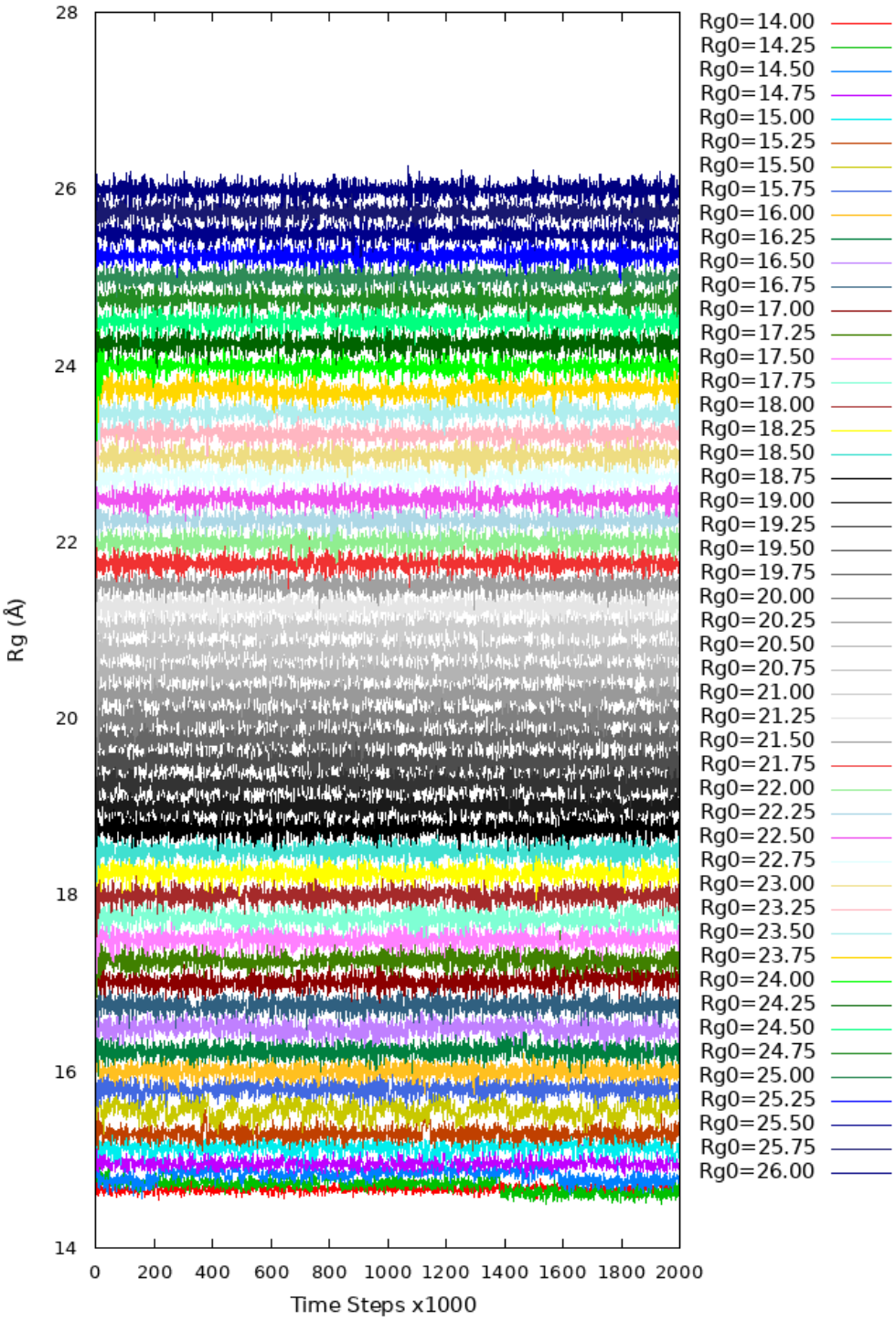

Figure 2.7. Radius of gyration (R_g) variation as a function of the simulation time. Rg was varied from 14 to 26 Å in increment of 0.25. Each color represents the time variation of the Rg of

a single window. Simulations were performed for 2,000,000-time steps using λ_{FM} = 0.1, $W_2^{1\mathrm{CLL}}$ = 8. I calculated R_g with respect to the extended crystal structure of Ca^{2+}-CaM (PDB ID 1CLL) for the C_α atoms of the backbone residues. The different plots show that Rg fluctuates around the biasing Rg_0. [14]

2.5 DATA ANALYSIS

We described the conformational changes of the Ca^{2+}-CaM during the simulations by calculating the radius of gyration (R_g) and the pairwise comparison (Q_w) from the crystal structure (PDB ID: 1CLL).

2.5.1 *Radius of gyration ($\boldsymbol{R_g}$)*

R_g describes the compactness of the whole system and its definition is given in Equation 2.8,

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i \, \Delta r_i^2}{\sum_{i=1}^{N} m_i}} \tag{2.8}$$

In the case of PDB structures, m_i is the mass of the i^{th} atom, Δr_i is the distance between the i^{th} atom and the center of mass of the system. In the case of coarse-grained structures, only the C_α beads are used, m_i is the mass of the C_α bead of the i^{th} residue, Δr_i is the distance between the C_α bead of the i^{th} residue and the center of mass of the system.

2.5.2 *Pairwise comparison*

Q_w measures the degree of similarity of conformations from the trajectories with respect to the native structure through the pairwise comparison. In other words, it compares the pairwise

distances of C_α atoms (beads) among the residues in each instantaneous structure to counterpart in the native structure. This value is normalized to 1, with a higher value corresponding to the greater similarity to the native structure. In this case, Q_w for the extended crystal structure of Ca^{2+}-CaM (PDB ID: 1CLL), Q_w = 1. The expression of the order parameter Q_w is given in Equation 2.9,

$$Q_w = \frac{2}{(N-2)(N-3)} \sum_{|i-j|\geq 2} exp[-\frac{(r_{ij} - r_{ij}^m)^2}{2\sigma_{ij}^2}] \quad (2.9)$$

where N is the total number of residues, r_{ij} represents the instantaneous distance between C_α atoms (beads) of residues *i* and *j*, r_{ij}^m is the same distance in the reference structure (PDB ID: 1CLL) and σ_{ij} is obtained from the following relation $\sigma_{ij}=(1+|i-j|)^{0.15}$.

2.5.3 *Definition of collapsed and extended conformations of CaM*

In the analyses, we used R_g to define a collapse or extended conformation. The R_g cutoff was set according to the potential of mean force (PMF) and distribution of R_g in the simulations (see the sessions below). A collapsed conformation is defined if R_g < 18.5Å; elsewise an extended conformation is defined.

2.6 MODELING THE MULTIPLE CONFORMATIONS OF CALMODULIN

The goal is to create a computational model for CaM accommodating its several key functions including divalent ion binding, conformational dynamics, as well as target recognition. In order to

achieve this, we adopted a pool of 60 CaM or CaM complex structures from our previous study [98] representing full-length CaM without mutations and its complex with 24 unique target proteins/peptides. As stated in the previous section, our approach is to tune the two scaling parameters λ_{FM} and W_n^m to adjust the global and each segment of the multiple memories, respectively, relative to other terms in the total energy function. After optimizing the memory parameters, three models of calcium are compared to describe the divalent ion binding.

Each of the 60 memories has a specific conformation of calmodulin which could be helpful in target recognition and target selection, therefore, clustering the 60 memories may leave out important conformations that accommodate recognition of a specific target protein. Potentially there could be as many as 181 fragment memory parameters to be determined in the most complicated scenario. However, in general practice of AWSEM modelling, a search in the available determined structures in the database according to the match in the desired amino-acid sequence is conducted to generate a library of memory structures. Therefore, to avoid excessive tuning of the parameters, the individual memory weight W_n^m are kept as a constant default value. Because the unbound extended form of the Ca^{2+}/CaM (PDB code: 1CLL) was found to be most dominant experimentally, we adjusted its local memory weight parameter correspondingly, while keeping others the default $W_n^{others} = 1$. The actual parameters we tuned eventually reduces to λ_{FM}, and W_n^{1CLL} (n = 1, 2, 3 to represent N-domain, central linker, and C-domain of the protein).

Parameterization of our CG model was guided by the existing following experimental measures: R_g of Ca^{2+}-CaM spans from 20.0 to 22.5 Å measured by small angle X-ray scattering (SAXS) experiments [103-105]; in addition, the percentage of CaM's collapsed population is about 10% measured by NMR experiments [72] (i.e. the ratio between the collapsed population and the extended population is $\gamma_{c:e} \sim 0.11$).

2.7 THE GLOBAL PARAMETER WEIGHT CONTROLS THE ENERGY BARRIER BETWEEN EXTENDED AND THE COLLAPSED STATES OF Ca^{2+}-CaM

Firstly, we determined the magnitude of the fragment memory interaction by investigating its effect on the folding properties of Ca^{2+}-CaM. We carried out simulated annealing using the unfolded structure of the Ca^{2+}-CaM (Q_w ~ 0.2 as shown in Figure 2.8) with three orders of magnitudes of the global scaling parameter λ_{FM} = 1, 0.1, 0.01. At $\lambda_{FM} = 1$, the system transitions abruptly from the unfolded to the folded state (Figure 2.9 (A, B)); at $\lambda_{FM} = 0.1$, smooth transitions of Ca^{2+}-CaM from the unfolded to the folded structure were observed with Q_w reaching a maximum value of 0.67 (Figure 2.10 (C, D)); at $\lambda_{FM} = 0.01$, the maximum value of Q_w was less than 0.55 and the total energy of the system did not stabilize on a long timescale (Figure 2.11 (E, F)).

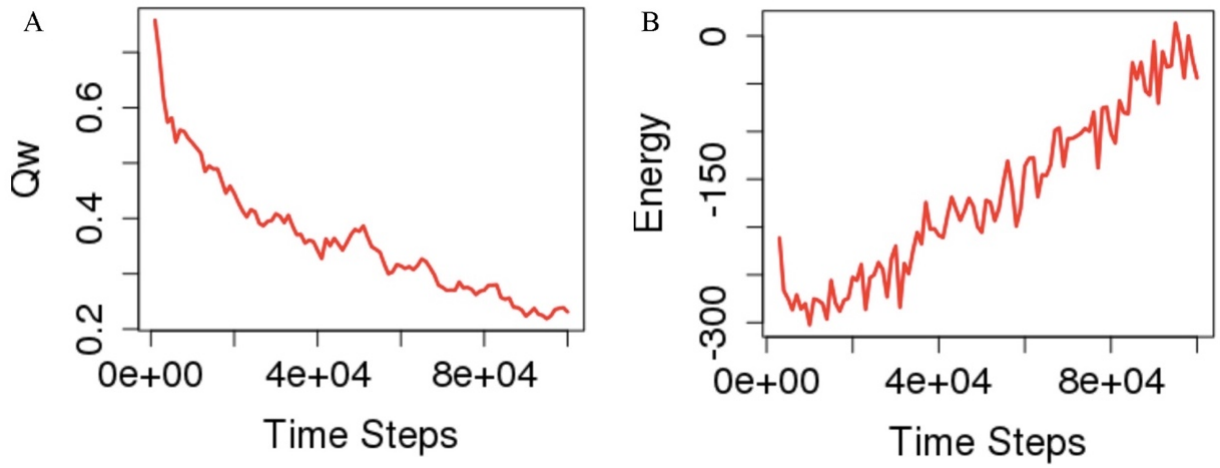

Figure 2.8. Time variation of the order parameter Q_w (A) and the associated energy (B) of the denatured CaM before the simulated annealing process. [14]

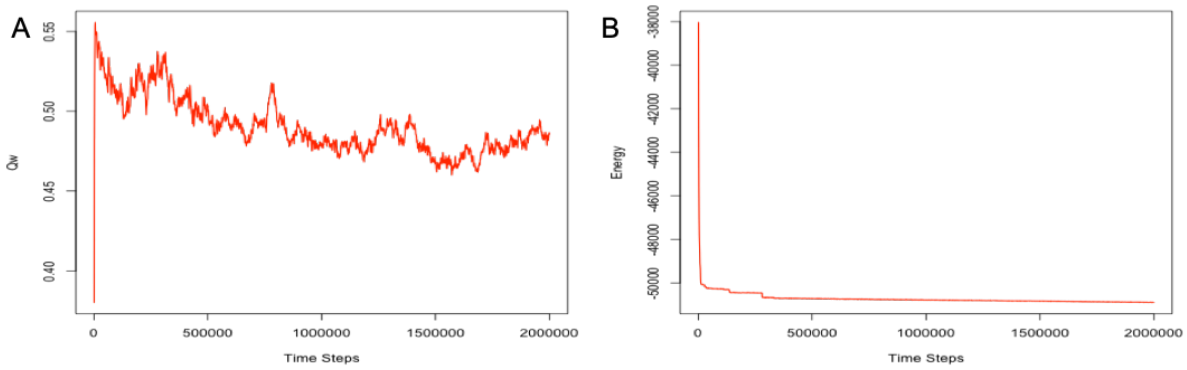

Figure 2.9. Time variation of the order parameter Q_w and the energy obtained from the simulated annealing of CaM using $\lambda_{FM} = 1$.

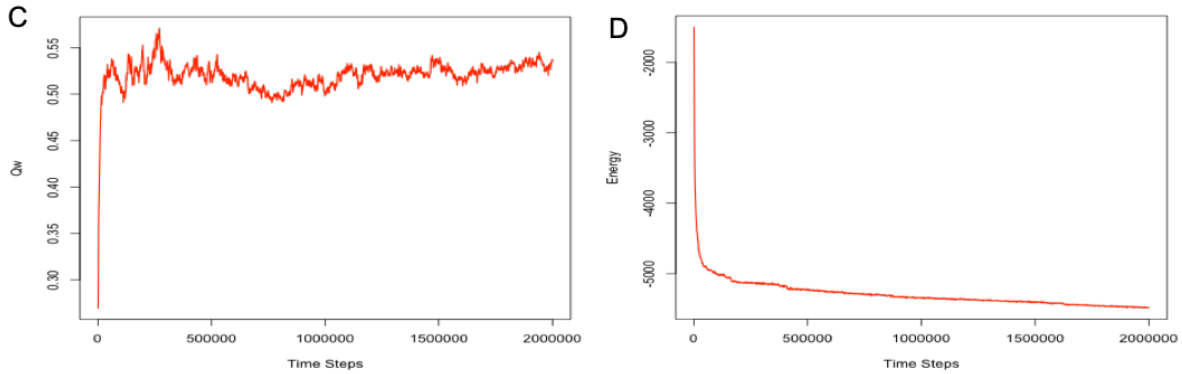

Figure 2.10. Time variation of the order parameter Q_w and the energy obtained from the simulated annealing of CaM using $\lambda_{FM} = 0.1$.

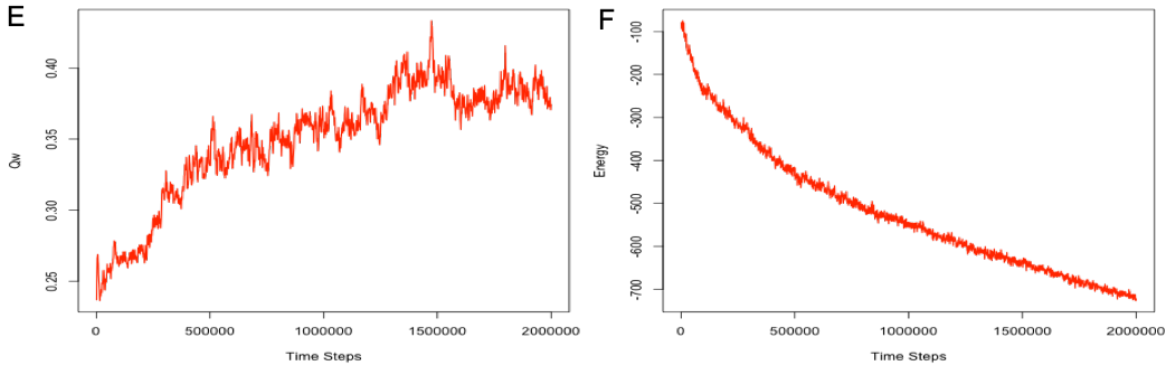

Figure 2.11. Time variation of the order parameter Q_w and the energy obtained from the simulated annealing of CaM using $\lambda_{FM} = 0.01$. [14]

In addition, we carried out US simulations to evaluate thermodynamic properties of Ca^{2+}-CaM with the global scaling parameter in three orders of magnitudes λ_{FM} = 1, 0.1, and 0.01, while keeping the local memory scaling parameters W_n^m = 1. The PMF and the probability distribution as a function of R_g were computed using the MBAR estimator (Figure 2.12 and Figure 2.13) and the average R_g ($\overline{R_g}$) as well as the ratio between the population of collapsed conformation and that of the extended conformation ($\gamma_{c:e}$) were evaluated from the PMF. For $\lambda_{FM} = 1$, Ca^{2+}-CaM mostly samples the extended conformations (with R_g ~ 21.5 to 22.5 Å) and the free energy barrier between the extended and collapsed conformations is quite large (~6 k_BT), therefore, the system lacks the capacity of shifting between multiple conformations (Figure 2.14). For $\lambda_{FM} = 0.1$, there are four minima in the PMF of Ca^{2+}-CaM with R_g ranging from 15 to 22 Å (Figure 2.15), and the free energy barrier between the collapsed states and the extended state is ~ k_BT; the $\overline{R_g}$ of Ca^{2+}-CaM is within the experimental range ($\overline{R_g}$ = 20.51 Å), however, $\gamma_{c:e}$ (~ 0.48) is much higher than the experimental one (~ 0.11). It is worthwhile to note that the minimum at R_g ~ 15 Å resembles a completely collapsed structure as found in our previous study [106]. For $\lambda_{FM} = 0.01$, there are noticeably two minima located at R_g ~ 17 Å and R_g ~ 20.5 Å (Figure 2.16), sampling a somewhat collapsed conformation and the extended conformation respectively with free energy barrier < k_BT; $\overline{R_g}$ = 19.86 Å, below the experimental measure range and $g_{c:e}$ (~ 0.64) is much higher than the experimental one.

In summary, a large value of the scaling parameter λ_{FM} might tend to limit the malleability of Ca^{2+}-CaM in favor of the extended conformation (initial configuration of the simulations). Decreasing λ_{FM} leads to an increase of $\gamma_{c:e}$ and a decrease in $\overline{R_g}$. This is because decreasing the weight of the memory potential favors the hydrophobic interactions among residues of the two

domains of Ca^{2+}-CaM and reduces the free energy barrier between the two major conformations of Ca^{2+}-CaM (extended and collapsed forms); consequently, the system becomes more flexible and can shift easily between the extended and collapsed conformations.

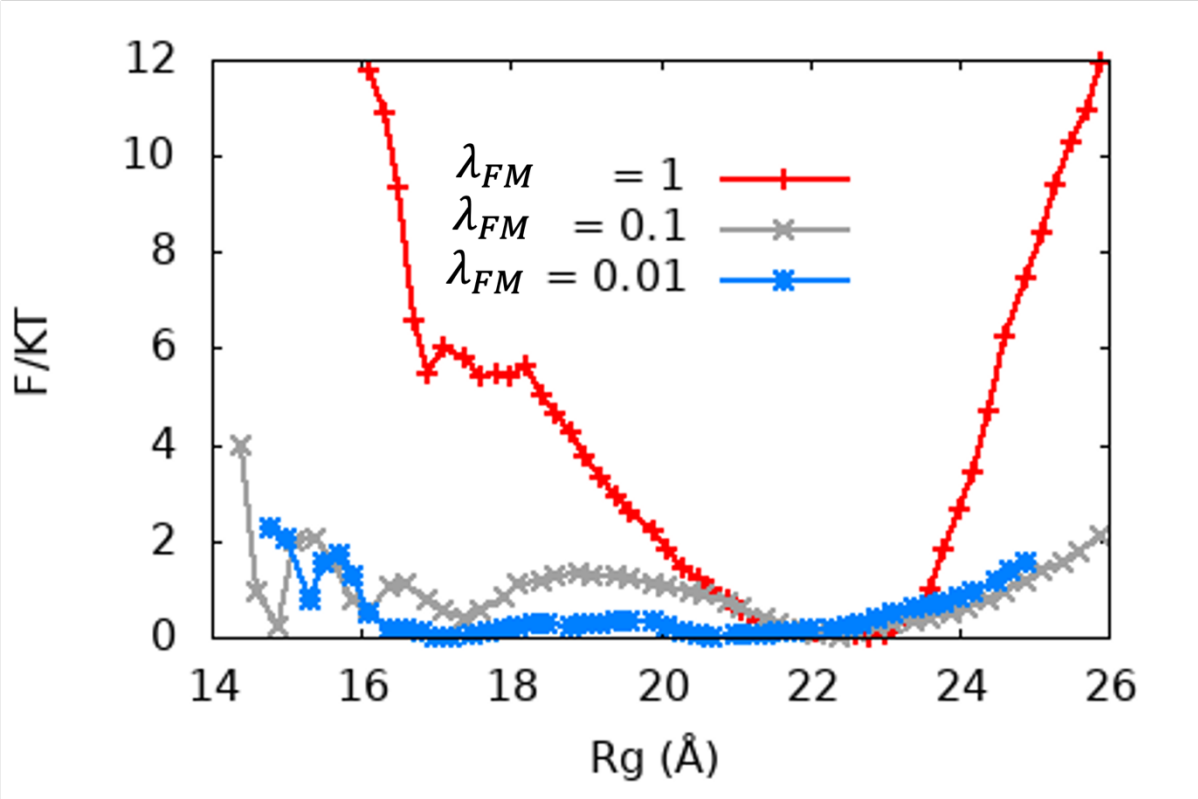

Figure 2.12. PMF of the Ca^{2+}-CaM models using the global memory weight of different order of magnitudes. The memory weight parameters of the Ca^{2+}-CaM models are listed in Table 2 (Models 1-3; λ_{FM} = 1, 0.1, 0.01, and W_n^{m} = 1). These different parameters were used to get the reweighted free energy profile as a function of R_g from US simulations with λ_{FM} = 1, 0.1, 0.01, respectively. [14]

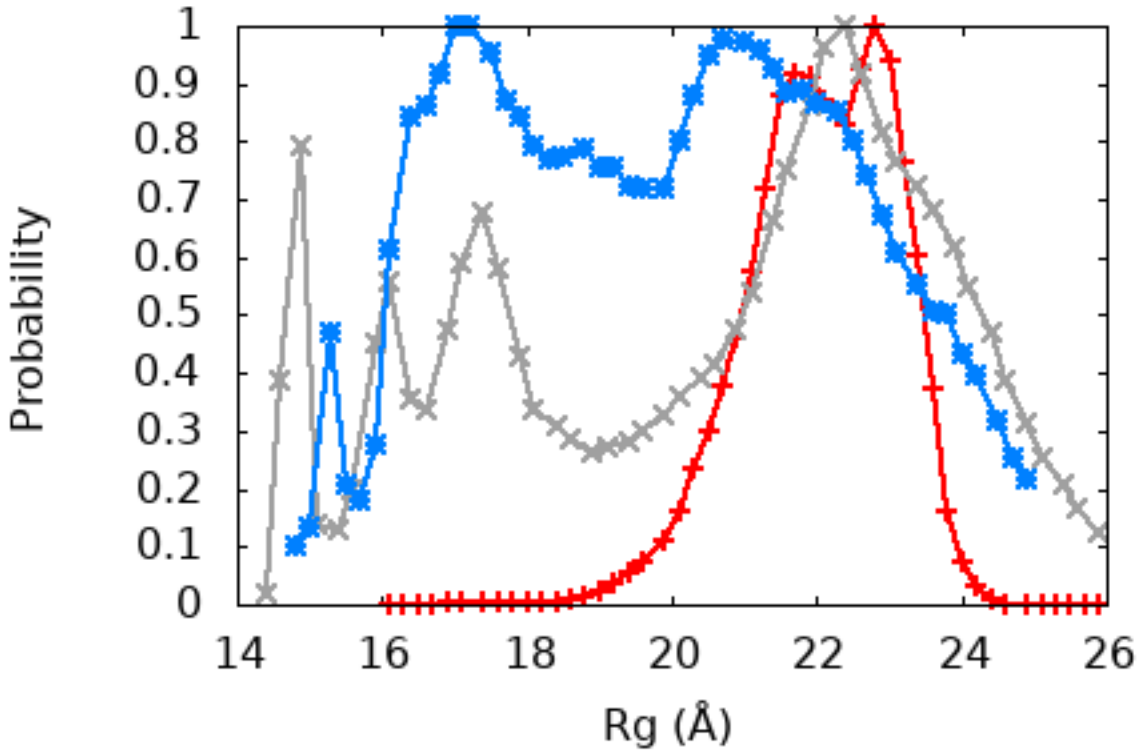

Figure 2.13. R_g distribution of the Ca^{2+}-CaM models using the global memory weight of different order of magnitudes. The memory weight parameters of the Ca^{2+}-CaM models are listed in Table 2 (Models 1-3; λ_{FM} = 1, 0.1, 0.01, and W_n^{m} = 1). These parameters were used to get the probability distribution as a function of R_g with λ_{FM} = 1, 0.1, 0.01, respectively. [14]

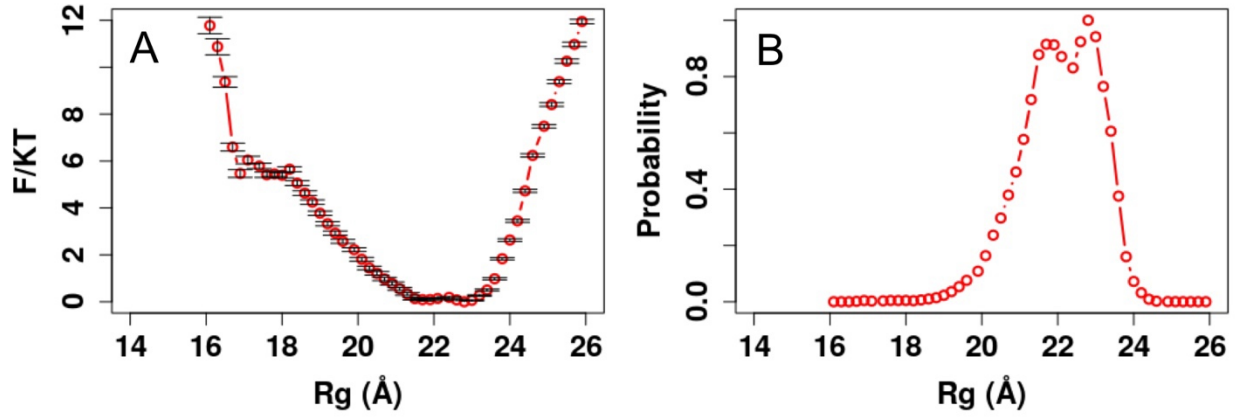

Figure 2.14. Standard errors of the PMF and R_g distribution of the Ca^{2+}-CaM models using the global memory weight λ_{FM} = 1 and W_n^{m} = 1. [14]

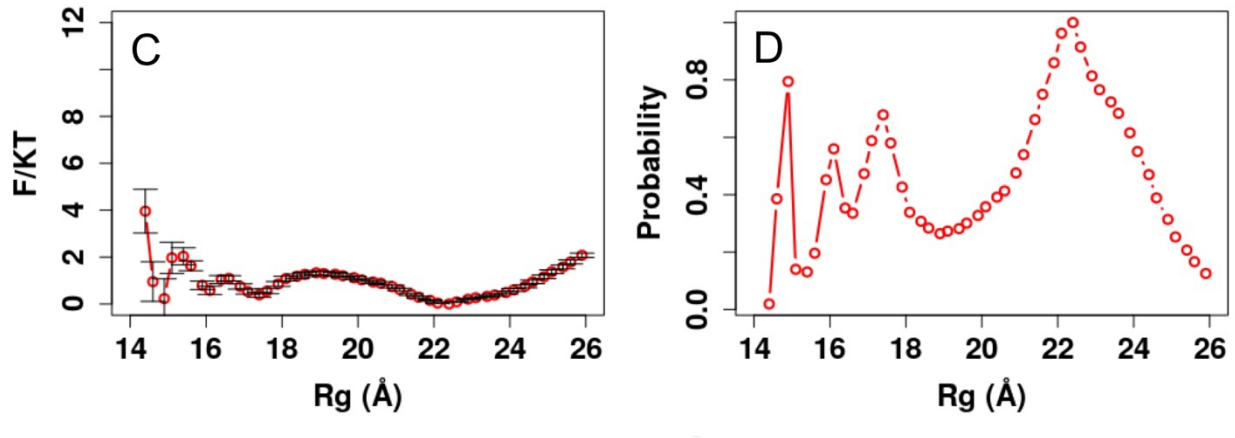

Figure 2.15. Standard errors of the PMF and R_g distribution of the Ca^{2+}-CaM models using the global memory weight λ_{FM} = 0.1 and W_n^{m} = 1. [14]

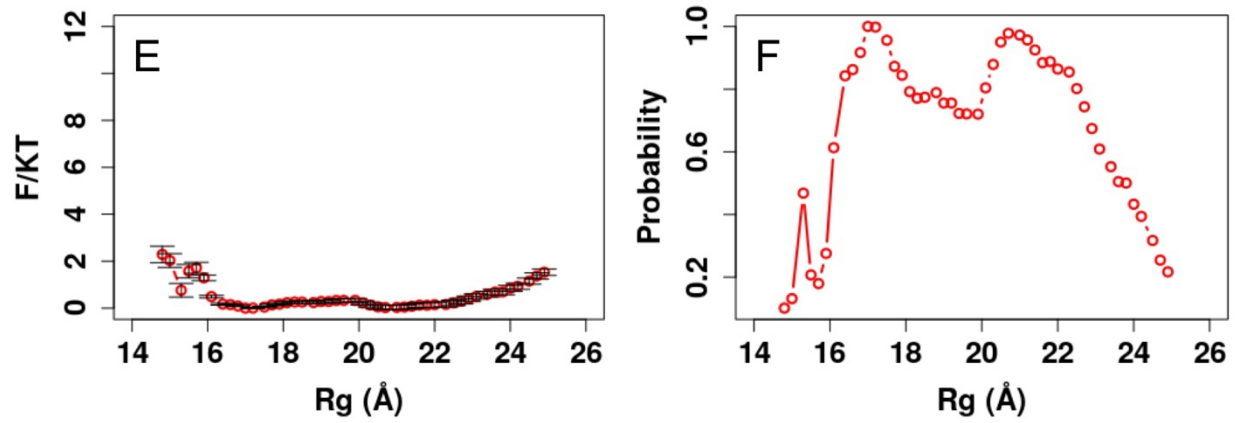

Figure 2.16. Standard errors of the PMF and R_g distribution of the Ca^{2+}-CaM models using the global memory weight λ_{FM} = 0.01 and W_n^{m} = 1. [14]

2.8 SINGLE MEMORY OPTIMIZATION CAPTURES THE DYNAMICS OF CA^{2+}-CAM.

Although the global fragment memory weight of $\lambda_{FM} = 0.1$ resulted in a system that can sample the major two conformations of the Ca^{2+}-CaM, and $\overline{R_g}$ is within the experimental range, the ratio

between the collapsed and the extended populations $\gamma_{c:e}$ did not match the experimentally measured ones. Therefore, finer tuning of the memory parameters is required.

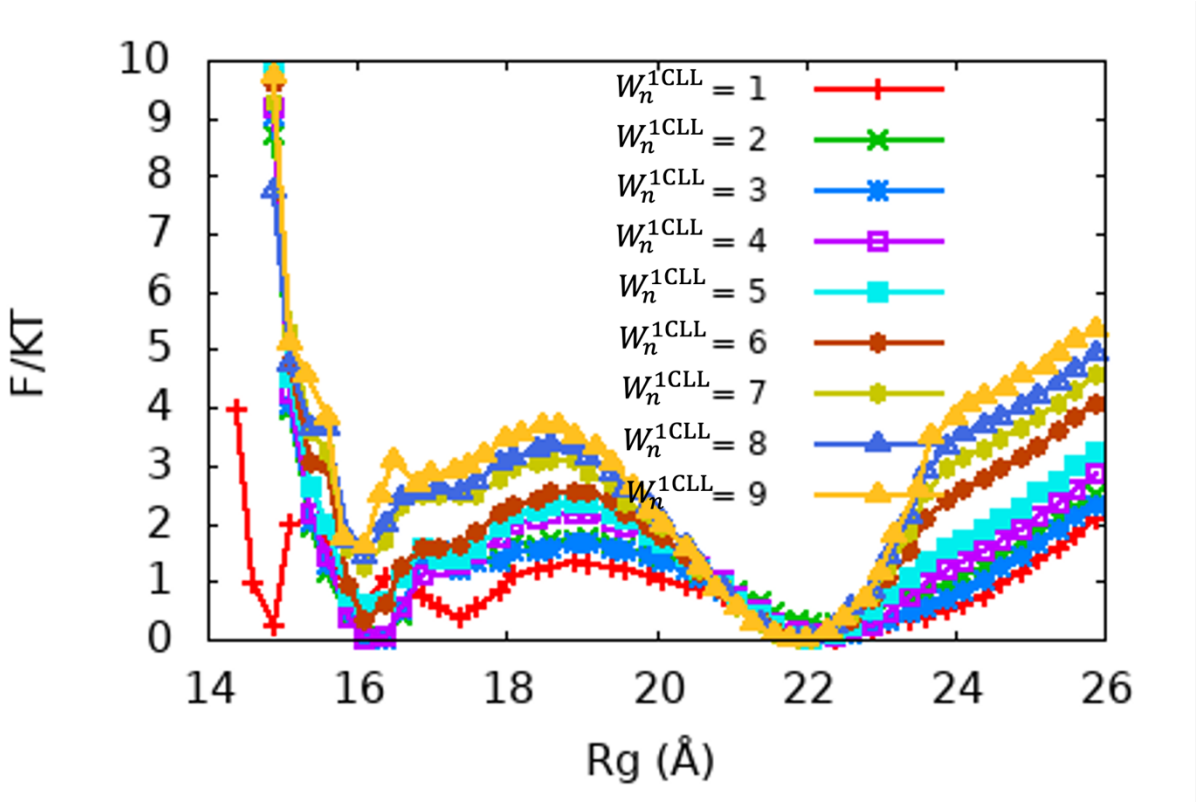

Figure 2.17. PMF of the Ca^{2+}-CaM models for different values of the local memory weight W_n^{1CLL} (n = 1, 2, 3). The memory weight parameters of the Ca^{2+}-CaM models are listed in Table 2 (Model II; $\lambda_{FM} = 0.1$, $W_n^{others} = 1$, and $W_n^{1CLL} = 1 - 9$) were used to get the reweighted free energy profile as a function of R_g from US simulations. [14]

Since the crystal structure of Ca^{2+}-CaM (PDB ID: 1CLL) demonstrates a value of R_g that is closest to the experimentally measured values and does not bind a target, we first examined whether variation in single memory weight of this structure (W_n^{1CLL}) could give rise to the conformational changes of the Ca^{2+}-CaM. We performed a series of molecular simulations using US with W_n^{1CLL} = 1, 2, …, 9 and maintaining $\lambda_{FM} = 0.1$ and $W_n^{others} = 1$ (Table 2.2 Model II). From the PMF

profiles and the probability distributions shown in Figure 2.17 and Figure 2.18, we show that the Ca^{2+}-CaM samples both the extended and collapsed conformations in all the models we explored. In general, with the increase of W_n^{1CLL}, stability of the collapsed state decreases and the barrier between the extended and collapsed states increases, hence the increase in $\overline{R_g}$ and decrease in $\gamma_{c:e}$. Specifically, without tuning any other local memory weights, $W_n^{1CLL} = 8, 9$ matches with experimental values of $\overline{R_g}$ and $\gamma_{c:e}$ quite well (Table 2.2).

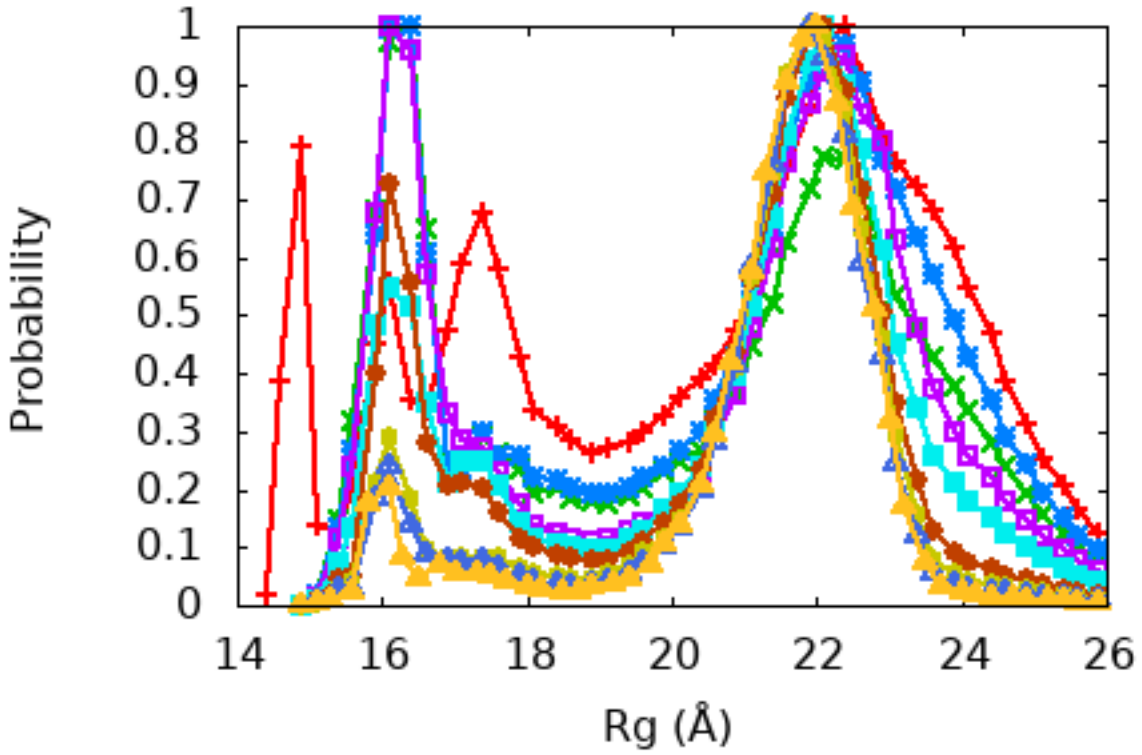

Figure 2.18. R_g distribution of the Ca^{2+}-CaM models for different values of the local memory weight W_n^{1CLL} (n = 1, 2, 3). The memory weight parameters of the Ca^{2+}-CaM models are listed in Table 2 (Model II; $\lambda_{FM} = 0.1$, $W_n^{others} = 1$, and $W_n^{1CLL} = 1 - 9$) were used to get the probability distribution as a function of R_g. [14]

2.9 CENTRAL LINKER OF THE Ca^{2+}-CAM DETERMINES THE CONFORMATIONAL CHANGE OF Ca^{2+}-CAM.

Although the X-ray crystallography experiment shows that the central linker that connects the two domains of Ca^{2+}-CaM has a helical structure, numerous studies including NMR have shown its high flexibility in solution. To explore the relationship between the flexible linker and the dynamics of Ca^{2+}-CaM in more details, we carried out a series of US molecular simulations with the memory weight of the central linker of the Ca^{2+}-CaM, $W_2^{1\mathrm{CLL}} = 1,2,\ldots,9$ while keeping $\lambda_{FM} = 0.1$ and $W_n^{\mathrm{others}} = 1$, and $W_n^{1\mathrm{CLL}}$(n=1,3) = 5 (Table 2.2 Model III). As shown in the PMF profiles (Figure 2.19 and Figure 2.20), $W_2^{1\mathrm{CLL}}$ shows a similar effect on the thermodynamic properties of the Ca^{2+}-CaM; with the increase of $W_2^{1\mathrm{CLL}}$, stability of the collapsed state decreases and the barrier between the extended and collapsed states increases, hence the increase in $\overline{R_g}$ and decrease in $\gamma_{c:e}$ (Table 2.2). This is likely because the two globular domains of CaM consist of stable secondary structures (mostly α-helices) and the involving fragments are fully represented and heavily weighted by W_n^{others} (n = 1, 3) in other 59 memories, whereas the central linker connecting the two domains of CaM is modeled as an intrinsically disordered peptide; hence stability of the Ca^{2+}-CaM is more sensitive to the parameter $W_2^{1\mathrm{CLL}}$ corresponding to the central linker of CaM. Therefore, tuning $W_n^{1\mathrm{CLL}}$ for the whole CaM is mostly equivalent to tuning $W_2^{1\mathrm{CLL}}$ for the central linker. This can also be seen in the three highlighted models in Table 2.2, which reproduce the experimentally measure R_g and $\gamma_{c:e}$ well.

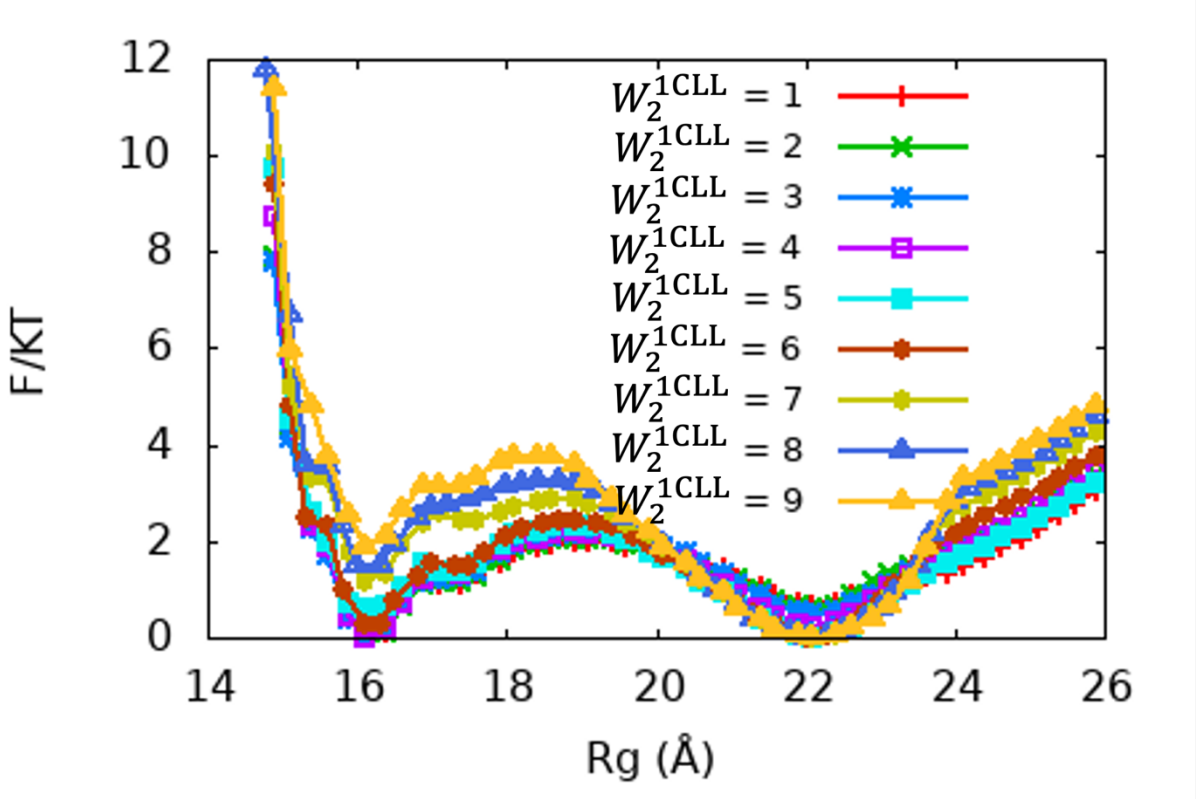

Figure 2.19. PMF of the Ca^{2+}-CaM models for different values of the local memory weight W_n^{1CLL} (n = 2) for the central linker of CaM. The memory weight parameters of the Ca^{2+}-CaM models are listed in Table 2 (Model III; $\lambda_{FM} = 0.1$, $W_n^{others} = 1$, $W_{1,3}^{1CLL} = 5$, $W_2^{1CLL} = 1 - 9$, which were used to get the reweighted free energy profile as a function of Rg from US simulations. [14]

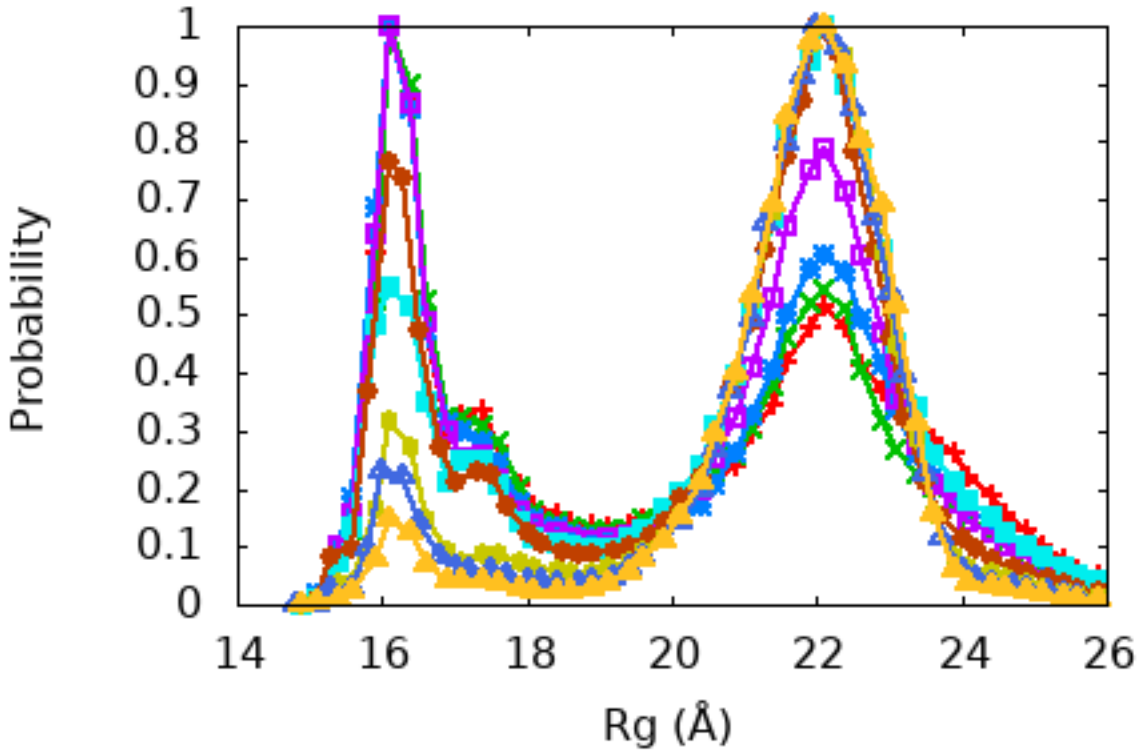

Figure 2.20. R_g distribution of the Ca^{2+}-CaM models for different values of the local memory weight W_n^{1CLL} (n = 2) for the central linker of CaM. The memory weight parameters of the Ca^{2+}-CaM models are listed in Table 2.2 (Model III; $\lambda_{FM} = 0.1$, $W_n^{others} = 1$, $W_{1,3}^{1CLL} = 5$, $W_2^{1CLL} = 1 - 9$, which were used to get the probability distribution as a function of R_g.

In order to illustrate the conformational changes of Ca^{2+}-CaM in solution, we selected representative snapshots from the simulations using the Ca^{2+}-CaM model with $\lambda_{FM} = 0.1$, $W_n^{others} = 1$, $W_{1,3}^{1CLL} = 5$, and $W_2^{1CLL} = 8$, as highlighted in Table 2.2. The unstructured central linker of the extended structure (Figure 2.21 (D)) unwraps and bends to allow the two domains of the Ca^{2+}-CaM to interact with each other, as shown in Figure 2.21 (A – C). Figure 2.21 (A) represents a completely collapsed structure, like the compact structure (PDB ID: 1PRW) or in CaM-target complexes, however, the chance for Ca^{2+}-CaM to sample this structure is low; Figure 2.21 (B)

represents a collapsed structure; and Figure 2.21 (C) represents a structure in the transition state between the collapsed and extended states.

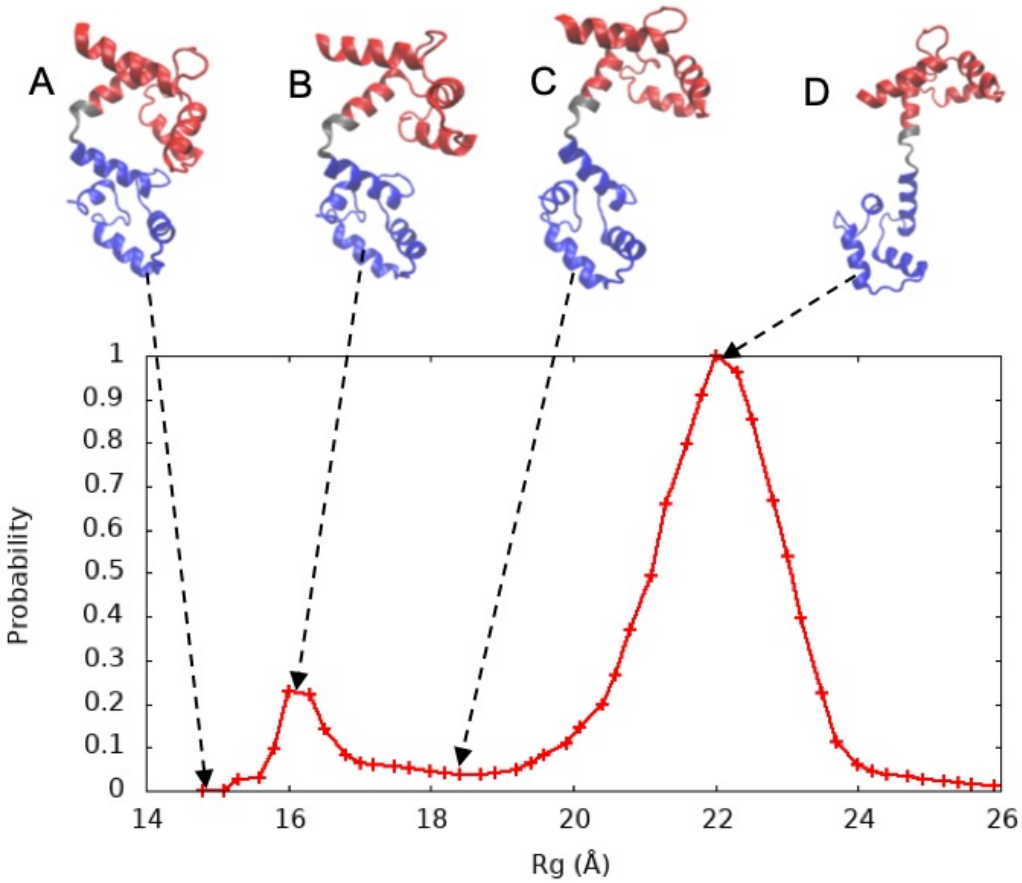

Figure 2.21. Representative snapshots of the Ca^{2+}-CaM model obtained from our simulations. The model that best reproduces experimental measurements were used ($\lambda_{FM} = 0.1$, $W_n^{\text{others}} = 1$, $W_{1,3}^{1\text{CLL}} = 5$, and $W_2^{1\text{CLL}} = 8$). (A) the completely collapsed conformation of the Ca^{2+}-CaM; (B) the collapsed conformation; (C) the intermediate conformation; (D) the extended conformation. These structures are represented along the probability distribution of R_g. [14]

2.10 COORDINATION GEOMETRY IS NECESSARY FOR REPRESENTING DIVALENT IONS IN COARSE-GRAINED MODELS

We further manifest the importance of coordination chemistry by comparing the three approaches to modelling Ca^{2+} as described in Method section 2.2 in Figure 2.22 and Figure 2.23. With approach I, the Ca^{2+} charges were evenly distributed to all the negatively charged residues [97] and Ca^{2+}-CaM presents equal probability of sampling the completely collapsed conformation and the extended conformation ($\gamma_{c:e}$ = 1.01). $\overline{R_g}$ ~ 18.64 Å is well below the experimental values ~ 21 Å. With approach II by applying constraints on the Ca^{2+} loops, Ca^{2+}-CaM samples favorably the completely collapsed conformation with Rg ~ 15 Å ($\gamma_{c:e}$ = 22.87). Consequently, the binding selectivity of Ca^{2+}-CaM could be impeded due to the lack of the flexibility of calcium binding loop in the system. Being confined in this completely collapsed conformation, Ca^{2+}-CaM lacks the capacity of recognizing, binding, and activating the targets. With approach III (w/ coordination chemistry), Ca^{2+}-CaM samples both extended and collapsed conformations while preferring the former.

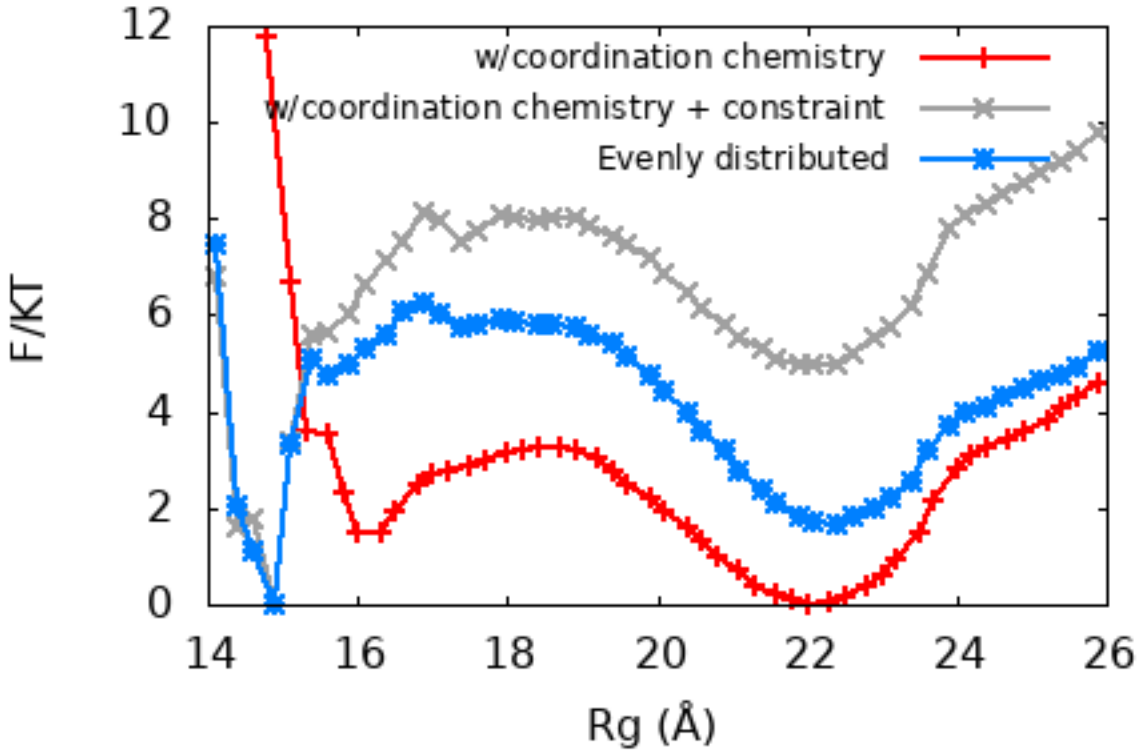

Figure 2.22. PMF of the Ca^{2+}-CaM for different Ca^{2+} models. The optimized memory parameters for the Ca^{2+}-CaM model were used ($\lambda_{FM} = 0.1$, $W_n^{\text{others}} = 1$, $W_{1,3}^{1\text{CLL}} = 5$, and $W_2^{1\text{CLL}} = 8$) to get the reweighted free energy profile as a function of R_g from the US simulations. The three charge models are described in the previous section. [14]

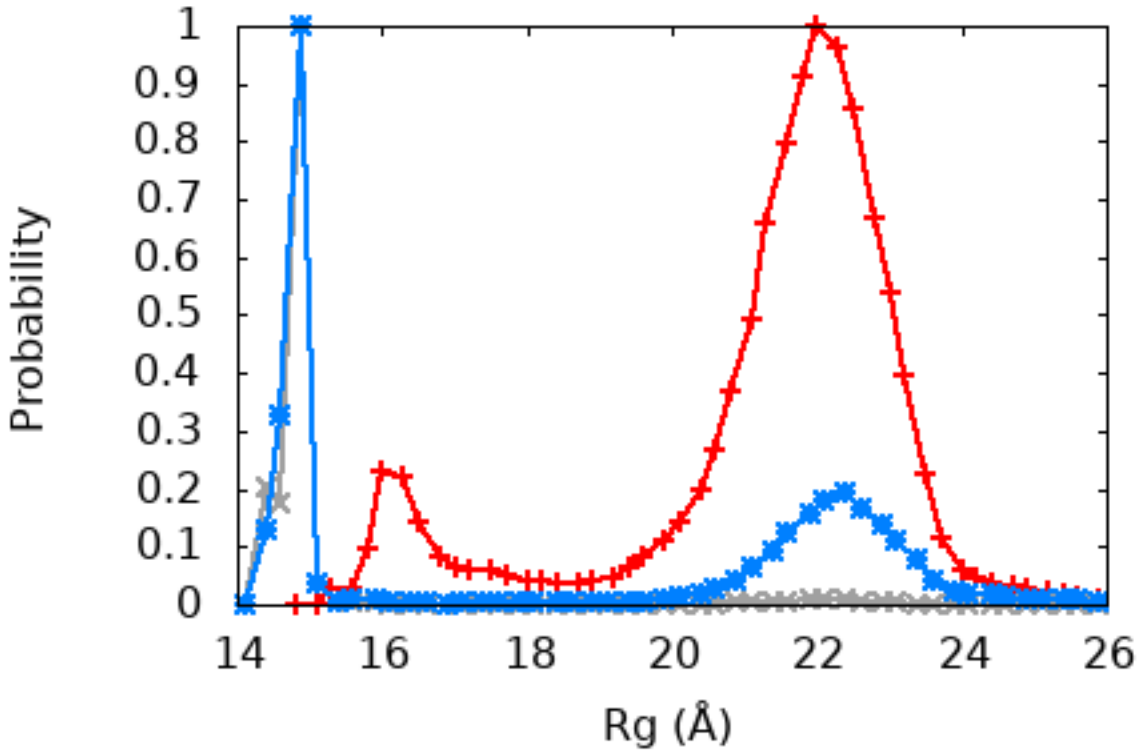

Figure 2.23. R_g distribution of the Ca^{2+}-CaM for different Ca^{2+} models. The optimized memory parameters for the Ca^{2+}-CaM model were used ($\lambda_{FM} = 0.1$, $W_n^{\text{others}} = 1$, $W_{1,3}^{1\text{CLL}} = 5$, and $W_2^{1\text{CLL}} = 8$) to get the probability distribution as a function of R_g. The three charge models are described in the previous section. [14]

Although it can be tempting to use the available intrinsically disordered protein (IDP) force fields such as AWSEM-IDP [96] (due to the disordered nature of the central linker) to study the dynamics of Ca^{2+}-CaM, one has to keep in mind the heterogeneity of CaM. The high rigidity of the two domains upon calcium binding limits the use of IDP force fields, thus making the computational investigation of CaM even more complex. However, force field development for calcium is mostly limited to Ca^{2+} ion in aqueous solution rather than in protein-bound environments. [107] One of the main challenges in computer simulations of macromolecules remains the development of force fields with the appropriate implementation of the metal ion. [46, 60, 108-111] Due to the high

sensibility of the molecules upon bivalent metal ion binding (example of the Ca^{2+}-induced conformational changes in CaM), a lack of appropriate approach to model its effects on the molecules could impede the capability of the target molecule to transduce its signals, thus, leading to the inactivation and/or mal-function of the signaling pathways. [95, 112]

Most calcium-binding proteins such as CaM contain the EF-hand (helix-loop-helix) structural domain as the calcium binding motif [113, 114]. In this motif, the calcium ion is coordinated in a pentagonal bipyramidal configuration (Figure 2.3). The canonical positions 1, 3, 5, 7, 9, and 12 represent the positions of the residues involved in calcium coordination [11, 60]. Table 2.3 shows a simple and efficient ways to consider the effect of the Ca^{2+} ions in the force field, where we distributed its charges to the negatively charged residues present in the pentagonal bipyramidal configuration. In contrast to other approaches where harmonic constraint is applied to the calcium coordinated residues in the loop, or the charges are distributed to all the negatively charged residues in the system, our approach (III) not only considers the heterogeneity of CaM, but also takes into account the coordination chemistry of calcium by giving the coordinated residues the freedom to participate to the different conformations of the system, which are relevant to the target selectivity of Ca^{2+}-CaM.

2.11 THE DISTRIBUTION OF CA^{2+}-CAM CONFORMATIONS IS BEYOND THE EXTENDED AND THE COLLAPSED STRUCTURES

Previous studies including NMR have shown that the high flexibility of the linker in solution allows the two domains to independently undergo multiple conformations [18, 20, 67, 68, 71, 86, 115]. Structural differences can be inferred from crystallography, but the dynamical insights are pivotal to understand Ca^{2+}-CaM interactions with the targets. Previous studies including NMR

spectroscopy, X-ray crystallography, small-angle X-ray scattering, molecular dynamics simulations unveil important details of the dynamics of Ca^{2+}-CaM, and paint a clear portrait of its conformational motions at the origin of the binding selectivity. The available structures of Ca^{2+}-CaM show that it crystallizes into two major conformations: extended [63] and collapsed [76]. However, solution structure studies including NMR have shown that Ca^{2+}-CaM adopts multiple conformations in solution [16, 18, 20, 64, 69, 116-119]. The conformational changes of CaM upon Ca^{2+} binding have been under intensive investigation [18, 20, 80, 119, 120]. The goal of such studies is to understand the molecular mechanism that leads to the Ca^{2+}-CaM target binding selectivity. How Ca^{2+}-CaM interacts with its CaMBTs is of major interest, since a good understanding of this complex mechanism involving rich conformations beyond the two major extended or the collapsed conformations could further our interpretation of how target binding may tune CaM's affinity for calcium ions [26, 112].

Ca^{2+}-CaM undergoes substantial conformational changes upon binding to a target; in most cases, the change in conformation is accompanied by the Ca^{2+}-CaM that adopts a collapsed form by wrapping around the target peptide with the two domains coming in proximity [81, 82, 121]. One can see that for the small molecules, Ca^{2+}-CaM will undergo less structural rearrangement during the complex formation. It is also interesting to notice that the collapsed conformation of the intact Ca^{2+}-CaM free target (e.g., Figure 2.21) assembles a vast variety of structures including structures that resemble the compact conformation of the Ca^{2+}-CaM/target peptide (Figure 2.1 (D)). This was also observed by X-ray crystallography [76], and NMR studies [72]. This collapsed structure of the Ca^{2+}-CaM (Figure 2.21 (A)) confirms the earlier experiments that show the existence of multiple conformations of the Ca^{2+}-CaM in solution, as well as the communication between opposing domains of CaM [16, 72, 122].

2.12 CONCLUSION

The present study was designed to numerically investigate the dynamics of Ca^{2+}-CaM. Due to the heterogeneity of CaM, we broke down CaM into three segments to capture both its flexibility at the linker and its physical properties that hold the EF hands upon calcium binding. Our coarse-grained simulations successfully reproduced the different motions of Ca^{2+}-CaM in solution and revealed the detailed pathways of its dynamical behavior. We have shown that Ca^{2+}-CaM adopts multiple conformations in solution, ranging between the extended and the collapsed conformations. The fluctuations in the calculated radius of gyration between that of the extended and the compact crystal structures of the Ca^{2+}-CaM provide evidence of the interdomain interactions. Furthermore, the average value of the radius of gyration, 21.35 Å, agrees with that previously reported from the SAXS measurements, 20.0 - 22.5 Å. The population ratio between the extended and the collapsed conformations is approximately 0.12, respectively, in accord with the previous experimental findings. The structural information presented in this study could be used to investigate the interactions between Ca^{2+}-CaM and the CaMBTs, such as CaMKII, which is the main protein involved in the memory formation.

Here we demonstrate a workflow of development of our AWSEM model for CaM guided by knowledge and experimental data. This model can sample various conformations of CaM and carries the memory of binding with various targets. The model would be useful for studying the interplay between target binding, calcium binding and conformational dynamics. Should new structures of CaM (or CaM/target complex) structures become available, the model can be easily expanded by including the structures in the memory library

Chapter 3. COMPUTAIONAL STUDY OF THE CALCIUM-CALMODULIN AND THE CALMODULIN DEPENDENT PROTEIN KINASE II PEPTIDE INTERACTIONS

Calmodulin (CaM) and calmodulin-dependent kinase II (CaMKII) interaction, modulated by the changes in intracellular calcium levels, is fundamental for various physiological processes, serving as a central mechanism in cellular signaling, with implications for neuronal function, synaptic plasticity. CaMKII, a serine/threonine kinase, exists in multiple isoforms and is prominently found in the central nervous system. It plays a crucial role in long-term potentiation (LTP), a cellular mechanism underlying learning and memory, by modulating the strength of synaptic connections. The role of CaMKII in synaptic plasticity and broader cellular processes underscores its significance in physiological and pathological conditions. The goal of this chapter is to use a combine experiment and simulations to investigate the reciprocal relationship between calcium, CaM, and CaMKII, which is of paramount importance to understand the above biological functions.

3.1 INTRODUCTION

Calcium/calmodulin (Ca^{2+}/CaM)-binding target $Ca^{2+/}$/CaM-dependent kinase II (CaMKII) is a multi-subunit (Figure 3.1) molecular complex that has interesting biological functions [15, 123]. Each subunit is composed of a kinase domain, a CaM-binding domain (also known as regulatory domain), an association domain, and a linker that connects the regulatory and the association

domains [124-126]. When activated by Ca^{2+}/CaM, it autophosphorylates (on a specific residue, T286/287), which increases its binding affinity for Ca^{2+}/CaM by several orders of magnitude [127, 128] – termed CaM-trapping [129, 130]. Studies using synthetic peptides that mimic its CaM-binding region (Figure 3.2 B) (CaMKIIp) successfully reproduce the low- and high-affinity of Ca^{2+}/CaM for the CaMKII holoenzyme [127, 129, 131-133]. Using a family of these peptides, along with the crystal structure of Ca^{2+}/CaM in complex with CaMKIIp (Figure 3.2 C), it is possible to identify sequence of residues within the peptide that produce the low and high affinity of Ca^{2+}/CaM for CaMKIIp [128, 134, 135]. More specifically, the minimum peptide sequence, CaMKIIp (296-309), leads to the low binding affinity of Ca^{2+}/CaM for CaMKIIp [134, 135], whereas CaMKIIp(290-309) mimics the high binding affinity of Ca^{2+}/CaM for CaMKIIp [136]. As a result, these peptide models are useful in defining specific residue sequences that show the molecular basis of CaM-trapping [129].

A substitution of residues 296-RRK-298 by alanine 296-AAA-298 in the bound Ca^{2+}/CaM/CaMKIIp complex with mutant peptide 296-AAA-298 induces a significant (up to 3000-fold) decrease in the Ca^{2+}/CaM's affinity for the AAA peptide relative to that of the Ca^{2+}/CaM/CaMKIIp (293-312) in the bound complex with the wild-type peptide [129]. To define the molecular basis for the altered binding affinity of CaM with CaMKIIp, a previously study established all-atom molecular dynamics simulations to assess the structural bases for these kinetic findings and validated them with the spectra from the circular dichroism experiment [133]. Specifically, three peptides: wild-type (RRK) and its 1-amino-acid (RAK) and 3-amino-acid (AAA) (Figure 3.2 B) mutants were analyzed. The simulations show the existence of a dominant beta-strand secondary structure in the AAA peptide that does not exist in RRK and RAK peptides. This suggests that each charged amino acid mutation reduces the disordered content of the

peptide's structure ensemble. As a result, the charged amino acid mutation, AAA, presents a significant conformational change from a disordered structure in RRK to an ordered structure in AAA. The computational studies explained the changes in the CD spectra and provided the necessary structural ensemble for probing the molecular basis for how the interactions with the target peptides tune the CaM's affinity for Ca^{2+}.

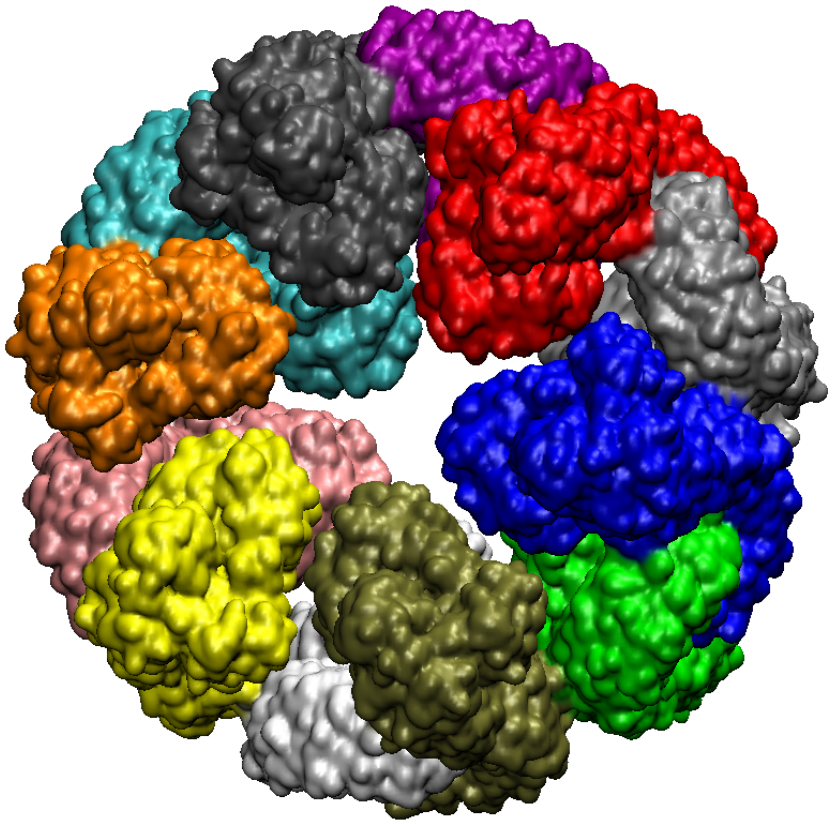

Figure 3.1. Cryo-EM structure of the mouse CaMKII dodecamer, PDB ID 7VTQ.

Calcium (Ca^{2+}) signaling is a dynamic system where Ca^{2+} concentration fluctuates in range of 0.1 to 10 μM with time [137]. These short transient Ca^{2+} around the entry sites activate Ca^{2+}-binding proteins such as calmodulin (CaM) [138]. CaM is a dumbbell-shaped protein that is grouped into two (N- and C-) domains [14] where each domain contains two helix-loop-helix (EF-hand) binding

sites for Ca^{2+} (Figure 3.2 A). The prototypical pathway describes CaM as encoding a Ca^{2+} signal by selectively activating downstream CaM-dependent proteins through molecular binding [139, 140]. However, CaM's intrinsic Ca^{2+}-binding properties alone appear insufficient to decode rapidly fluctuating Ca^{2+} signals [141]. We propose that CaM-target interactions directly tune the CaM's affinity for Ca^{2+} through reciprocal interactions that utilize multi-body molecular interactions as a basis to achieve high specificity among vast selection of targets.

The reciprocal relationship between calcium, CaM and CaM-target was also manifested in CaM's affinity for Ca^{2+} in the presence or absence of the binding peptides, CaMKIIp and neurogranin, Ng(13-49). The experiment shows a lower CaM's affinity for Ca^{2+} in the presence of Ng(13-49) and a considerable increase in the CaM's affinity for Ca^{2+} in the presence of CaMKIIp [132, 142]. With an integrated approach of quantum mechanical calculations and steered all-atom molecular dynamics simulations, we computed the free energy differences of CaM's affinity for Ca^{2+} in the presence of CaMKIIp or Ng(13-39), relative to CaM's. We showed that CaM's binding of CaMKIIp stabilizes the two Ca^{2+}-binding loops [143] via the formation of beta-sheet scaffold between the two EF-hands in each of the N or C-domain of CaM, contributing to a large increase in CaM's affinity for Ca^{2+}. However, because the acidic region in Ng(13-39) includes a proline residue that breaks the helical content of the peptide, CaM's binding of Ng(13-39) further destabilizes the two Ca^{2+}-binding loops in CaM. Hence, target binding of CaMKIIp or Ng tunes CaM's affinity for Ca^{2+} in opposite directions.

Here, we established a residue-level understanding for how target binding reciprocally influences CaM's affinity for Ca^{2+} by capitalizing on a tractable synthetic peptide model of the Ca^{2+}/CaM-

binding domain of Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) (Figure 3.2 B) [15]. Altering a subset of highly charged amino acids nearby the core CaM-binding site (residues 296-309) from the CaMKIIp, where amino acids 296-RRK-298 are substituted by three alanines, 296-AAA-298 (Figure 3.2 B), led to a significant decrease (up to 3000-fold) in its Ca^{2+}/CaM-binding affinity when compared to the wild-type CaMKIIp [129]. This example and several others have pointed that Ca^{2+}/CaM/target interactions are a multi-body problem where target binding enhances or diminishes CaM's affinity for Ca^{2+} [132, 142]. We hypothesized that the Ca^{2+}-binding loop of CaM's EF-hand tunes the local electrostatics by varying the loop conformations. In the integrative framework of the four EF-hands across two domains, binding to an individual target potentially tunes CaM's affinity for Ca^{2+}, which accounts for the multi-body interactions.

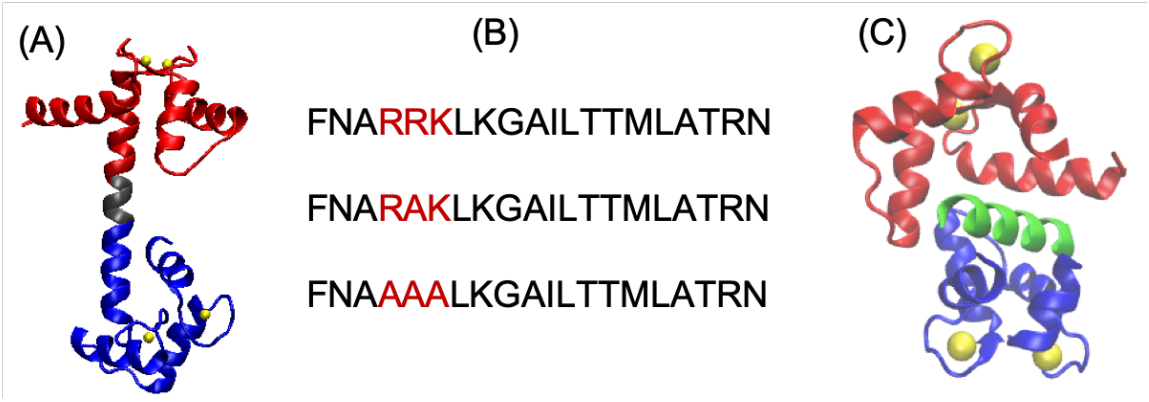

Figure 3.2. (A) The structure of Ca^{2+}/CaM in an extended conformation (PDB ID 1CLL) [144]; (B) A family of peptide sequences from the Ca^{2+}/CaM-binding domain of CaMKII used in this study; (C) Ca^{2+}/CaM in complex with the CaMKII peptide (PDB ID 1CDM) [128]; N- and C-domains of Ca^{2+}/CaM are shown in red and blue respectively. The flexible linker is shown in gray. CaMKIIp is shown in green. The yellow spheres represent the calcium ions. CaM is dumbbell shaped. One Ca^{2+} ion binds to each of the four EF-hands, two in the N-domain and two in the C-domain [144]. Ca^{2+}-binding produces conformational changes in each domain from a closed to an open conformation [14] (A), exposing a hydrophobic patch that binds a target. It is critical to appreciate that the on- and off-rates of $Ca^{2+}$are unique for each domain of Ca^{2+}/CaM. The N-

domain has fast on- and off-rates of Ca^{2+} while the C-domain has slower on- and off-rates [145]. When Ca^{2+} is oscillating, the N-domain exhibits more fidelity in tracking the rising and falling phase of the oscillation while longer sustained Ca^{2+} elevations are required to saturate the C-domain. [22]

We employed a combined approach of stopped-flow experiments and molecular dynamic simulations to test this hypothesis. It is computationally prohibitive to explore a coupled event of Ca^{2+}/CaM folding and target binding with all-atom molecular dynamics (MD) simulations[86, 146-149]. Therefore, several groups adopted simplified, coarse-grained models that sample the transitions only between the known extended and the collapsed (target-wrapped) states of Ca^{2+}/CaM (i.e. dual-basin model) [88, 150-152]. The construction of such a model strictly depends on the choice of the two states, which makes it less appealing to capture a broad ensemble of intermediate states attributing to varying Ca^{2+} signals [153]. In addition, we must deal with the status quo of current molecular dynamics force fields for Ca^{2+} that lack an adequate description for divalent-protein interactions, whereas first-principle quantum mechanical calculations involving Ca^{2+} and a flexible Ca^{2+}-binding loop are prohibitively expensive computationally [154].

To overcome the abovementioned technical challenges, we employed the approach of coarse-grained modeling and simulations using a well-established Associative memory, Water mediated, Structure and Energy Model (AWSEM) [155-157] that enables learning the coarse-grained force fields from experimental constraints to define residue specific interactions between $Ca^{2+/}$CaM, and target peptides. We trained the AWSEM force fields from the protein database and experimental measurements as constraints, as well as from our prior investigations on CaM's binding targets including the CaMKIIp wide-type 296-RRK-298, and the two mutants 296-RAK-298 and 296-

AAA-298. For example, the AWSEM force field of Ca^{2+}/CaM was determined from known protein structures and their properties in solution (Figure 3.3 A). That includes the coordination chemistry of Ca^{2+} distinctive to each of Ca^{2+}-binding loops of CaM (Figure 3.3 B). In addition, the propped Ca^{2+}-binding loops in Ca^{2+}/CaM were chosen from top-ranked conformations determined by quantum mechanical calculations [154] (Figure 3.3 C). Regarding the conformations of CaMKIIp variants, we selected them from dominant polymorphs characterized by the circular dichroism (CD) spectra and all-atom molecular dynamics simulations [158] (Figure 3.3 D). Now armed with trained coarse-grained force fields, we investigated the molecular underpinning of CaM's affinity for Ca^{2+} in the stopped-flow experiments, tuned by specific mutations in the CaMKIIp.

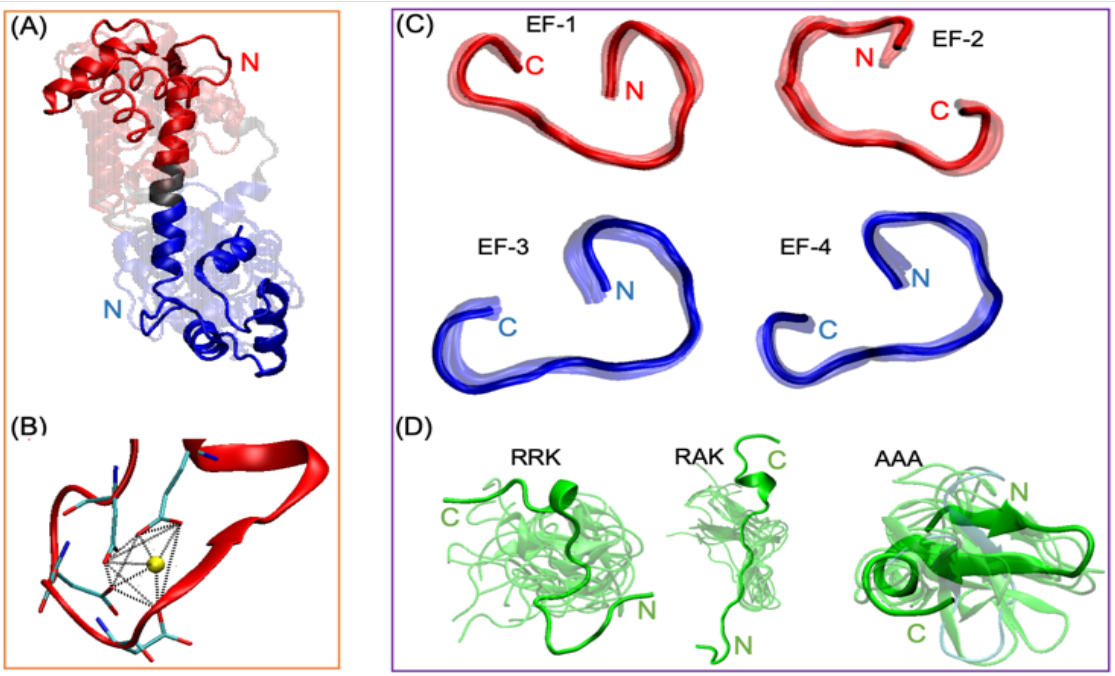

Figure 3.3. Structural variations of the associative memories to recall the conformations in the coarse-grained simulations. They do not bias the formation of associated complexes. The orange

box encloses (A) and (B) representing structures derived from experiments. The purple box encloses (C) and (D) representing the structures derived from the first principle calculations or all-atomistic simulations. (A) Representative structures of multiple conformations of CaM. N-CaM and c-CaM are shown in red and blue, respectively, and the linker is shown in gray [14]. (B) An extracted Ca^{2+}-binding loop 1 (EF-hand 1) of the Ca^{2+}/CaM that shows the pentagonal bipyramidal geometry of the residues that coordinate Ca^{2+}. (C) EF-hand structures of the Ca^{2+}-binding loops of CaM obtained from the first principle calculations [159]. Each loop is the superposition of the 10 most populated conformations. EF-1 and EF-2 are the Ca^{2+} binding loops 1 and 2 of the N-domain of CaM. EF-3 and EF-4 are the Ca^{2+} binding loops 3 and 4 of the c-CaM. (D) 10 most populated conformations of the CaMKIIp variants obtained from all atomistic simulations guided by circular dichroism experiment [158]. RRK represents the wild type peptide; RAK represents the peptide with 1-amino acid mutation; AAA represents the peptide with 3-amino acid mutation. N and C represent the N- or C-terminus of each structure.

3.2 STOPPED-FLOW FLUORESCENCE MEASUREMENT.

We quantified the rates of Ca^{2+}-dissociation from CaM using the increase in fluorescence from the fluorescent Ca^{2+}-chelator, Quin-2. An Applied Photophysics model SV.17 MV sequential stopped-flow fluorimeter was employed with excitation set at 334 nm and emission collected using a 435 nm cut-on filter. The buffer for all experiments was 20 mM MOPs, pH 7.5, 100 mM KCl and CaM and peptides were diluted in this stock buffer. CaM or the isolated C-terminal domain of CaM were prepared as previously described [160] and were used at a final concentration of 4 uM with Ca^{2+} concentration at 25 uM to ensure saturation of CaM's Ca^{2+}-binding sites. Individual peptides were added to these reactions at 10 uM final concentration to ensure saturation of binding to Ca^{2+}/CaM. Dissociation was initiated via mixing with 300 uM Quin-2 and resulting fluorescence traces were fit with one or two exponential components, as described in our previous studies [127, 129]. For each set of experiments, Quin-2 was rapidly mixed (1.7 msec instrument

deadtime) with either CaM, or the isolated C-terminal domain of CaM, in the presence or absence of the RRK, RAK or AAA peptides. The timeframe of data acquisition was adjusted to meet increased or decreased rates of Ca^{2+}-release from CaM. Data from five to six injections were averaged and all experiments were repeated at least two times. We further quantified the moles of Ca^{2+} released by calibrating the intensity of the Quin-2 dye with a series of known Ca^{2+} concentrations before each experiment. CaM and the C-terminal domain of CaM were produced by bacterial expression and purified exactly as previously described [160]. Peptides were synthesized by LifeTein to a purity greater than 95% and identify verified via mass spectroscopy. Their sequences are FNARRKLKGSILTTMLSTRN, FNARAKLKGSILTTMLSTRN and FNAAAALKGSILTTMLSTRN which represent amino acid residues 293-312 of the CaM-binding domain of CaMKII (numbering from the alpha isoform of rat CaMKII).

3.3 COARSE-GRAINED MOLECULAR SIMULATIONS

3.3.1 *Initial configuration*

The initial system for the simulations is composed of Ca^{2+}/CaM in the extended conformation (PDB ID: 1CLL) [144] (Figure 3.2 A) and the extracted peptide from the bound complex of the Ca^{2+}/CaM/CaMKIIp (PDB ID: 1CDM) [128] (Figure 3.2 C). For the peptide, we used the wild-type from CaMKIIp, the RAK peptide (peptide with 1-residue mutation, R297A), and the AAA peptide (peptide with 3-residue mutation, R296A, R297A, and K298A). We placed both Ca^{2+}/CaM and CaMKIIp in the simulation box at the distance of 20 Å apart from each other, at the temperature of 300 K, with an initial velocity randomly assigned using Boltzmann distribution ($3k_BT/m$), where k_B is the Boltzmann constant, m is the mass of C-alpha atoms, and T is the temperature in Kelvin.

3.3.2 *AWSEM force fields for Ca^{2+}/CaM and CaMKII peptides*

We leveraged the AWSEM coarse-grained force field [14, 97, 155] for our simulations. We used the structures from the past studies on Ca^{2+}/CaM or CaMKIIp [14, 158, 159, 161] as constraints (aka associative memories) to investigate the interactions between Ca^{2+}/CaM and CaMKII peptides. We used a total of 60 memories of Ca^{2+}/CaM from the protein databank [14], 80 memories of Ca^{2+}-binding loops (20 for each EF-hand loop) from quantum mechanical calculations [159], and 30 memories of the CaMKIIp variants (10 for each peptide) from all-atom molecular dynamics simulations guided by CD spectra [158].

Ca^{2+} binding proteins in these cases contain the EF-hand Ca^{2+} binding motif, forming a pentagonal bipyramidal geometry chelating the Ca^{2+} ion. We considered the effect of Ca^{2+} by assigning the Ca^{2+} charges to the negatively charged residues in the Ca^{2+} binding motifs such as defined in the previous Chapter 2 [14]. In the AWSEM parameter optimization for CaM, we treated the N- and C-domains of CaM as globular proteins and the flexible linker that connects both domains as an intrinsically disordered peptide [14]. This specific parameter optimization allows us to model the dynamics of CaM that samples the major conformations, extended and collapsed, with numerous intermediate conformations. I further considered the effect of the Ca^{2+} by assigning its charges to the negatively charged residues in its pentagonal bipyramidal geometry, which allowed us to model the dynamics of CaM.

The justification of using coarse-grained models for the simulations of the folding and/or dynamics of proteins is driven by the prohibitively large cost of computational time. As a result, all atomistic molecular dynamics simulation is often limited to a relatively small protein system

with a short biological time scale in dynamics [162]. Since the size of an average protein is over 400 residues and important biological functions occur in the range of microsecond to second, the coarse-grained model has emerged as a good alternative [47, 162]. The early development was based on Go type lattice model [163], which assumes that the folding energy landscape of an ideal protein resembles a perfect funnel picture. In this situation, monomer contact energies are biased to their pre-defined native structure. Nowadays, there are models beyond the Go type lattice models including MARTINI [164], CABS [165], and AWSEM [43]. Their force fields are purely sequence-based, but that sometimes use a Go-model as control for exploring the limit of strong specificity. AWSEM uses a mix of physical and bioinformatics approaches, as compared to other sequence-based models, which are exclusively based on knowledge-based (CABS) or physics-based (MARTINI). AWSEM learns distinctive structural biases from a fragment of short sequences of nine residues or less in a protein database. These short fragment memories allowed us to model the two domains of CaM (N- and C-domains) independently as globular proteins while keeping the flexible linker as an intrinsically disordered-like protein [14].

3.3.3 *Simulations*

We performed the coarse-grained simulations with the coarse-grained force field AWSEM [97, 155], which is implemented in the open-source MD package LAMMPS [166]. We used the cutoff distance of 3.5 Å and κ (inverse of the Debye length) = 1.4 $Å^{-1}$ in the electrostatic term. We further used enhanced sampling based on the umbrella sampling (US) [167-169] with the radius of gyration of Ca^{2+}/CaM as the reaction coordinate (Figure 3.4 and Figure 3.5). We conducted the production simulations in a canonical ensemble (NVT) where we ran 5 simulations at each US window using the same initial configuration and different random initial velocities, for a total of

50 runs per simulation set. We kept the integration time step at 2 fs and coordinates were recorded every 1000-time steps for further analyses. We then used the multi-state Bennett acceptance ratio (MBAR) [170] estimator, implemented in python, pyMBAR package, to remove the constraint on the data before the free energies calculation and other analyses.

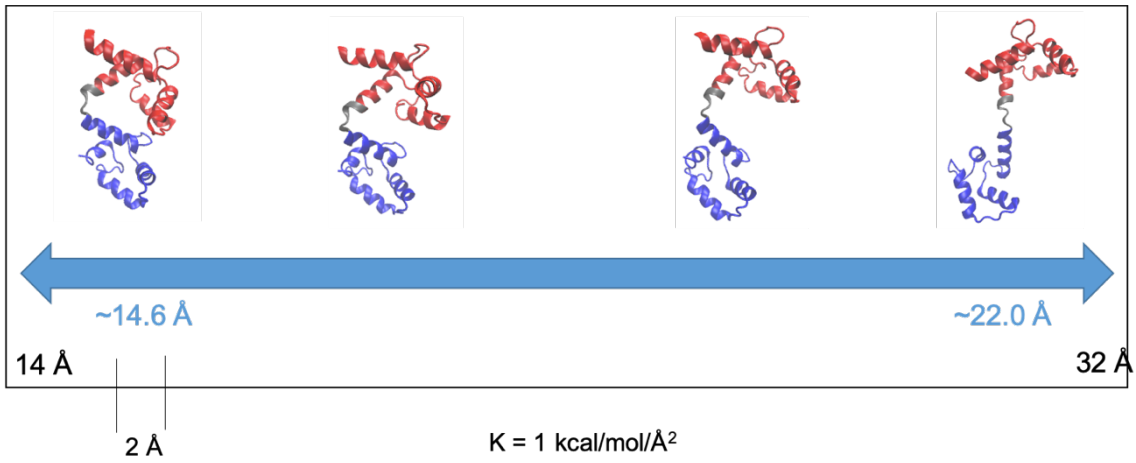

Figure 3.4. The setup of umbrella sampling [22]

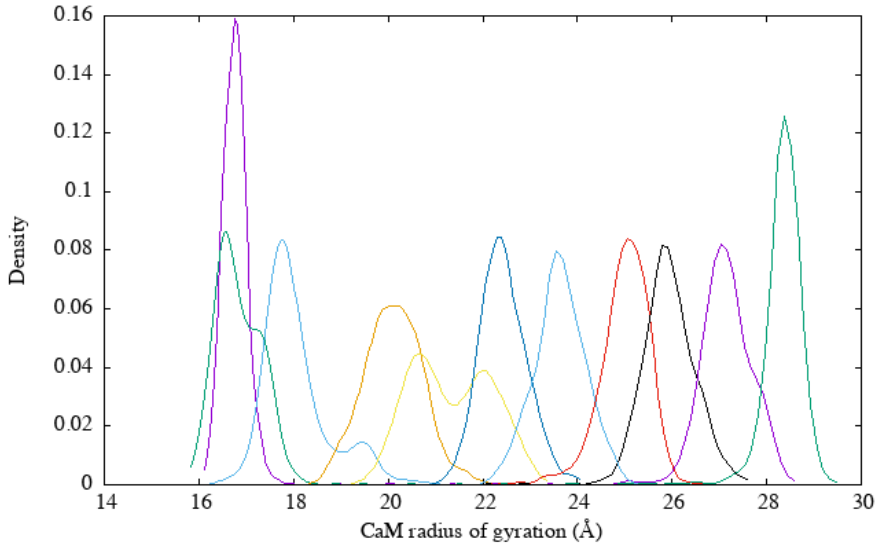

Figure 3.5. The density distribution function of radius of gyration. Each color represents a specific window from the umbrella sampling simulations, spanning from 14 to 32 Å in increment

of 2 Å. The overlaps among windows show that the increment is sufficient to cover the sampling of the entire configurational space. [22]

3.3.4 *Data analysis*

We analyzed the structural variations by calculating the radius of gyration (Rg), the potential of mean force (PMF), the pairwise comparison (Q-interface) at the interface of the bound complex. Q-interface compares the contacts formed at the interface between Ca^{2+}/CaM/CaMKIIp from the simulations and that from the experimental structure (PDB ID 1CDM). If Q-interface = 1, then contacts formed between Ca^{2+}/CaM and CaMKIIp are the same as that of the 1CDM; if Q-interface = 0, then no contact is formed. We further used the residue root-mean-square fluctuation (RMSF) of the Ca^{2+}/CaM in the presence of the CaMKIIp variants to assess the effects of the peptide binding on the CaM's calcium properties. The cutoff distance to define a close contact is 7 Å. We studied the binding and the interactions between Ca^{2+}/CaM and CaMKIIp variants by plotting the potential of mean force (PMF), the local pairwise comparison, and the contact map of the Ca^{2+}/CaM/CaMKIIp bound complex for each simulation set.

For the 1-D free energy profile, we used the radius of gyration, Rg, of Ca^{2+}/CaM as the reaction coordinate, which is restrained by a harmonic potential $E^{i}_{Restr} = \frac{1}{2}\mathrm{k}(\mathrm{Rg^{i}} - \mathrm{Rg^{i}_{0}})^{2}$ with a force constant k = 1 kcal/Å^2. The setup of the force constant and the equilibrium positions of the windows were justified by the fluctuations of Rg around the corresponding equilibrium as well as the overlap in the distribution of Rg between neighboring windows (Figure 3.5). We constrained the radius of gyration of CaM in increment of 2 Å; for each window, I ran 5 US simulations using

the same initial configuration and different random initial velocities, for a total of 50 runs per simulation set.

For the 2-D free energy profile, in addition to the Rg, we used the pairwise comparison at the interface (Q-interface) between the two monomers to get the 2-D free energy profile. The protocols to execute 2-D US simulations and reweighing are the same as those for 1-D US sampling.

3.4 Quantification of the dissociation rate constants between Ca^{2+} and CaM in the presence of peptides that mimic the Ca^{2+}/CaM-binding domain of CaMKII

The release of Ca^{2+} from the n-terminal domain of CaM in the absence of target peptides is too fast to be measured in the employed stopped-flow apparatus (see Table 3.1). If Ca^{2+}/CaM is preequilibrated with the RRK peptide, two large differences are identified in the kinetics of Ca^{2+}-release (Table 3.1). First, the release of Ca^{2+} from the n-terminal domain of CaM has significantly slowed down (8 s^{-1}) to the point that it can now be measured. Because we calibrate the intensity of the Quin-2 dye (see Methods Section), we now see that 4 total moles of Ca^{2+} (each mole of CaM can bind 4 moles of Ca^{2+}) recovered from data fits with rates of 1.7 s^{-1} and 2.3 s^{-1} for the n-CaM and c-CaM, respectively (Table 3.1). It is possible that we are detecting some overlap of kinetic release which is why the data is not balanced into 2 and 2 or that there is preferential tuning of the Ca^{2+}-binding sites in the n-CaM and c-CaM. Second, the rate of release from c-CaM has slowed by over 50-fold (Table 3.1 and Figure 3.6). This is an immense change and reflects the fact that binding of RRK to CaM has had reciprocal effects on CaM's interactions with Ca^{2+}. We conclude that RRK has had a significant impact on the Ca^{2+}-binding properties of both the n-CaM and c-

CaM. Further, the slowed rates of Ca^{2+} release from the c-CaM could be identified because of the similarity in kinetics using the isolated c-CaM (Table 3.1).

Table 3.1. Summary of the Ca^{2+} dissociation rate from the n-CaM and c-CaM in the absence or presence of a family of three target peptides. [22]

	Peptide	Ca/mol CaM from N-domain	Rate of release from N-domain	Ca/mol of CaM from C-domain	Rate or release from C-domain	Total mol Ca recovered/mol CaM
CaM	None		TFTM	2.00	9.73	2
	RRK	1.68	8.09	2.29	0.17	3.97
	AAA	1.68	8.09	2.15	4.95	2.98
C-CaM	None			1.99	12.53	1.9
	RRK			2.20	0.65	2.2
	AAA			2.09	2.55	2.1

rates are in s^{-1}
TFTM: too fast to measure

We extended these studies to evaluate the impact of two mutant peptides of the Ca^{2+}/CaM-binding domain of CaMKII (i.e. RAK and AAA mutants of CaMKIIp) (Figure 3.2 B). Another study recently showed significant differences in their secondary structure [158] (Figure 3.3 D) as polymorphs. These polymorphic mutants have decreased binding affinity with Ca^{2+}/CaM [129] relative to the wild-type CaMKIIp (RRK). We proposed that the weakened affinity of binding to Ca^{2+}/CaM [129] would have a corresponding impact (an increase) in Ca^{2+}-dissociation from CaM. Figure 3.6 shows the results of the Ca^{2+}-dissociation rate constants from CaM in the presence or absence of the peptides from our measurements. Indeed, the rate of Ca^{2+} dissociation from c-CaM is much faster than with the RRK peptide. Remarkably, Ca^{2+}-dissociation from n- CaM is indistinguishable when bound to either the AAA or RRK peptides (Table 3.1). This is perhaps not surprising because the crystal structure obtained with Ca^{2+}/CaM bound to a peptide analogous to RRK (Figure 3.2**)** shows an anti-parallel binding where the residues including RRK interact largely

with c-CaM instead n-CaM. Therefore, the altered interactions of the RRK region in the AAA/RAK peptides are the likely reason for the weakened Ca^{2+}-binding affinity of c-CaM. Next, we used coarse-grained molecular simulations to determine the residues on CaM affected by the mutations on the CaMKIIp, and how these affected residues on CaM disturb the chemical coordination of Ca^{2+} in the Ca^{2+}-binding loops.

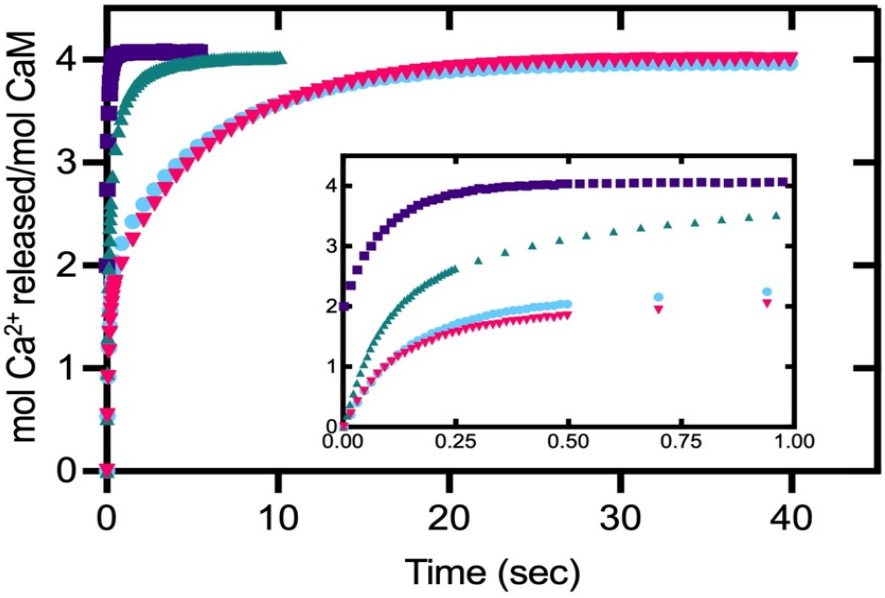

Figure 3.6. Measurement of Ca^{2+} dissociation rate constants from the time course for Ca^{2+} dissociation from CaM in the presence or absence of CaMKIIp variants using stopped-flow fluorimeter. Ca^{2+} dissociation in the presence of CaMKIIp wide type (RRK) (red); Ca^{2+} dissociation in the presence of CaMKIIp mutant with 1 mutation (RAK) (blue); Ca^{2+} dissociation in the presence of CaMKIIp mutant with 3 mutation AAA (green); Ca^{2+} dissociation from CaM in the absence of peptides (purple). [22]

3.5 MUTANT AAA PEPTIDE DEFORMS C-CAM IN THE BOUND COMPLEX WITH CA^{2+}/CAM

To determine the structural basis for the changes in CaM's affinity for Ca^{2+}, coarse-grained molecular simulations were undertaken. 1-dimentional free energy of the observed change as CaMKIIp binds to Ca^{2+}/CaM is shown in Figure 3.7. shows that the free energy minimum of the Ca^{2+}/CaM/RRK complex is at Rg ~18.5 Å while that of the Ca^{2+}/CaM/AAA complex is at Rg~16.5 Å. We explain the shift of the minima from 18.5 Å to 16 Å for the Ca^{2+}/CaM/CaMKIIp complex as follows: both Ca^{2+}/CaM and the peptides require partially unfolding before the formation of the bound complex [78, 161]. For the RRK peptide, an unstructured conformation binds to Ca^{2+}/CaM at its first encounter at Rg of Ca^{2+}/CaM ~ 22 Å, before it adopts a helical structure in the final bound complex with Ca^{2+}/CaM (Rg of Ca^{2+}/CaM ~ 18.5 Å). This helical structure of the RRK peptide then leads to the higher stability of the bound Ca^{2+}/CaM/RRK complex (Figure 3.7). On the other hand, the Ca^{2+}/CaM/AAA complex adopts a collapsed conformation, that prevents the canonical form of the bound complex to be reached. Since the formation of the complex Ca^{2+}/CaM/AAA favors fewer bond breaking in AAA, Ca^{2+}/CaM/AAA will tend to have higher stability, or lowered free energy, than the Ca^{2+}/CaM/RRK complex (Rg of Ca^{2+}/CaM ~ 16 Å). Interestingly, the free energy of the complex Ca^{2+}/CaM/RAK remains in between that of Ca^{2+}/CaM/RRK and Ca^{2+}/CaM/AAA, which further shows the consistency of our results (Figure 3.7).

The radius of gyration (Rg) of a folded Ca^{2+}/CaM in the absence of a peptide is ~ 22 Å from the previous simulations [14], presented in Chapter 2. Here, Rg of Ca^{2+}/CaM > 28 Å signified an unfolded Ca^{2+}/CaM structures. When Ca^{2+}/CaM transitions from the unfolded to the folded structure in the bound complex, the differences in free energies became noticeable at Rg between

18 and 22 Å. Their differences are as high as 4 k_BT around Rg = 20 Å (Figure 3.7). We speculated that this would be the region where both Ca^{2+}/CaM and CaMKIIp undergo mutual conformational changes when Ca^{2+}/CaM prompts to wrap around the CaMKIIp to form a bound complex.

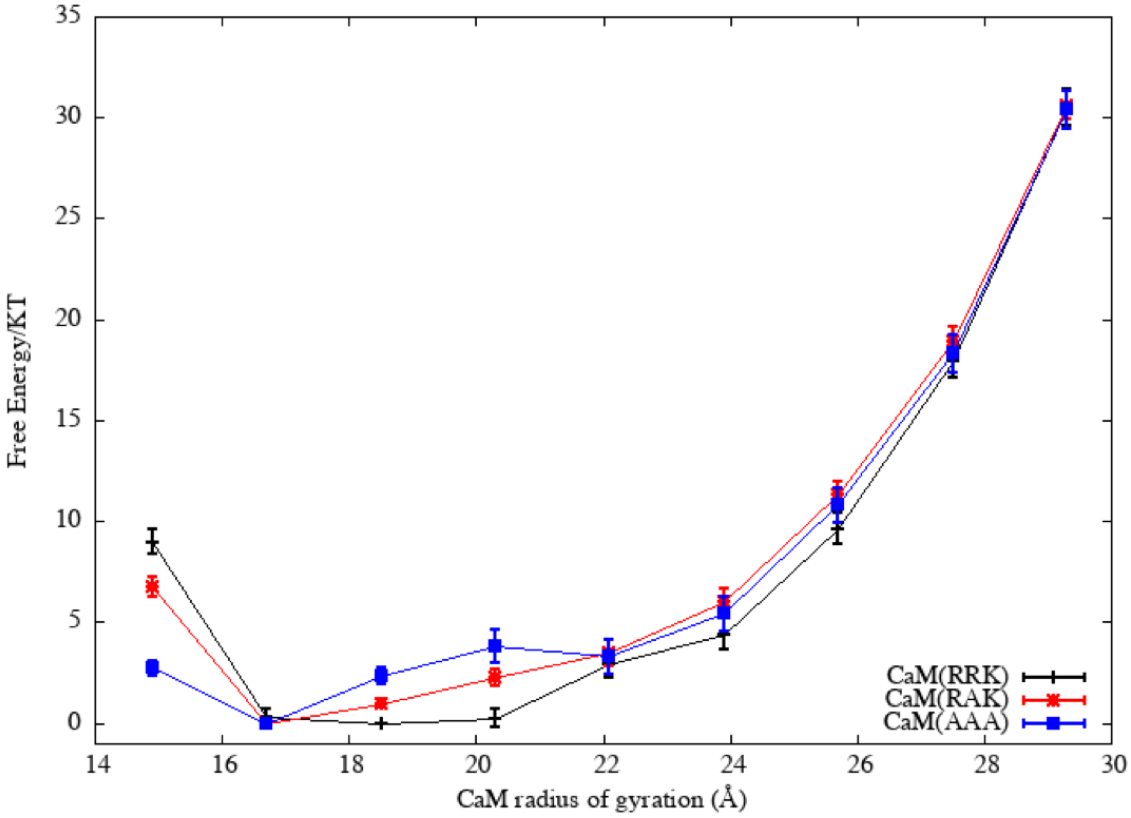

Figure 3.7. Free energy variation as a function of the radius of gyration (Rg) of Ca^{2+}/CaM and the errors from the enhanced sampling simulations. I performed the umbrella sampling simulations using the Rg of Ca^{2+}/CaM as the reaction coordinate of the complex, which constrained the Rg of Ca^{2+}/CaM from 14 to 30 Å. I used the multistate Bennett acceptance ratio (MBAR) implemented in python, pyMBAR [171, 172], to estimate the unbiased free energy of the Ca^{2+}/CaM/CaMKIIp complexes. Each color represents CaM interacting with a CaMKIIp variant, where CaMKIIp represents the wide-type RRK (black), one-residue mutant RAK (red), and three-residue mutant AAA (blue). [22]

The 2-dimensional free energy of each complex was further plotted using the Rg of Ca^{2+}/CaM and the probability of contact at the interface between Ca^{2+}/CaM and CaMKIIp (Q-interface, please see the previous section for a detailed definition) (Figure 3.8**)**. The complexes of Ca^{2+}/CaM/RRK as well as Ca^{2+}/CaM/RAK have extended the regions of stability over Rg ~18 Å and Q-interface ~0.25 (Figure 3.8 A and Figure 3.8 B), respectively. However, the stability of the Ca^{2+}/CaM/AAA complex was highly localized between Rg ~16 Å and Q-interface ~0.32 (Figure 3.8 C). It inferred that the binding of AAA peptide to Ca^{2+}/CaM has deformed Ca^{2+}/CaM in the bound complex. Together with Figure 3.7, the Ca^{2+}/CaM/AAA complex has led to a higher free energy than the wild-type, which could be unfavorable for retaining Ca^{2+}. Next, we evaluated the probability of contact between Ca^{2+}/CaM and the peptide that provided further insights on the observed changes.

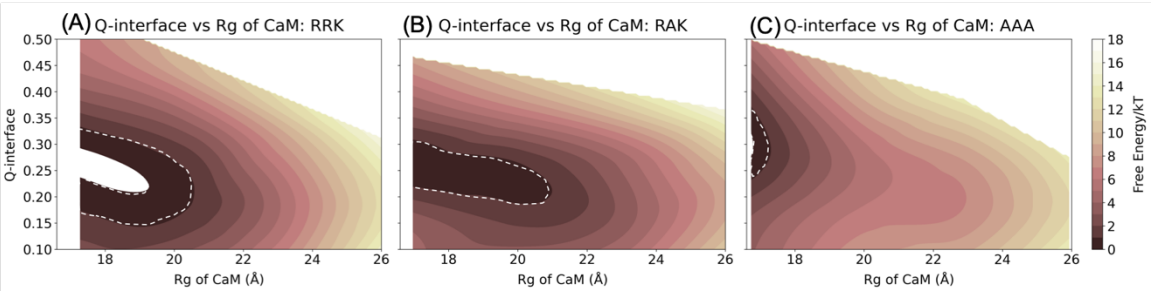

Figure 3.8. 2D-free energy profiles for Ca^{2+}/CaM binding with CaMKIIp plotted against Q-interface and the radius of gyration (Rg) of Ca^{2+}/CaM. (A) Ca^{2+}/CaM binds with the wide type (RRK). (B) Ca^{2+}/CaM binds with CaMKIIp with 1 mutation (RAK). (C) Ca^{2+}/CaM binds with CaMKIIp with 3 mutations (AAA). The areas of high stability for each complex are highlighted with dashed lines. The white regions represent regions with high free energy greater than 18kT.

3.6 CONTACT MAP OF THE BOUND COMPLEX Ca^{2+}/CAM/AAA SHOWS NEW CONTACTS WITHIN THE DOMAINS OF Ca^{2+}/CAM

To evaluate how the peptides binding modulates the properties of the Ca^{2+}/CaM/CaMKIIp bound complex in detail, we plotted in Figure 3.9 the contact maps of each complex with data from the highlighted areas in Figure 3.8. The lower triangles represent the contact map, and the upper triangles represent its variances. The results showed the similarity between Ca^{2+}/CaM/RRK and Ca^{2+}/CaM/RAK complexes (Figure 3.9 A and Figure 3.9 B). RRK and RAK peptides do not form a helix prior to binding to Ca^{2+}/CaM and positively charged residues on RRK or RAK peptides make contacts with the negatively charged residues in the c-CaM. Agreeing with a previous work [161], we observed that to form a bound complex, following the initial encounter, contacts from both Ca^{2+}/CaM and the target have to be broken to make way (around Rg of Ca^{2+}/CaM at 20 Å) for new contacts to form (around Rg of Ca^{2+}/CaM at 17 Å).

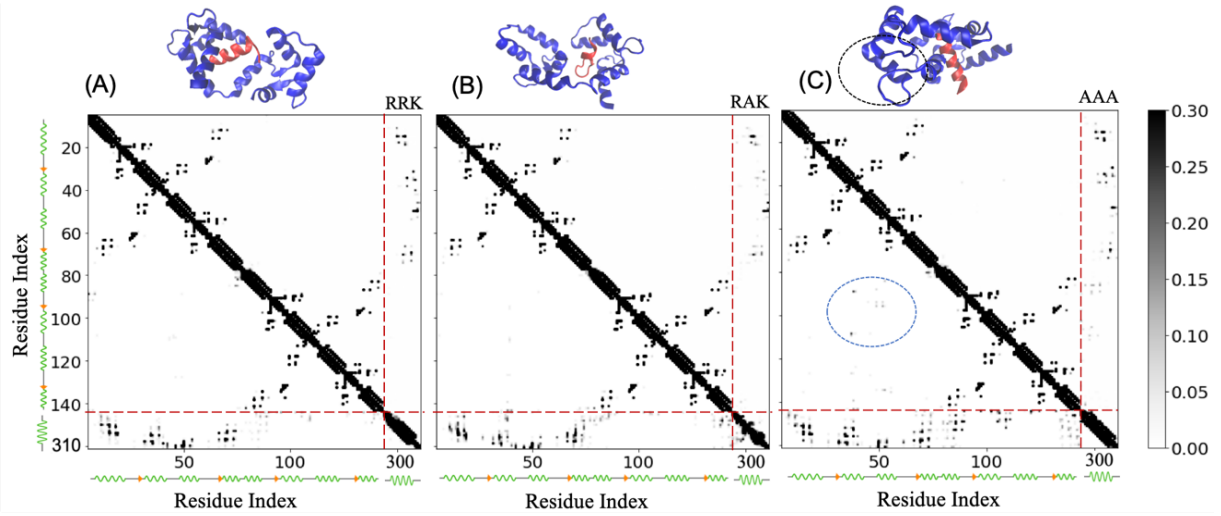

Figure 3.9. Probability of the contact formation (in lower triangles) and its variances (in upper triangles) from the highlighted areas in Figure 3. (A) For the complex of Ca^{2+}/CaM/RRK in the region defined by Q-interface = 0.16 to 0.34 and Rg = 18 to 20 Å; (B) For the complex of Ca^{2+}/CaM/RAK in the region defined by Q-interface = 0.22 to 0.32 and Rg = 18 to 20 Å; (C) For

the complex of Ca^{2+}/CaM/AAA in the region defined by Q-interface = 0.2 to 0.38 and Rg = 16 Å. The representative structure for each model is shown on top. The secondary structure for the crystal structure 1CDM is also shown with helices and beta strands (arrows). The highlighted areas in (C) show the formation of new contacts between Ca^{2+}/CaM domains (blue) and the loss of beta-sheet in the representative structure (black circle).

However, for the Ca^{2+}/CaM/AAA bound complex, contacts are formed within the domains of Ca^{2+}/CaM (Figure 3.9 C) that prevents subsequent conformational changes. The representative structure from Figure 3.9 shows that the AAA mutant peptide remains folded and the two failed to form a subsequent canonical bound complex in which Ca^{2+}/CaM wraps around the target. Furthermore, the Ca^{2+}/CaM/AAA bound complex lost the "beta-sheet scaffold" formed by the two Ca^{2+}-binding loops in the c-CaM (shown by a circle in the representative structure in Figure 3.9 C). The formation of beta-sheet scaffold is necessary to further stabilize the overall stability of Ca^{2+}/CaM/CaMKIIp wrap-around complex [132]. Figure 3.10 confirms the observations that the new contacts within the domains of Ca^{2+}/CaM in the presence of AAA (Figure 3.10 and Figure 3.10) are not present in the complexes of Ca^{2+}/CaM/RRK or of Ca^{2+}/CaM/RAK (Figure 3.10).

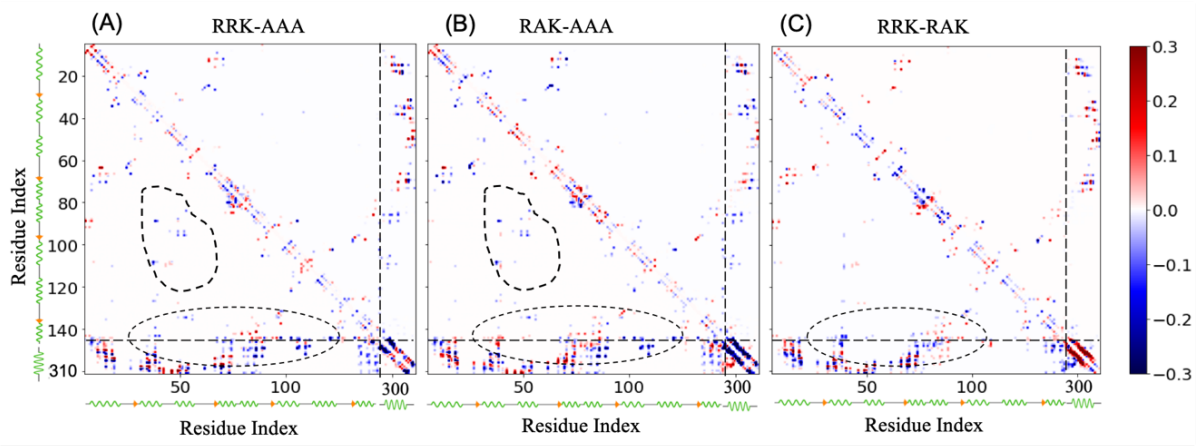

Figure 3.10. Differences in the probability of contact formation (lower triangles) and its variances (upper triangles) from the highlighted areas in Figure 3. (A) represents the contact map difference between RRK and AAA; (B) represents the contact map difference between RAK and AAA; (C) represents the contact map difference between RRK and RAK. The highlighted areas show the formation of new contacts within Ca^{2+}/CaM domains in the presence of AAA peptide. [22]

From Figure 3.10, the lower triangles (upper triangles) represent the contact map difference between two complexes (and its variances). The highlighted regions from Figure 3.10 and Figure 3.10 showed the new contacts that are absent in Figure 3.10. These observations supported the results from the probability of contact maps that the interactions between Ca^{2+}/CaM and the AAA mutant peptide led to the formation of new contacts within the domains of Ca^{2+}/CaM.

Next, we aimed to reveal the mutation from CaMKIIp variants that affects the structures of Ca^{2+}/CaM. For each complex, we clustered the structures with the Rg of Ca^{2+}/CaM ranging from 16 to 20 Å based on the root mean square deviation (RMSD) (Figure 3.11) and colored the residues with the root mean square fluctuation (RMSF) in Figure 3.12. c-CaM in both the presence of AAA and the N-terminus of AAA show the largest RMSF in Figure 3.11 C, especially residues 293-

FNA-295 of the peptide and the Ca^{2+}-binding loops of the c-CaM, relative to the N-domain of Ca^{2+}/CaM in the presence of AAA peptide, the complex of Ca^{2+}/CaM/RRK in Figure 3.11 A, or the complex of Ca^{2+}/CaM/RAK Figure 3.11 B. The specific mutation of these charged residues not only removes the electrostatic interactions among the residues (296-AAA-298) and the negatively charged residues in the c-CaM (Ca^{2+}-binding loops 3 and 4), but also deforms the beta sheet scaffold that props the Ca^{2+}-binding loops of the c-CaM including residues 90 to 94 and residues 126 to 129. As a result, such large RMSFs in these residues account for the increase rate of Ca^{2+} release from CaM as it interacts with AAA peptide, such as observed in the stopped-flow analyses (Figure 3.6). The visualization of each complex shows differences in binding between Ca^{2+}/CaM and CaMKIIp variants.

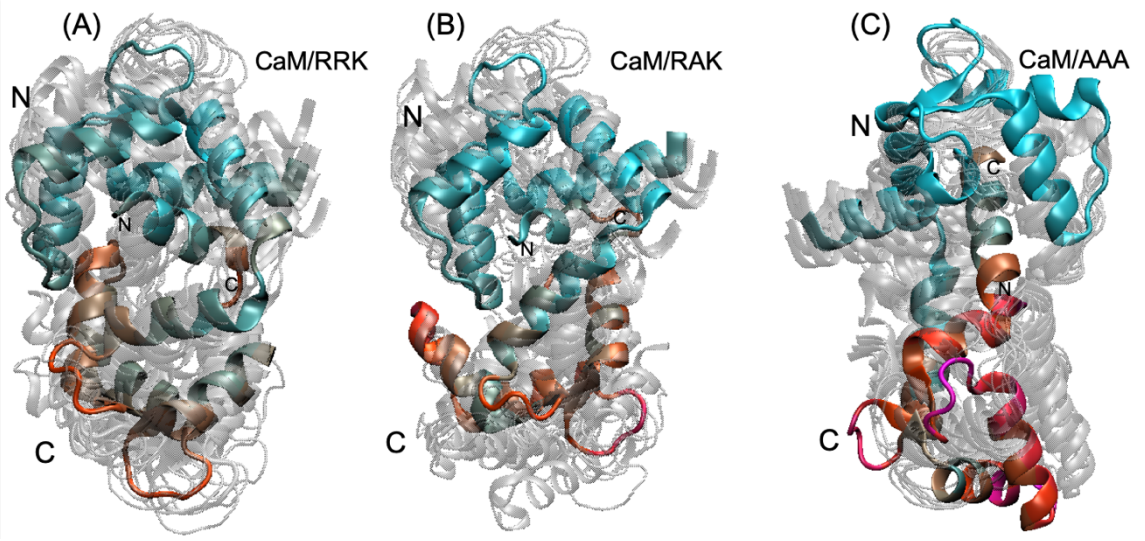

Figure 3.11. Overlapped structures of the Ca^{2+}/CaM in complex with (A) CaMKIIp wild type (RRK), (B) CaMKIIp with 1 mutation (RAK), and (C) CaMKIIp with 3 mutations (AAA). We superimposed 10 conformations from the most populated clusters, which shows the fuzziness of each bound complex. One of the representative structures (CaM and CaMKIIp) is colored by the root-mean-square frequency (RMSF) of each residue from Figure 3.12. The teal represents the regions with RMSF less than ~ 0.4 nm, orange or magenta represent the regions with moderate or

high fluctuations with RMSF greater than 2.3 nm. The residues in the c-CaM and the N-terminus of AAA from the complex Ca^{2+}/CaM/AAA in (C) show the highest RMSF. [22]

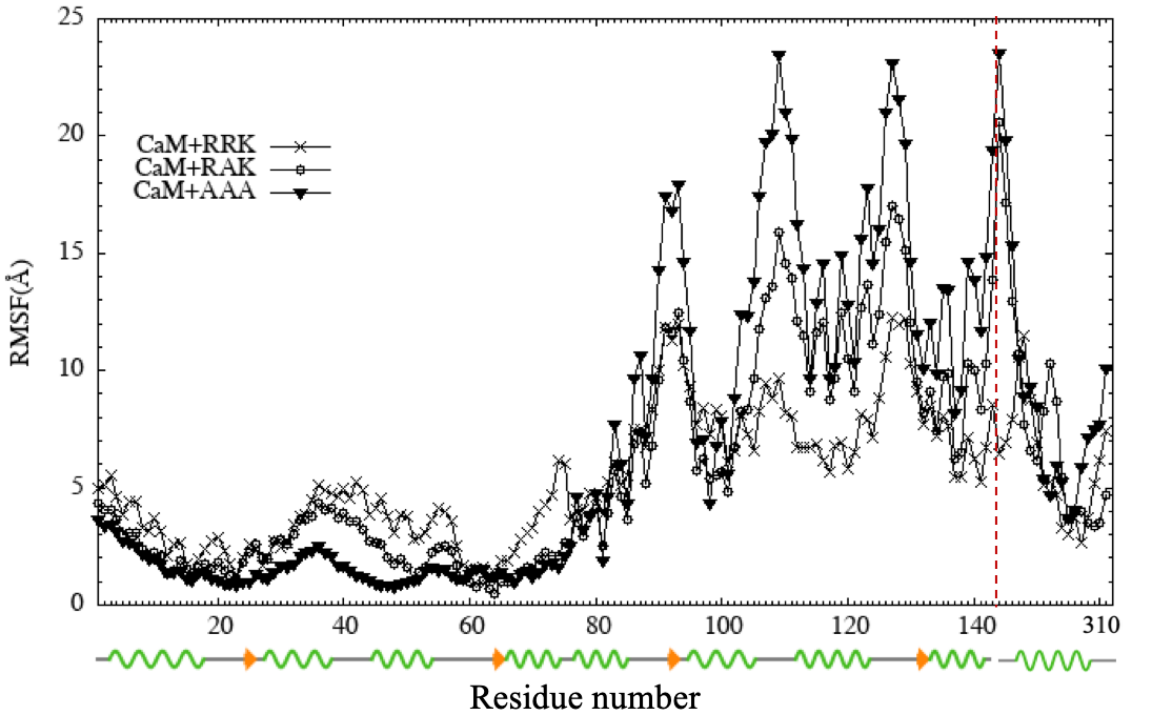

Figure 3.12. Root-mean-square-fluctuations (RMSF) of the C-alpha atoms from the structures of the bound complexes in Figure 7. RMSF spectra of each complex (Ca^{2+}/CaM/RRK, Ca^{2+}/CaM/RAK, Ca^{2+}/CaM/AAA) are shown with different symbols. [22]

3.7 AAA MUTATIONS INTERRUPT BOTH THE ELECTROSTATICS AND THE CONFORMATIONS IN THE Ca^{2+}/CaM/CaMKIIP BOUND COMPLEX FOR Ca^{2+}

The stopped-flow experiment in our study shows that although the CaM affinity for Ca^{2+} remains relatively similar for the three binding peptides at the N-terminal domain of Ca^{2+}/CaM, it dropped by over 50-fold in the C-terminal domain of CaM for the AAA peptide (Table 3.1). It shows that relative to the RRK wide-type peptide, the AAA mutant peptide has a significant negative impact on the CaM's affinity for Ca^{2+} [127].

Using the complementary coarse-grained molecular simulations, we showed that the difference in free energy of both bound complexes with either a wild type or a AAA mutant peptide, is as high as $4k_BT$ (Figure 3.7). We investigated the differences in the residual interactions between Ca^{2+}/CaM and the CaMKIIp variants from the simulations. These comparative studies (Figure 3.12 and Figure 3.13) showed that contact formation between hydrophobic residues such as 293-FNA-295 of CaMKIIp and M124 from the C-terminal domain of Ca^{2+}/CaM follows those contact formations between the oppositely charged residues such as 296-RRK-298 of CaMKIIp and several others from Ca^{2+}/CaM. Specifically, R296 of CaMKIIp forms contact with E120, E123, E127 on the EF-hand 3 from the c-CaM. R297 of CaMKIIp forms contact with E7, E11, E14 on the EF-hand 1 from the N-domain of Ca^{2+}/CaM. R298 forms contact with residues from both the n-CaM and c-CaM -- with E14 on the EF-hand 1 and E114 on a Ca^{2+}-binding loop 3, respectively (Figure 3.13). Mutations from RRK to AAA certainly diminished the electrostatic interactions.

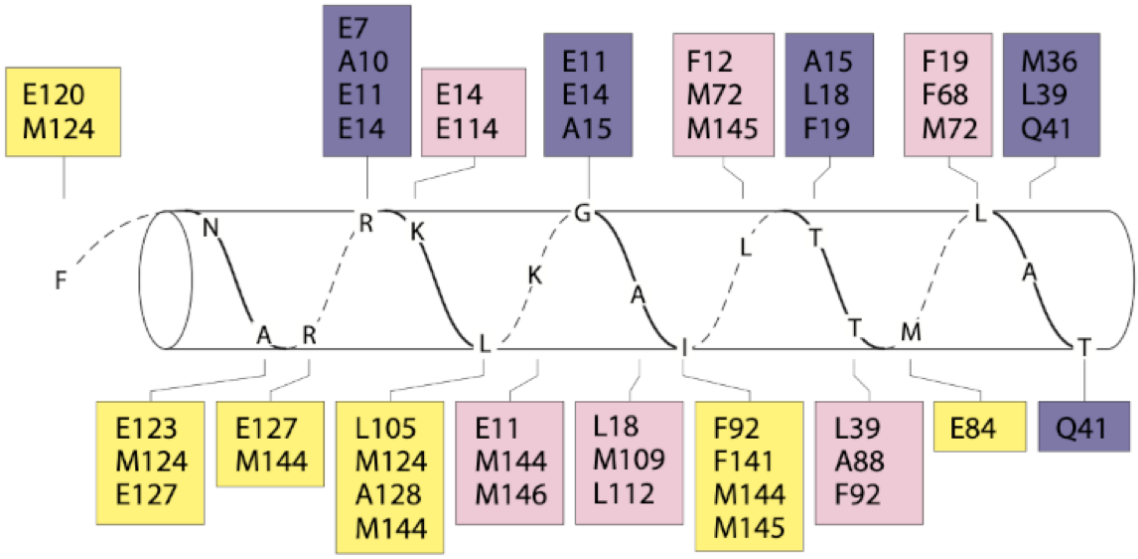

Figure 3.13. Diagram of CaMKIIp (293-314) binding to Ca^{2+}/CaM derived from a crystal structure data. Purple are amino acids from the N-terminal domain of Ca^{2+}/CaM that interact with CaMKIIp; yellow are amino acids from the C-terminal domain of Ca^{2+}/CaM that interact with

CaMKIIp; pink are amino acids from both N- and C-domains of Ca^{2+}/CaM that interact with CaMKIIp. [22]

Moreover, when we compared the structures of the Ca^{2+}/CaM/AAA (Figure 3.11 C) against the Ca^{2+}/CaM/RAK complex (Figure 3.11 B) and the Ca^{2+}/CaM/RRK complex (Figure 3.11 A), I discovered significant structural deformations in the c-CaM than the n-CaM. The large RMSF deviation in the c-CaM is caused by binding with a very different polymorph of the AAA mutant peptide that forms non-specific contacts following 293-FNA-295.

Conclusively, mutations of RRK into AAA not only diminish these electrostatic interactions between the charged amino acids with Ca^{2+}/CaM, but also provide a different polymorph that disrupted the beta-sheet scaffold in the EF-hands of Ca^{2+}/CaM. In addition, the Pearson correlation between the domains of Ca^{2+}/CaM and the peptides (Figure 3.14 and Figure 3.15) shows the positive correlation between the c-CaM and the AAA mutant peptide (Figure 3.14 C), which explains their high RMSF observed in the C-terminal domain of Ca^{2+}/CaM as it interacts with AAA (Figure 3.11), underscoring the enhanced dissociation of Ca^{2+} observed in the stop-flow experiments (Table 3.1).

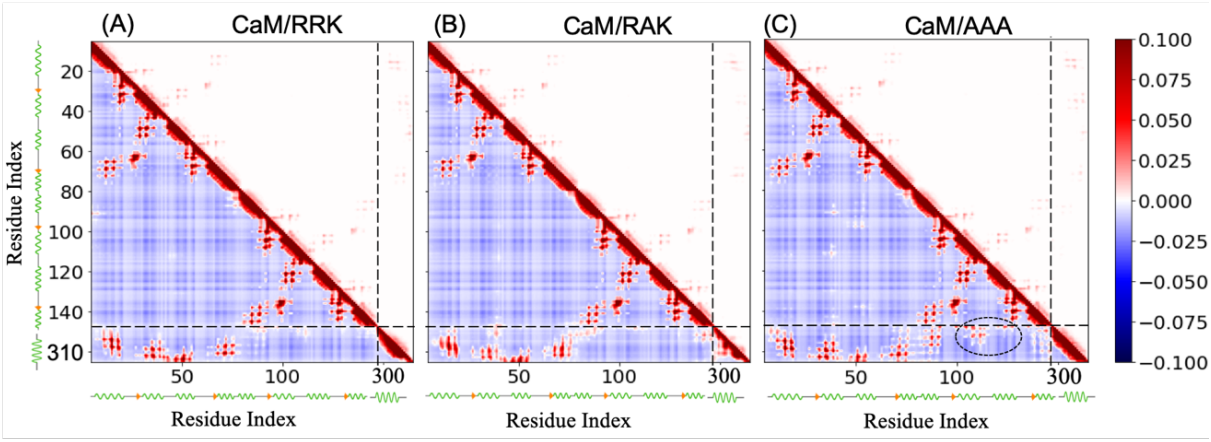

Figure 3.14. Pearson correlation of the probability of contact formation from the highlighted areas in Figure 3. Lower triangles represent the Pearson correlations and upper triangles represents their variances. (A) shows the complex of Ca^{2+}/CaM/RRK; (B) shows the complex of Ca^{2+}/CaM/RAK; (C) represents the complex of Ca^{2+}/CaM/AAA. The highlighted areas in (C) shows positive correlations between EF-hand 3 of the C-terminal domain of CaM and the AAA peptide which are not seen in (A) and (B). [22]

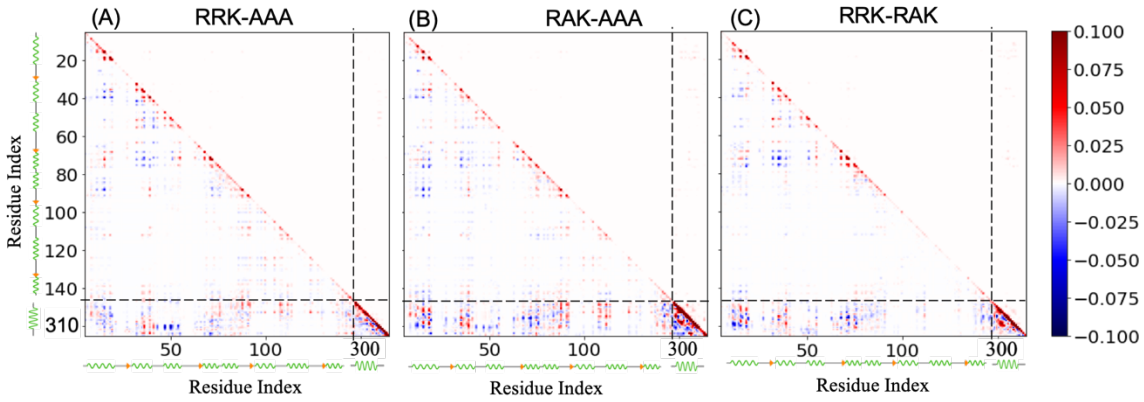

Figure 3.15. Pearson correlation differences in the probability of contact formation from the highlighted areas in Figure 3. Lower triangles represent the Pearson correlation differences and upper triangles represent their variances. (A) shows the results from the complexes of Ca^{2+}/CaM/RRK- Ca^{2+}/CaM/AAA; (B) shows the results from the complexes of Ca^{2+}/CaM/RAK- Ca^{2+}/CaM/AAA; (C) shows the results from the complexes of Ca^{2+}/CaM/RRK- Ca^{2+}/CaM/RAK. [22]

3.8 A COMBINED APPROACH OF EXPERIMENTS AND COMPUTER SIMULATIONS PROVIDES INSIGHTS INTO THE RECIPROCAL RELATIONS IN Ca^{2+}/CaM/TARGET-BINDING

CaM has been a subject of interest from the computational community because of its coupling of structures and functions: folding, conformational changes, and Ca^{2+}-binding. A few molecular dynamics simulations focus on separate incidences because each of these topics requires enormous computational resources and modeling efforts: CaM's allostery [74, 152, 173-176], the CaM-target binding [78, 148, 177], or the coordination chemistry of Ca^{2+} in the Ca^{2+}-binding loops in CaM [159, 178-180]. The energy landscape view [153] for the interplay among folding, allostery, and binding of CaM domains was provided to investigate possible routes CaM might explore in response to external stimuli.

These investigations are the first to report how the interaction of two "fuzzy" components of Ca^{2+}/CaM and CaMKIIp tunes the Ca^{2+} affinity at a residue level by using the AWSEM force fields that learn from the provided structures as associative memories. Although the coarse-grained protein representation lacks atomistic details, those key conformational motifs are retained in the associative memories of the AWSEM force fields. For example, the AWSEM force fields recall the conformations of the Ca^{2+}-binding loops of CaM propped by the coordination chemistry of Ca^{2+} (Figure 3.3 C) from the quantum mechanical calculations. This capability to learn from the propped loop structures enables the investigation of how specific mutations on a target that binds Ca^{2+}/CaM impact the interplay among the Ca^{2+}-binding loops of CaM on EF-hand 3 and EF-hand 4.

3.9 CONCLUSION

In this work, we have used a series of peptides that mimic the Ca^{2+}/CaM-binding domain of CaMKII, namely RRK (wild type), RAK (1-residue mutation), and AAA (3-residue mutation), to show that their interactions with Ca^{2+}/CaM modulate CaM's Ca^{2+}-binding properties. The wild-type peptide binding to Ca^{2+}/CaM leads to high CaM affinity for Ca^{2+} while binding the peptide with the 3-residue mutation leads to an over 3000-fold decrease in CaM affinity for Ca^{2+} relative to the wild type. On the other hand, a peptide with a 1-residue mutation did not show a considerable change relative to the wild type, although a disturbed C-domain of Ca^{2+}/CaM was observed. Although we devised coarse-grained molecular simulations specifically for Ca^{2+}/CaM/CaMKIIp by using the AWSEM force fields that learn from the provided structures as associative memories, the integrative approach of experiments and simulations can be adopted to investigate the regulatory mechanism of other fuzzy complexes that lack distinctive features and are difficult to characterize by experiments.

This first part of this thesis involved either one protein (CaM) or two proteins (CaM/CaMKIIp). However, inside the cell, most biological functions are driven by more than two proteins, known as protein complexes. Therefore, how these proteins come together in a crowded environment like cell to form a specific protein complex, unique to a biological function remains incompletely understood. In the second part of the thesis, I will introduce an ongoing effort to develop an integrative modeling approach to investigate the three-dimensional structures of protein complexes.

Chapter 4. INTEGRATIVE MODELING OF THE MACROMOLECULAR COMPLEX ASSEMBLY

Protein complexes stand as the molecular architects of life, orchestrating a vast array of biological processes essential for cellular function and organismal viability. Comprising multiple protein subunits intricately bound together, these complexes exhibit a remarkable diversity in structure, function, and dynamics, playing pivotal roles in cellular signaling, metabolism, gene regulation, and structural support. Understanding the organization and the 3D structural assembly process of protein complexes is fundamental to unraveling the molecular mechanisms underlying complex biological phenomena, from enzymatic catalysis to cellular communication and disease pathogenesis. By developing an integrative modeling approach that combines experimentally guided computational techniques, we aim to unravel the three-dimensional structures of protein complexes, which will guide the current experimental approaches such as cryo-electron macroscopy during the refinement process, indeed deepening our understanding of their biological roles and regulatory mechanisms.

4.1 INTRODUCTION

The elucidation of three-dimensional (3D) structures of protein complexes represents a pivotal endeavor in structural biology, offering profound insights into the molecular machinery governing key biological processes.[31, 181, 182] Protein complexes, comprised of multiple protein subunits intricately arranged in space, orchestrate a myriad of cellular functions.[182] Understanding the architecture and dynamics of these complexes is essential for unraveling their roles in health and

disease, as well as for designing targeted therapeutics aimed at modulating their activity with precision.

The determination of protein complex structures presents unique challenges compared to individual proteins due to the increased complexity arising from the interactions among multiple components.[181, 182] Traditional structural biology techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) have been instrumental in revealing the three-dimensional arrangements of protein complexes (Table A.1). These methods have enabled ones to decipher the spatial organization of complex assemblies, identify critical interface regions mediating protein-protein interactions, and unveil the mechanisms underlying their biological functions.[181] However, the highly dynamics nature of protein complexes, their large sizes, and the transiently interacting components have limited the utilization of these experimental techniques. The use of multiple techniques, taking advantage of the strength of each individual technique, has been proposed to overcome these difficulties, but they also lack the ability to capture the transient interactions or intermediate steps during protein complex formation.[183, 184]

Complementing experimental techniques, computational methods play a crucial role in predicting, modeling, and refining protein complex structures.[185, 186] Molecular docking, protein-protein interaction prediction algorithms, and molecular dynamics simulations provide invaluable tools for exploring the conformational space of protein complexes, predicting binding modes between interacting partners, and simulating dynamic processes occurring within complex assemblies.[186] These computational approaches not only aid in interpreting experimental data but also facilitate the rational design of protein complexes with desired functionalities for biomedical and

biotechnological applications. Integrative modeling approach, which combines the strength of individual modeling approach and the high-quality data produced experimentally for individual protein, has been proposed to further assess the 3D structure determination of protein complexes, benefiting from the significant advancements in experimental and computational methodologies, which have propelled the field of protein complex structure determination to new heights.[186, 187] Integrative modeling approaches combine data from multiple experimental sources, such as cryo-EM, X-ray crystallography, and cross-linking mass spectrometry to revolutionize our ability to model complex structures with unprecedented accuracy and completeness. In this Chapter, we delve into our proposed integrative modeling approach to determine the 3D structures of protein complexes, while using the chromatin remodeling INO80 complex as a case example.

Chromatin remodeling complexes represent an intricate network of molecular machinery (Figure 4.1) essential for regulating gene expression, DNA repair, and other fundamental cellular processes [36, 183]. Found within the nucleus of eukaryotic cells, chromatin, the complex of DNA and proteins that make up chromosomes, serves as the architectural framework for DNA organization, consisting of DNA wrapped around histone proteins [181]. However, the dynamic nature of chromatin necessitates mechanisms that allow access to DNA for various cellular activities [188, 189]. Therefore, chromatin remodeling complexes play a pivotal role in modulating chromatin structure by altering nucleosome positioning, histone modification patterns, and DNA accessibility [36]. These complexes are comprised of multiple protein subunits that work synergistically to modify chromatin architecture (Figure 4.1). Through ATP-dependent processes, they exert their effects by disrupting histone-DNA interactions, facilitating nucleosome sliding, or evicting histones altogether [183].

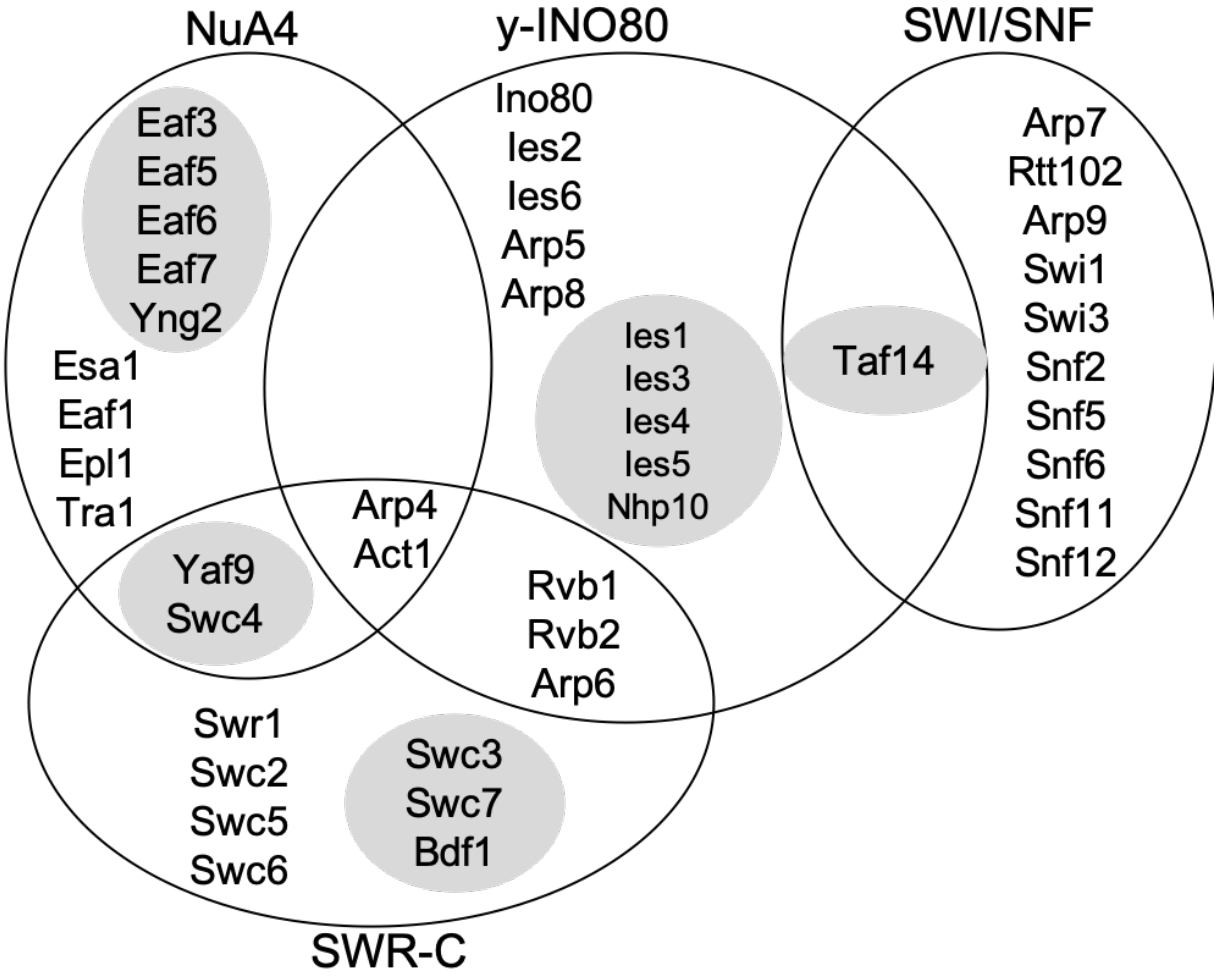

Figure 4.1. Different families of the yeast chromatin remodeling complexes.

The sizes of the remodelers vary from a single subunit, CHD (chromodomain helicase DNA-binding) to over 15 subunits, INO80 (Inositol requiring 80) [36, 181]. Although the functions of the full-length complexes have been studied intensely over the past decade and are now known to some extend [36, 37, 190], the contribution of individual subunit remains elusive. This is due in part to the lack of functional 3D structures of the complexes, because of their high flexibility [35,

188, 191]. This Chapter introduces an integrative modeling approach to determine the 3D structure of the macromolecular complexes, while using INO80 complex as a case study. It provides a step-by-step algorithm (see appendix B3 for the tutorial) on how the subunits of the macromolecular complexes assemble into functional complexes, which are intimately related to their functions.

Remodelers are commonly grouped based on sequence similarity into four families that are highly conserved throughout evolution [36, 37]: (1) SWI/SNF (switch/sucrose nonfermentable) remodelers, which include SWI/SNF and RSC, are composed of 8 to 14 subunits [192]. They establish the nucleosome-depleted regions and position the +1 nucleosome for transcription initiation. (2) ISWI (imitation switch) remodelers, which include ISW1a, ISW1b, and ISW2, are composed of 2 to 4 subunits. (3) CHD remodelers are composed of 1 to 10 subunits. ISWI and CHD families are involved in nucleosome maturation and spacing (they are involved in creating nucleosomal arrays with fixed distances) [37, 191, 193]. (4) INO80 remodelers, which include INO80 and SWR1, are composed of over 10 subunits. They are involved in nucleosome editing by exchanging histone variants [189, 190, 194-200]. However, how these highly complex systems are organized in the crowded environment like cell and their mechanisms of action remain debatable.

Multiple subunits, including acting related protein (ARPs), Rvb1, and Rvb2 are often conserved across complexes (see Figure 4.1) [191, 201-204]; however, many questions remain unanswered. For example, how do the shared subunits interact with different complexes? Simultaneous interactions would imply scaffold subunits where each interacting complex has a specific binding site. As a result, such subunits could serve as bridges that allow intercommunication between two or more complexes. On the other hand, there could be some binding competition among complexes

that share the same binding sites. Therefore, a process that explains the binding cascade among subunits of each complex will further shed light on the assembly and regulatory mechanisms of the remodelers.

In this study, we will focus on the yeast INO80 complex.[201] The INO80 complex, named after its founding member, the yeast INO80 ATPase, is an evolutionarily conserved macromolecular assembly found in eukaryotic cells. Its multifaceted functions encompass nucleosome sliding, histone variant exchange, and DNA resection, making it an indispensable player in various cellular processes such as DNA replication, DNA repair, and transcriptional regulation (Figure 4.3) [183]. However, elucidating the structural blueprint of this complex has remained a formidable challenge due to its sheer complexity and dynamic nature. INO80 complex, unlike other remodelers, has a conserved core of 8 subunits between human and yeast organisms (Figure 4.2 A), with many other subunits specific to each organism with still undefined functions [190, 205-209]. We hypothesize that INO80 subunits may pre-assemble into modules which are further assembled into the functional complex [188, 201]. To test this hypothesis, we will first study the assembly of individual module, which will provide further details on the binding sites and functions of the individual subunit. Yeast INO80 (y-INO80 or simply INO80) has been studied intensely and salient results are present in the literature [184, 189, 190, 197, 201, 206, 210-212]. Crosslinking mass spectrometry data show that the full-size INO80 complex is organized into 4 modules [201, 211] (Figure 4.2 B). This structural organization facilitates the complex assembly and/or regulatory mechanism. However, to the best of our knowledge, besides the mass spectrometry crosslinking data that show the connectivity among subunits [201], a complete 3D structure of the functional INO80 complex is still lacking. This has been hampered by the lack of experimental

and/or computational approaches capable of capturing transient interactions among subunits, due in part, to their high mobility.

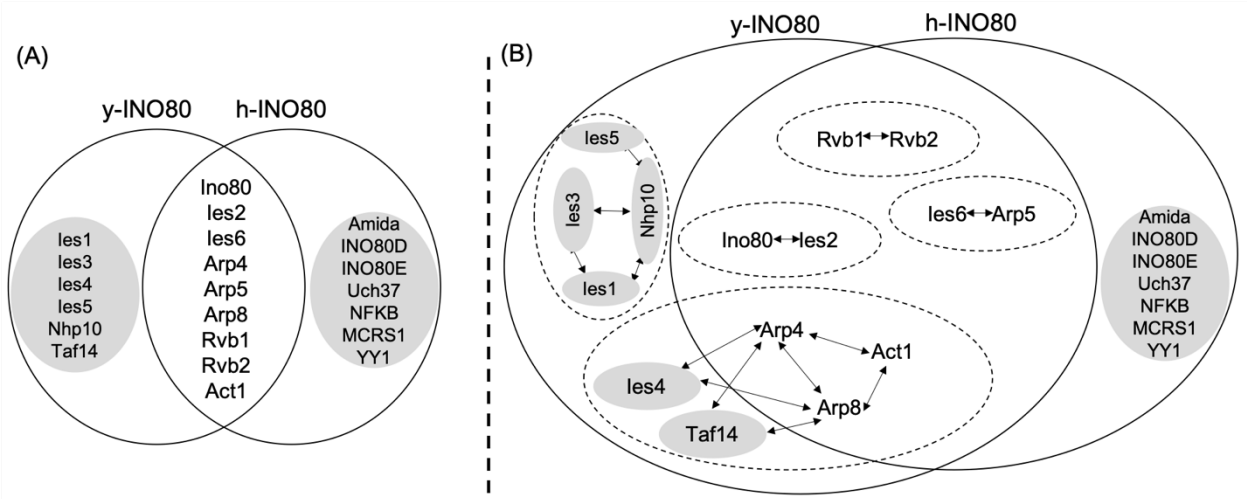

Figure 4.2. INO80 subunits organization. Yeast and human share over 8 core subunits (A) which are grouped into functional modules (B).

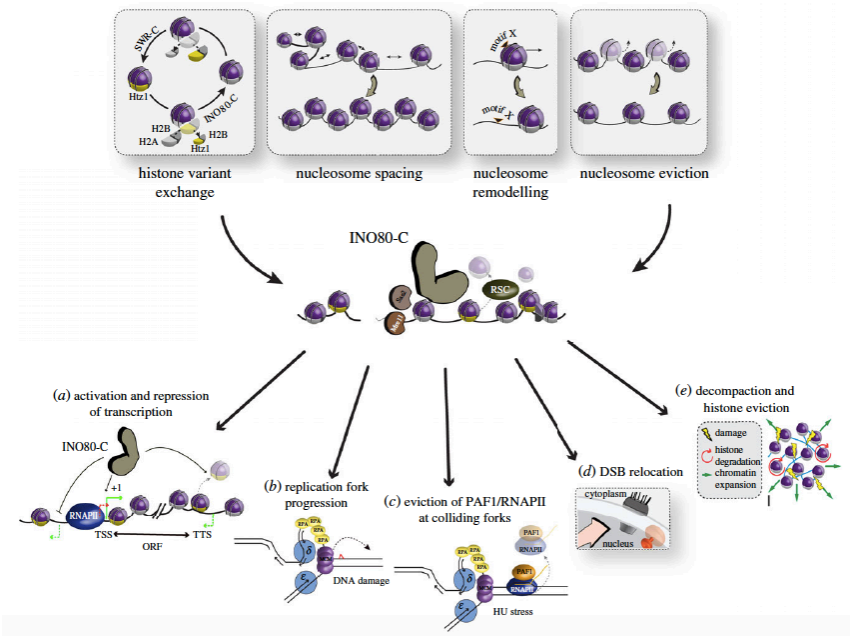

Figure 4.3. Functions of the INO80 complex. INO80 complex participate to cellular functions including transcription, replication, and DNA repair. [183]

There has been a considerable effort to determine the functional 3D structure of the INO80 complex using available experimental techniques such as cryo-electron microscopy and cross-linking mass spectrometry (see Table 4.1) [189, 196, 201, 202, 206, 208, 210]. One of the attempts shows that the y-INO80 complex is organized into 4 modules: head, neck, body, and foot [201]. The head, which is composed of subunits Rvb1 and Rvb2 (Rvb1/2 module); the neck with subunits Arp5 and Ies6 (ARP5 module); the body with subunits Nhp10, Ies1, Ies3, and Ies5 (NHP10 module); the foot with subunits Arp8, Ies4, Taf14, Act, and Arp4 (ARP8 module). These modules interact with the core subunits Ino80 and Ies2, such as shown by the crosslinking data (Figure 4.2 and Figure 4.4), where the Ino80 subunit serves as docking sites for the rest of the subunits in the complex. Furthermore, the N-terminal of the catalytic Ino80 subunit folds back to interact with the NHP10 module, located in the body of the complex. However, this configuration has been debated by many groups. They proposed that the evolutionary conserved Ino80 subunit across species would imply a more elongated configuration at the N-terminal domain, such as observed in the cryo-EM structure of the human INO80 core complex [189, 196, 211]. As a result, the NHP10 module would be located at the foot of the complex instead of the body. These controversies come from the fact that the INO80 complex is highly dynamic with modules that show high mobility, making it difficult to use the available experimental techniques to solve its full-length functional 3D structure. The integrative modeling approach proposed in this Chapter utilizes the known information from the experiment, machine learning, and molecular simulations to determine and characterize the functional 3D structure of the y-INO80 (Figure 4.5). This will be possible by first building the 3D structure of the functional modules, which are then assembled into a final complex.

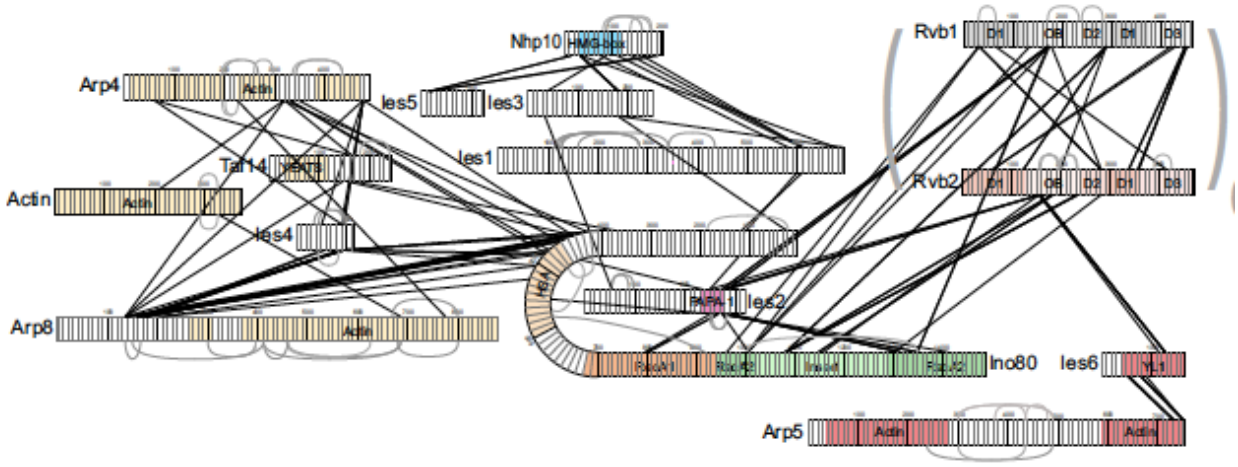

Figure 4.4. Crosslinking mass spectrometry shows the structure and subunit topology of the INO80 chromatin remodeler. Subunits are organized into modules. [201]

Integrative approaches to study protein complexes have become increasingly popular in structural biology, as they combine the strengths of different experimental and computational techniques to obtain a more comprehensive and accurate structure of the complex [182, 186, 187]. It has been used to study a wide range of biomolecular complexes, including protein-protein interactions and membrane proteins [182]. Computational approaches combine data from different experimental techniques to generate a high-resolution 3D structural model of the complex [186]. These approaches are particularly useful for large, dynamic, of flexible complexes that are difficult to study using a single experimental technique. Indeed, it relies on the complementary information provided by different experimental techniques, such as X-ray crystallography, NMR spectroscopy, and cryo-EM. Each technique provides different types of information about the complex, such as the overall shape, subunit interactions, and local structural details. Due to the high flexibility of the INO80 complex as well as that of its subunits, the determination of the high-resolution 3D

structural model of the full-length complex is very challenging with current experimental techniques.

The 3D structure of the yeast INO80 core complex present in the literature is composed of the heterododecamer Rvb1/2 and the Arp5-Ies6 subunits that interact with the insertion region within the ATPase domain of the Ino80 catalytic subunit, and Ies2 [206, 208, 211]. In addition, the 3D structure of the incomplete ARP8 module, which contains subunits Arp8, Act, and Arp4 as well as the HSA domain of the Ino80 subunit has been solved and is present in the protein data bank (PDB) under the identification 5NBN [206]. This incomplete structure of the ARP8 module is certainly due to the high mobility of its subunits, where the missing subunits, including Taf14 and Ies4, are washed out during the refinement process. On the other hand, the N-domain of the Ino80 subunit, which is an intrinsically disordered region, is often removed from the study, which explains the absence of the NHP10 module in the complex [211]. How the NHP10 module interacts with the Ino80 core subunits and its contribution to the chromatin remodeling activities of the INO80 complex remains unknown. This is due not only to the intrinsically disordered contain of the Ino80 N-domain but also to the high mobility of the subunits present in the NHP10 module [201]. It has been proposed that due to the conserved nature of the INO80 complex across species, the NHP10 module in yeast and its human analog, the AMIDA module, may have at least a similar shape [189], despite the differences in the sequence and composition of each module, but no evidence is present in the literature to support such claim. In this study, we will use the proposed integrative modeling approach to build and characterize the 3D structure of the NHP10 module, which will hopefully provide inside into that of AMIDA module necessary to fully characterize the human INO80 complex.

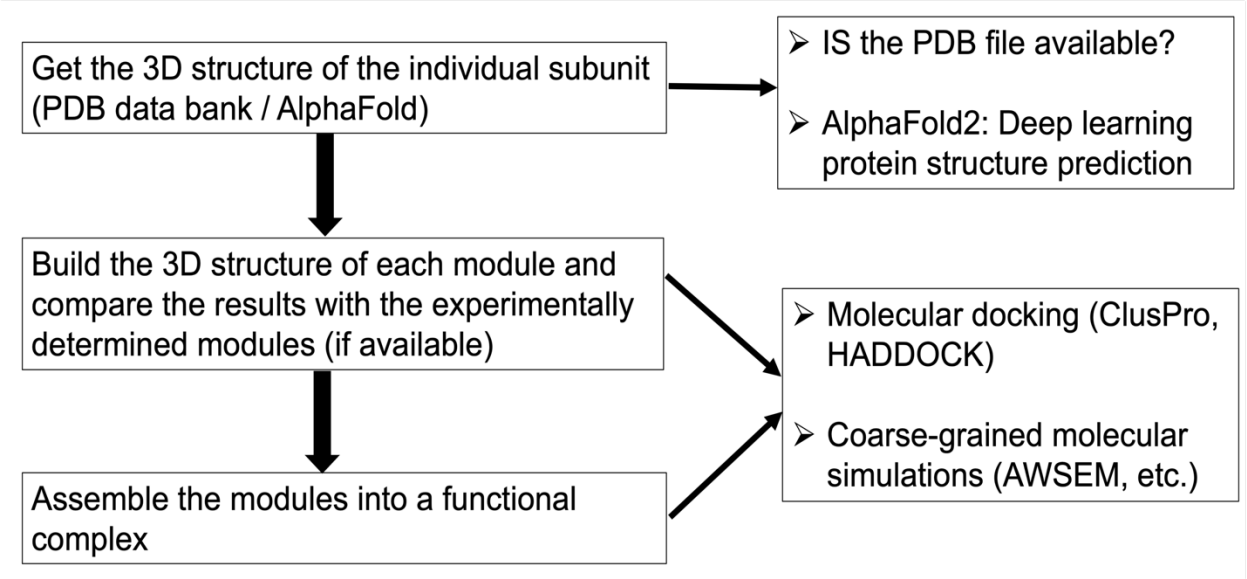

Figure 4.5. Workflow of the proposed integrative modeling approach developed in this study.

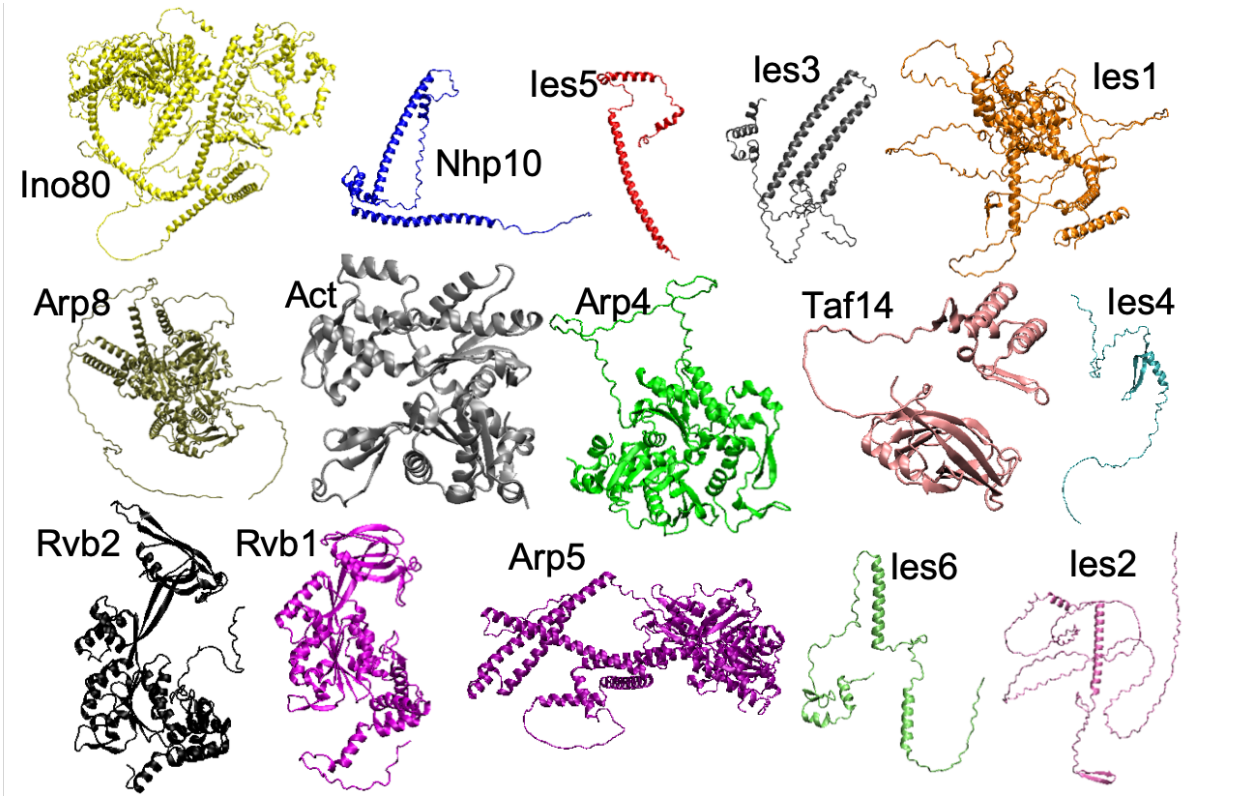

Figure 4.6. Different subunits of the yeast INO80 complex. These structures were predicted using AlphaFold2 and visualized with VMD.

Table 4.1. Subunits of the INO80 complex

Prot. Name (yeast)	Gene info.	Seq.	PDB ID	Function	Prot. Name (human)	Gene info.	Seq.	PDB ID	Function
Ino80	6321289	1489	5jxr		Ino80	33469139	1556		
Arp8	2492678	881	4am6		Arp8	39812115	624		
Arp5	1730738	755			Arp5	31542680	607		
Arp4	728794	489			Arp4	30089997	387		
Rvb1	6320396	463			Rvb1	4506753	456		
Rvb2	6325021	471			Rvb2	5730023	463		
Ies2	6324114	320			Ies2	13775202	356		
Ies6	6320791	166			Ies6	34916002	192		
Taf14	461510	244			Amida	7019371	253		
Nph10	6320202	203			Ino80E	27734727	244		
Ies1	14318508	692			NFRKB	23346420	1324		
Ies3	6323081	250			MCRS1	2384717	462		
Ies4	6324763	116			Ino80D	38488718	1027		
Ies5	6320938	125							
Act1					Act1				

4.2 DATA COLLECTION

4.2.1 *Experimental Data*

We downloaded the available experimentally solved 3-dimentional (3D) structures of the subunits of the IN80 complex from the protein databank (PDB), which includes data obtained with the following experimental techniques: Cryo-EM, X-ray crystallography, and NMR. The structures were either individual subunits or complexes (Table 4.1**)**. By comparing the fasta files of the downloaded structures with the ones obtained from the National Center for Biotechnology Information (NCBI), we found out that the coverage was extremely low for the individual and/or complex molecules with known structures. These structures will only serve as memories to guide the molecular simulations.

4.2.2 *Additional Data*

We used AlphaFold2 [213] to predict the structures of individual subunits of the INO80 complex from the sequence of amino acids downloaded from the NCBI database (Figure 4.6). Alternatively, the predicted structures of most individual subunits are now available online in the AlphaFold2 databank. Although AlphaFold2 offers the option to predict the structure of dimer using AlphaFold2-multimer [185], the accuracy of the prediction still depends on the amino acid sequence length. As a result, using AlphaFold2-multimer to predict the structure of complexes with the amino acid sequence length of 1500 or more provides structures, if possible, with a very low score. We further used the published data of the INO80 complex obtained using the XL-MS to guide and validate the 3D structure determination.[201] These XL-MS data show the pairwise interactions among subunits as well as the different modules of the INO80 complex, which are both useful to design the integrative modeling approach.

4.2.3 *Data Preprocessing*

The first part of the data preprocessing consisted in identifying all the PDBs of the subunits of the INO80 complex that have been experimentally solved and to annotate the coverage. For each subunit, we identified the regions that are present in the PDBs, which are specified in the memory file. These details about the coverage of each subunit will be used in the memory file to specify the regions of each subunit that have been solved to guide the simulations.

4.3 STRUCTURAL MODELING

4.3.1 *Initial Models*

After predicting the structures of individual subunits (structure with high score) of the INO80 complex with AlphaFold2, we then used ClusPro[214] to assemble the individual subunits into modules and HADDOCK [215] to assemble the modules into the final complex (see Figure 4.5 and more details below). The assembly was guided by the XL-MS data. [201]

4.3.2 *Integrative Modeling*

XL-MS data show that INO80 complex has over 3 functional modules. [201] Because of the complexity of the system and the energy that may be required to assemble all the subunits at once while respecting the modularity, we suggest that modules would first form before they are assembled into the final functional complex.

We utilized ClusPro to assemble the individual subunits of the INO80 complex into modules based on the following procedure: for a module with A, B, C, and D subunits, (1) If A and B interact with each other, based on the XL-MS data, we assembled them to form a complex AB; (2) If A or B or both A and B interact with C, we assembled them to form a complex ABC or CAB; the similar procedure is done for D to form either DABC or CABD. For each assembly step, we kept the structure from the highly populated cluster with the highest score. The chosen cluster is further checked by drawing the contact map that shows the connectivity among subunits, which, compared to the map of other clusters, is closed to the XL-MS map. This procedure can lead to some controversies. For example, the order by which subunits are assembled, since ABC and CAB will lead to different structures with different functions. To overcome these controversies, we conducted the coarse-grained molecular simulations on the chosen complex. Because our

simulations are guided by the experimentally solved structures and are long enough to overcome any barriers or any local minima, the final structure does not depend on any assembly order and the contact map resembled that of XL-MS data.

We further utilized HADDOCK to assemble the final frames from the molecular simulations of individual module to form the template of the INO80 complex, which is used as the initial step for the molecular simulations on the full-length complex. We kept the structure with the highest score from the highly populated cluster and utilized it to initiate the molecular simulations.

We conducted the coarse-grained molecular simulations using the associative memory, Water mediated, Structure and Energy Model (AWSEM) force field which is implemented in the open-source MD package Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS). We followed the same procedure such as described in the original AWSEM paper and our previous paper. [14, 22, 43] The initial systems for the simulations were composed of the final structures of each module as well as the full-length INO80 complex, obtained either with ClusPro or HADDOCK, respectively. We placed each system in the simulation box at the temperature of 300 K, with an initial velocity randomly assigned using the Boltzmann distribution ($3k_BT/m$), where k_B is the Boltzmann constant, m is the mass of C-alpha atoms, and T is the temperature in Kelvin. We used the experimentally solved structures of each subunit as constraints (aka associative memories) to guide the configurational search of the global minimum during the simulations. Therefore, we utilized a total of 64 different structures from X-ray crystallography, NMR, and Cryo-EM. For the subunits with unsolved structures, we used the structures predicted with AlphaFold2 as memories. We used the cutoff distance of 3.5 Å and κ (inverse of the Debye length) = 1.4 $Å^{-1}$ in the electrostatic term. We conducted the simulations in a canonical ensemble (NVT),

where we kept the integration time step at 2 fs and coordinates were recorded every 1000-time steps for further analyses. For the full-length complex, we further used the crosslinking information to constrain the subunits with known interactions to be in proximity, which facilitated the configurational search of the binding interfaces, thus, leading to fast convergence.

4.4 Model Validation

There are numerous approaches to check the accuracy of the final structure after the molecular simulations step, but they mostly rely on the known initial structure for the template matching, or they fit the final structure into the Cryo-EM/Cryo-ET density. Because the 3D structure of the INO80 is still unknown due to its high flexibility, we validated the final structure by comparing the contact maps that show the connectivity among subunits from the simulations with the one obtained from XL-MS data. We further used the variation in the radius of gyration, which becomes almost constant, to assess the convergence and/or stability of the system. Another approach, which we did not explore here, is to use the change in the TM-score ($\Delta_{\text{TM-score}}$). TM-score is a quantitative assessment of similarity between protein structures. It has value in (0,1], where 1 indicates a perfect match between two structures. $\Delta_{\text{TM-score}}$ is the difference between the TM-score of the current and the previous frames during the molecular simulations. When $\Delta_{\text{TM-score}}$ becomes extremely low, as compared to a predefined threshold, the simulations stop.

4.5 Model Visualization

We visualized the 3D structures using the Visual Molecular Dynamics (VMD) [52] software package. Contact maps and other plots were made using the python scripts, which are provided in the appendix B3.

4.6 MODELING THE 3D STRUCTURE OF THE INO80 COMPLEX

INO80 remodeler is a large complex comprising 15 unique subunits in yeast (Figure 4.6) with a combined molecular weight of 1.3 MDa [201] that has a core of 8 subunits conserved from yeast to human (Figure 4.2). Subunits of the INO80 complex are grouped into four independent subcomplexes, which are named according to the subunits that are required for the recruitment of other subunits into the complex, such as observed experimentally during the deletion study ([201]). Ino80 subunit serves as a platform that coordinates the assembly of the four functional subcomplexes: NPH10, ARP8, ARP5, and RVB1_2. Although the structures and functions of the ARP8, ARP5, and RVB1_2 subcomplexes are more of less known from the literature,[184] the higher flexibility as well as the limited structural information impeded the understanding of the NHP10 module assembly and its involvement in the chromatin remodeling activities of the INO80 complex. We first utilized the integrative modeling approach to provide insights into the subunit's organization of the NHP10 subcomplex as well as other subcomplexes of the of the INO80 complex.

4.7 ASSEMBLY OF THE NHP10 MODULE

NHP10 subcomplex is composed of four unique subunits Nhp10, Ies1, Ies3, and Ies5 where the 3D structures predicted using AlphaFold2 are shown in Figure 4.7. We used these individual structures to build the initial template of the NHP10 subcomplex. How subunits are recruited into the subcomplex and the order with which they are recruited can dictate its shape. To overcome these differences, we can either perform a deletion study to determine the correct recruitment order

or we can run longer simulations on the initial template, which will relax the system and probably converge to the same final structure despite different initial templates.

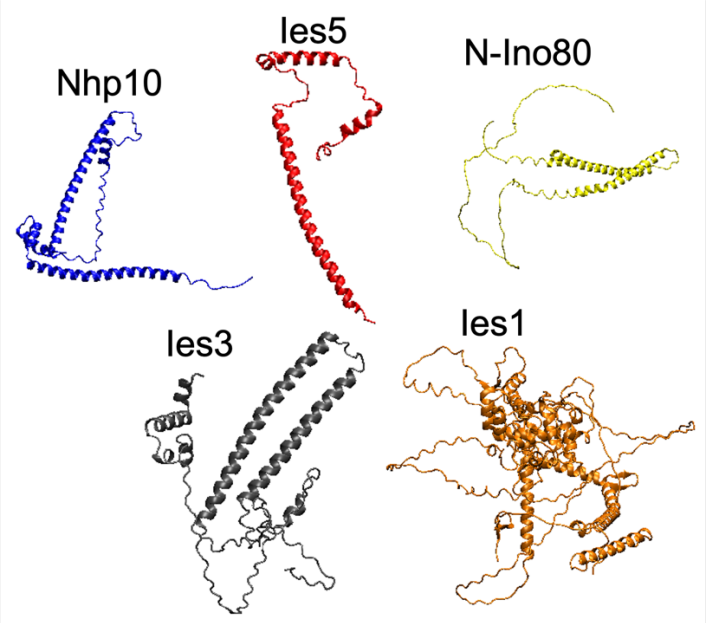

Figure 4.7. predicted structures of the subunits of the NHP10 subcomplex using AlphaFold2.

Nhp10 subunit constitutes the assembly scaffold of the NHP10 subcomplex, making distinct physical interactions with the other three subunits of the subcomplex (Figure 4.8). [201] The subcomplex is recruited into the INO80 complex through the interactions between the N-domain of the Ino80 subunit and the C-termini of the Ies1 and Ies3 subunits (Figure 4.8 and Figure 4.9). Additionally, the crosslinking data show the interactions between the N-terminal of the Ies2 and that of the Ies3 subunits. [201]

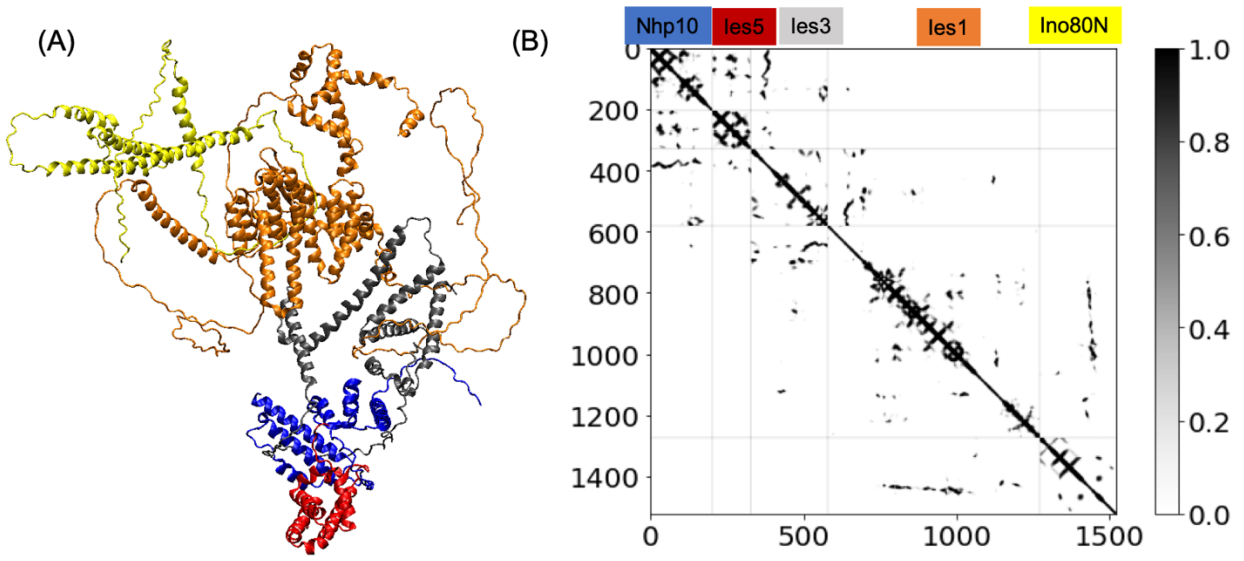

Figure 4.8. The individual structures are assembled to form the initial template of the NHP10 module using the molecular docking (A) and the associated contact map (B).

Ies5 peripherally associates with the Nhp10 subunit such as confirmed by the contact maps in Figure 4.8 and Figure 4.9. Furthermore, new contacts are observed after the coarse-grained molecular dynamics simulations (Figure 4.9). These new contacts, which were not observed in the initial template, are present in the crosslinking data, indeed, supporting our results and the crucial simulations step, which is required not only to get the equilibrated subcomplex (Figure 4.10), but also to relax the system and allows further contacts to be formed.

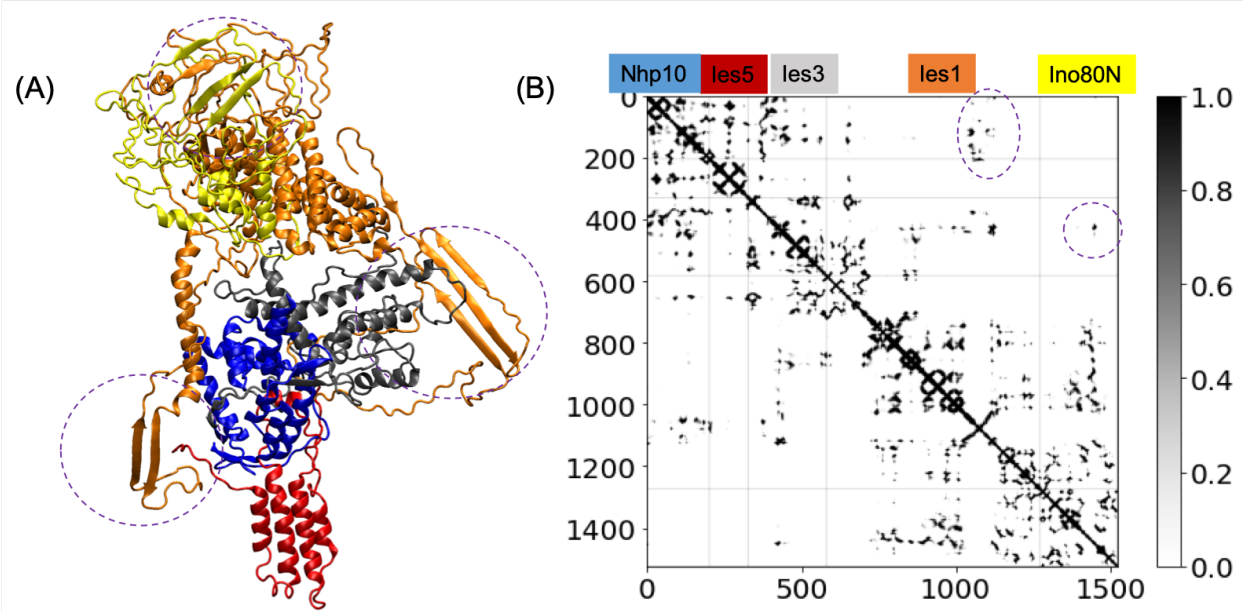

Figure 4.9. Final template of the NHP10 module after the converged coarse grained MD simulations. The contact map in (B) shows the formation of new contacts after the molecular simulations, which were not seen in the initial template before the MD simulations. These new contacts are validated experimentally. [201]

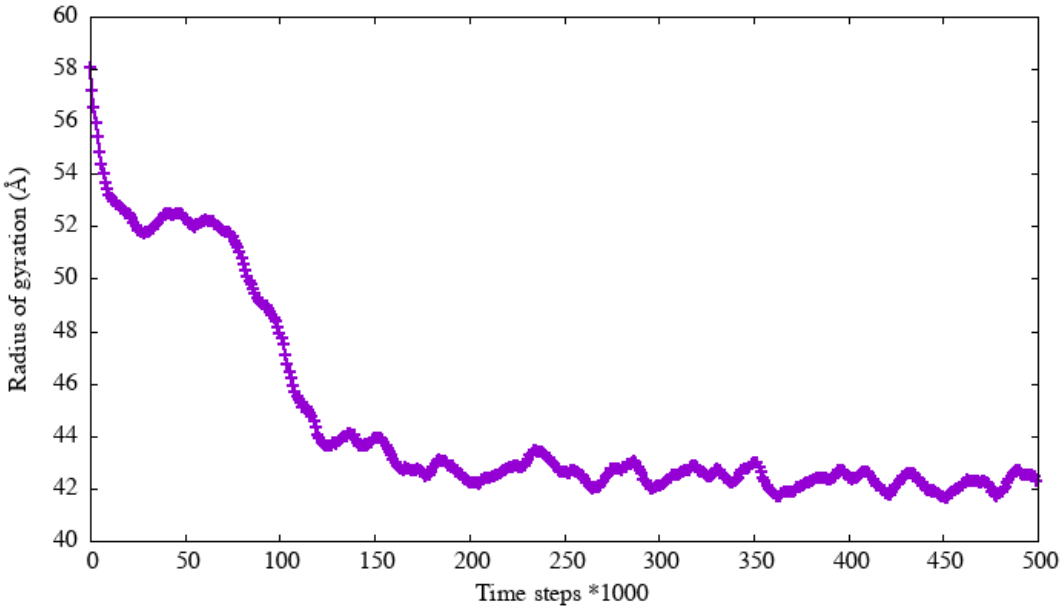

Figure 4.10. Radius of gyration variation versus time steps of the NHP10 module obtained after the coarse-grained MD simulations step.

4.8 ASSEMBLY OF THE ARP8 MODULE

Determining the three-dimensional (3D) structure of the ARP8 module within the INO80 complex is crucial for understanding its function and mechanism within chromatin remodeling. ARP8 is one of the essential modules of the INO80 complex, contributing to its overall function, and has been studied intensely over the past decade. [184]. The ARP8 module consists of the Arp8 protein along with four associated subunits, including Act, Arp4, Taf14, and Ies4, forming a distinct structural unit within the larger INO80 complex. Arp8 belongs to the actin-related protein (ARP) family and shares structural homology with actin, despite differences in their functions. The association of Arp8 with other subunits such as Arp4 forms a stable heterodimer, which contributes to the stability and functionality of the module. The 3D structures of individual subunits predicted using AlphaFold2 are shown in Figure 4.11. We used these individual structures to build the initial template of the ARP8 subcomplex. We further performed the coarse-grained MD simulations on the initial template, which relaxed the system to converge to a conformation close to the global one.

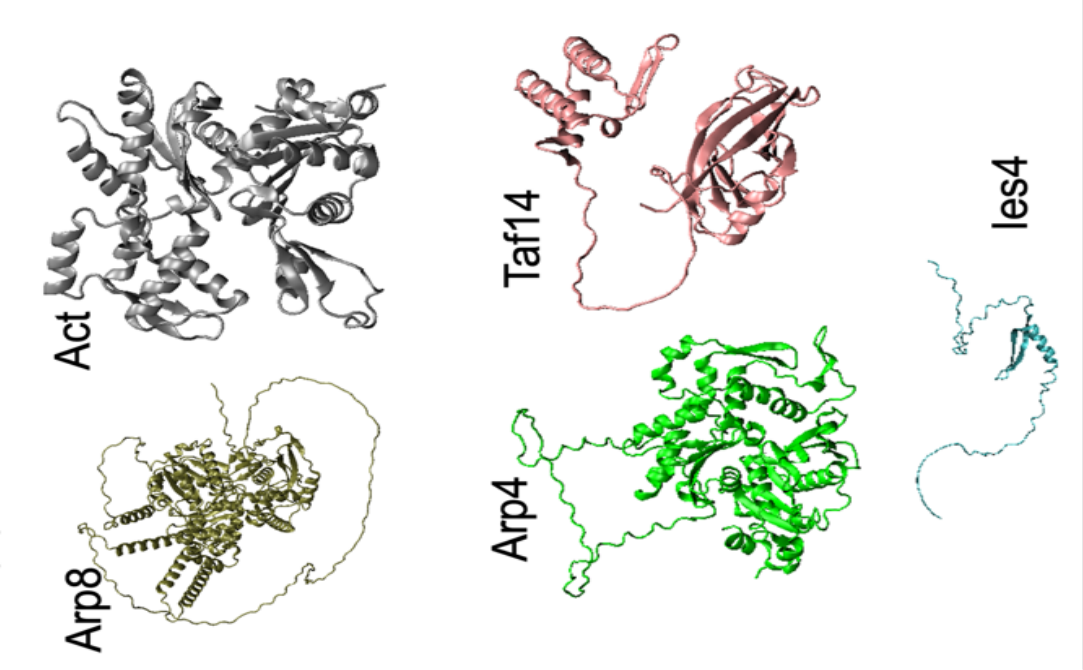

Figure 4.11. Predicted structures of the subunits of the ARP8 subcomplex using AlphaFold2.

Arp8 subunit constitutes the assembly scaffold of the ARP8 subcomplex, making distinct physical interactions with the other four subunits of the subcomplex (Figure 4.12). [201] The subcomplex is recruited into the INO80 complex through the interactions among the HSA domain of the Ino80 subunit and the C-termini of the Arp8 and Arp4 subunits (Figure 4.12 and Figure 4.13).

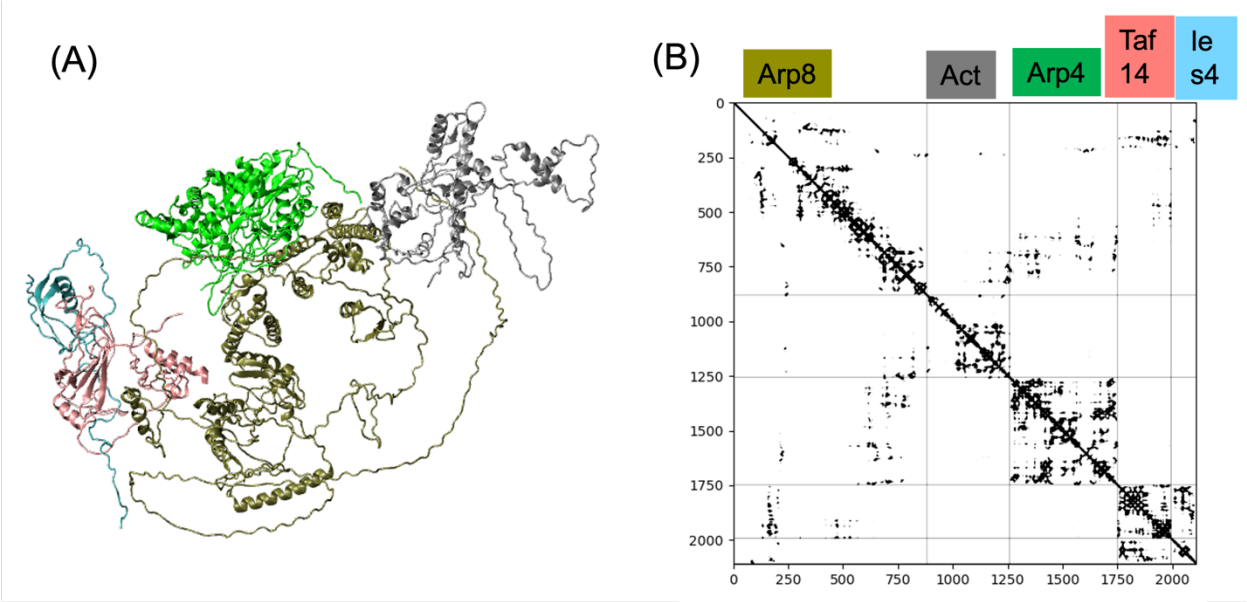

Figure 4.12. The individual structures are assembled to form the initial template using the molecular docking, 3D-structure (A) and the associated contact map (B)

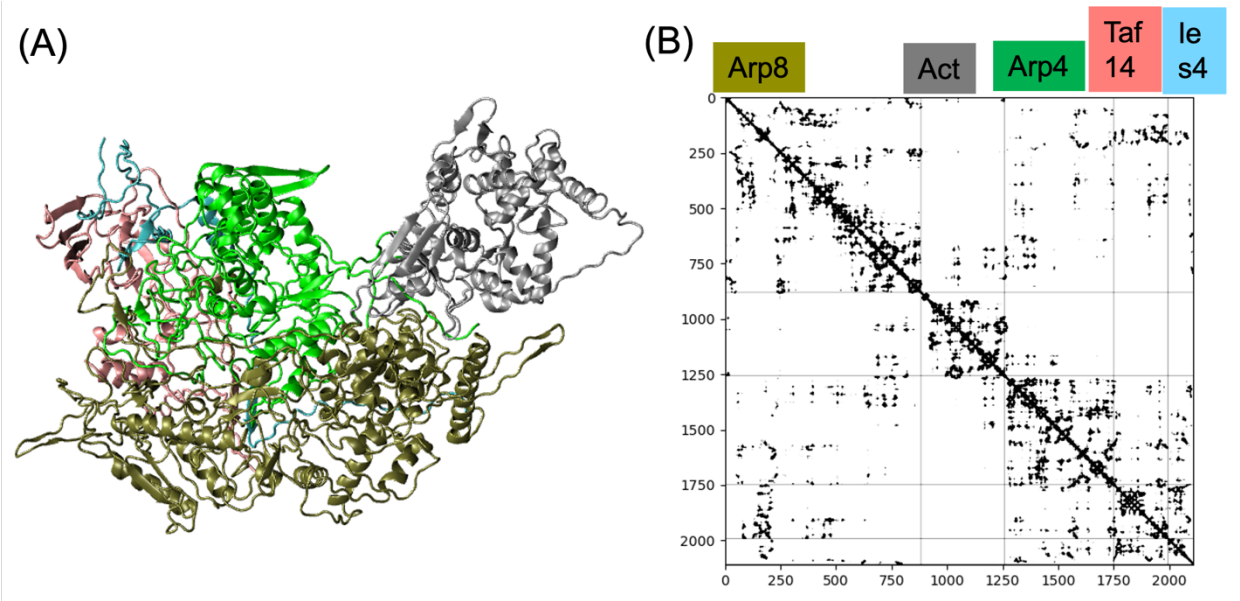

Figure 4.13. The initial template is subjected to the coarse-grained molecular dynamics simulations, 3D-structure (A) and the associated contact map (B)

Overall, determining the 3D structure of the ARP8 module within the INO80 complex requires a multidisciplinary approach, combining experimental techniques such as X-ray crystallography, Cryo-EM, and XL-MS with computational methods like molecular dynamics simulations, homology modeling, and integrative structural biology. These efforts not only elucidate the architecture and function of ARP8 but also contribute to our broader understanding of chromatin remodeling and its role in gene expression and genome maintenance.

4.9 ASSEMBLY OF THE ARP5 MODULE

The ARP5 module, composed of Arp5 and Ies6, is a structural unit within the INO80 chromatin remodeling complex. It represents one of the sub-complexes within the larger INO80 complex, contributing to its overall structure and function. [201] ARP5 module interacts with the insertion region within the ATPase domain of the Ino80 catalytic subunit, and Ies2. Arp5 is a member of the actin-related protein (ARP) family, which shares structural homology with actin but exhibits distinct functions. It participates in protein-protein interactions within the complex, facilitating its assembly and recruitment to chromatin. Ies6 (INO Eighty Complex subunit 6) is another subunit associated with the ARP5 module of the INO80 complex. While the precise role of Ies6 within the ARP5 module and the INO80 complex is not fully understood, it likely contributes to the stability, assembly, or regulation of the complex's chromatin remodeling activity. As the size of the ARP5 module is relatively reduced, compared to other modules, with only 2 subunits, we were able to predict the 3D structure of the functional module using AlphaFold2-multimer. [185] The 3D structures of individual subunit is shown in Figure 4.14 A, that of the module is shown in Figure 4.14 B and the associated contact map is shown in Figure 4.14 C.

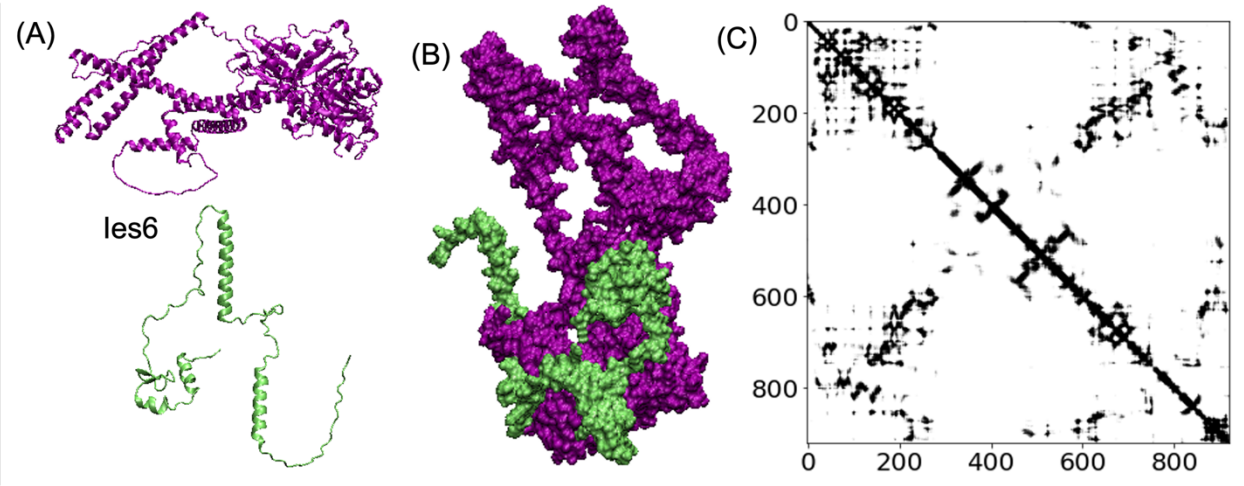

Figure 4.14. Workflow of the ARP5 module structure determination. (A) predicted structures of the subunits of the ARP5 subcomplex using AlphaFold2. (B) 3D structure of the ARP5 module and the associated contact map (C).

The ARP5 module plays a role in mediating the interaction between the INO80 complex and chromatin. By recognizing specific chromatin features, such as histone modifications or DNA sequences, the ARP5 module may target the INO80 complex to specific genomic loci for chromatin remodeling. Additionally, the ARP5 module may influence the activity or specificity of the INO80 complex, modulating its function in transcriptional regulation, DNA repair, or other chromatin-related processes. [201]

4.10 3D STRUCTURE OF THE FULL-LENGTH INO80 COMPLEX

Determining the 3D structure of the INO80 complex has been a significant endeavor in structural biology, aiming to unravel the molecular architecture underlying its diverse functions in chromatin remodeling and DNA metabolism.[184, 190, 195] The INO80 complex is a large multi-subunit protein assembly composed of about 15 distinct subunits (Figure 4.6), each contributing to its

overall structure and function. To determining the 3D structure of the INO80 complex, we utilized the final frames from the MD simulations on individual module to build the initial template, which then serves as the initial structure for the final MD simulations step. The preliminary results of the full-length INO80 complex are shown in Figure 4.15 A for the 3D structural representation and Figure 4.15 B for the associated contact map. These results require longer simulations to allow the system to further relax to a structure with the lowest energy.

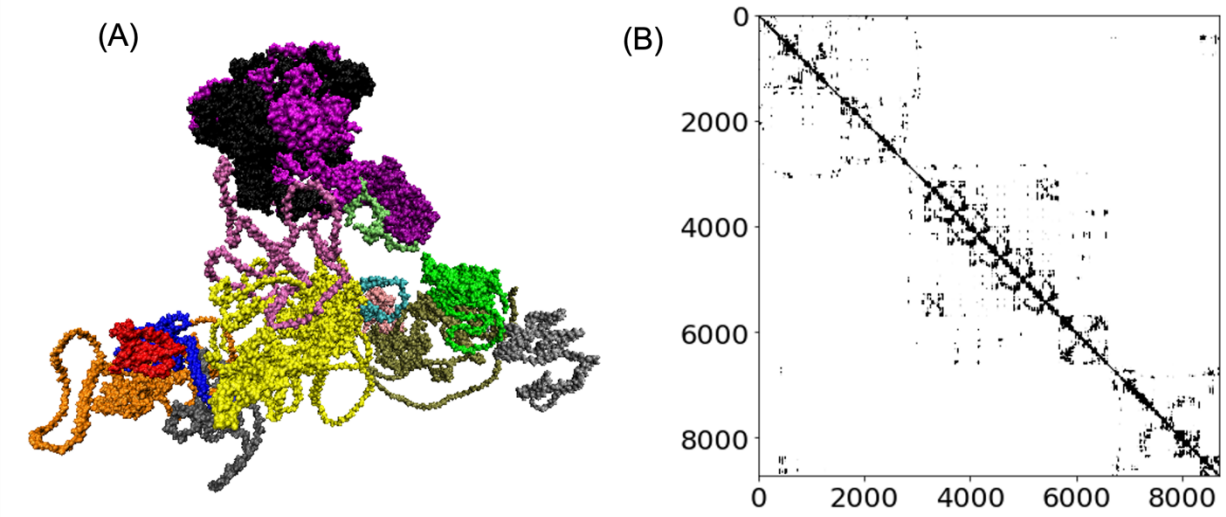

Figure 4.15. Workflow of the full-length INO80 complex 3D structure determination. (A) 3D structure of the INO80 complex (initial template for the molecular simulations) and the associated contact map (B).

By leveraging advanced experimental and computational methods, we introduced an integrative approach to determine the 3D structure of the full-length INO80 complex to unravel the molecular mechanisms underlying its chromatin remodeling activities and its roles in genome maintenance and gene regulation. Ultimately, insights gained from structural studies of the

INO80 complex deepens our understanding of chromatin biology and have implications for human health and disease.

4.11 Discussion on integrative modeling techniques to determine the 3D structures of protein complexes

Integrative modeling has emerged as a powerful approach for determining the three-dimensional (3D) structure of large and complex macromolecular assemblies like the yeast INO80 complex. [182, 186] Integrative modeling combines data from multiple experimental techniques, computational modeling, and bioinformatics to generate accurate structural models that provide insights into the architecture, dynamics, and function of the complex. In this Chapter, we introduced a step-by-step workflow on how integrative modeling has been utilized to elucidate the 3D structure of the yeast INO80 complex (Figure 4.15).

Integrative modeling begins by gathering diverse experimental data that capture different aspects of the INO80 complex structure and dynamics. This includes cryo-electron microscopy data providing low-resolution density maps, X-ray crystallography data offering atomic-level details of individual subunits or domains, chemical crosslinking mass spectrometry data revealing spatial proximity between residues, and other biochemical and biophysical assays characterizing protein-protein interactions, subunit composition, and conformational changes. Computational modeling techniques are then employed to integrate and interpret the experimental data and generate structural models of the INO80 complex. This involves molecular dynamics simulations to explore the dynamics and flexibility of the complex, homology modeling or machine learning to predict the structures of subunits or domains lacking experimental structures, and docking simulations to assemble individual structures into modules or complexes.

The structural models generated through integrative modeling provide valuable insights into the organization, assembly, and function of the protein complexes. They reveal the spatial arrangement of subunits within the complex, identify interaction interfaces and allosteric sites, and elucidate conformational changes associated with complex activity and regulation. [186, 187, 209] These structural insights deepen our understanding of the molecular mechanisms underlying chromatin remodeling, DNA repair, and transcriptional regulation mediated by the INO80 complex.

The use of integrative modeling to determine the 3D structures of protein complexes will continue to advance, driven by ongoing improvements in experimental techniques, computational algorithms, and data integration strategies. Future studies may aim to refine the structural models at higher resolutions, investigate dynamic transitions between different functional states of the complex, and explore its interactions with chromatin and regulatory factors in physiological contexts. Additionally, this integrative modeling approaches can be applied to study the structural and functional diversity of INO80 complexes across different organisms and biological systems as well as other macromolecular complexes.

4.12 CONCLUSION

Protein complexes represent fundamental entities in cellular biology, orchestrating a myriad of crucial functions essential for life. Through intricate interactions and precise arrangements, these complexes govern processes ranging from gene regulation to cellular signaling, ensuring proper cellular function and organismal development. The dynamic nature of protein complexes, their adaptability to environmental cues, and their involvement in various diseases underscore their significance. The determination of their three-dimensional structures represents a cornerstone in understanding the molecular basis of cellular processes and disease mechanisms. Through a combination of experimentally guided computational techniques, we have unveiled the intricate

architectures and dynamics interactions within the macromolecular assemblies, utilizing the chromatin remodeling INO80 protein complex as a case study. INO80 protein complex stands as a pivotal player in chromatin remodeling and genome maintenance, where its multifaceted roles in DNA repair, replication, transcriptional regulation, and nucleosome dynamics underscore its significance in orchestrating fundamental cellular processes, essential for genome stability and cellular homeostasis. Through its intricate assembly and coordinated activities, the INO80 complex demonstrates remarkable versatility, adapting to diverse cellular contexts and environmental stimuli. The structural insights provide invaluable knowledge regarding the functional mechanisms, substrate specificity, and regulatory networks of protein complexes, paving the way for rational drug design, therapeutic targeting, and the development of novel biotechnological applications.

Chapter 5. CONCLUSION AND FUTURE DIRECTIONS

In summary, proteins are the molecular workhorses of the cell, with diverse structures and functions crucial for the maintenance of life.[216] The dynamic nature of proteins, encompassing their structural flexibility and conformational changes, is fundamental to their biological activities.[217, 218] Protein-protein interactions play a pivotal role in orchestrating cellular processes, allowing proteins to collaborate in intricate networks and signaling pathways. The formation of protein complexes further highlights the cooperative nature of proteins, as they assemble into functional units that drive essential cellular functions.[216]

Proteins play a fundamental role in the structure and function of living organisms. Calmodulin, a highly conserved and ubiquitous calcium binding protein, serves as a crucial mediator in numerous cellular processes by modulating the activity of various target proteins, which in turn, affect the affinity of calmodulin for calcium.[14, 22] The dynamic nature of proteins, including calmodulin, is essential for their functionality.[219] The study of proteins, their dynamics, interactions, and complexes, represents a cornerstone in modern molecular biology, biophysics, and drug discovery. Advances in experimental techniques, such as X-ray crystallography, NMR spectroscopy, mass spectrometry, and cryo-electron microscopy, have enabled researchers to delve into the intricate details of protein structures and their dynamic behavior.[8] Understanding protein dynamics and interactions is not merely an academic pursuit; it has profound implications for human health and disease. Dysregulation of protein-protein interactions or disruptions in protein complexes are associated with numerous diseases, providing targets for therapeutic interventions. Drug development strategies that focus on modulating these interactions offer promising avenues for treating conditions ranging from cancer to neurodegenerative disorders.

In essence, the exploration of proteins and their molecular ballet—whether at the level of individual structures, dynamic motions, or intricate interactions—provides a comprehensive view of the inner workings of cells. This knowledge is foundational for deciphering the complexities of life and holds the key to developing innovative treatments for a myriad of human ailments.

Machine learning (ML) has become an invaluable tool in the study of proteins, offering new perspectives and accelerating research in areas such as protein structure prediction, analysis of protein dynamics, prediction of protein-protein interactions, and understanding protein complexes.[185, 220, 221] Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been employed to predict protein structures. AlphaFold, developed by DeepMind, is a notable example that uses deep learning to predict the 3D structure of proteins with remarkable accuracy.[222-224] ML algorithms can learn features from amino acid sequences and their evolutionary information to predict tertiary and quaternary protein structures. ML algorithms are used to analyze and interpret the massive datasets generated by molecular dynamics simulations. They can help identify key dynamic features and characterize protein motion more efficiently.[224] ML models can predict conformational changes in proteins based on experimental or simulated data, aiding in the understanding of how proteins transition between different functional states.[225] Its algorithms leverage various features, including sequence information, structural data, and evolutionary information, to predict potential protein-protein interactions. Support Vector Machines (SVMs), random forests, and deep learning architectures are commonly used for this purpose.[225] ML is employed to analyze and predict the architecture of protein interaction networks, providing insights into the organization and dynamics of cellular processes. ML models predict protein complexes by integrating diverse data types, such as interaction networks, structural information, and functional annotations. Clustering algorithms

and deep learning approaches are used to identify protein complexes from large-scale datasets. ML is employed to annotate the functions of proteins within complexes and predict the roles of individual proteins in specific cellular processes.[42, 226, 227]

Studying protein structures, dynamics, interactions, and complexes using science-informed machine learning is a cutting-edge and interdisciplinary approach that combines principles from bioinformatics, computational biology, and machine learning. Extract relevant features from protein structures and dynamics data.[227] This may include amino acid composition, physicochemical properties, structural motifs, solvent accessibility, and dynamic properties. Use domain knowledge to define features that capture the unique characteristics of calmodulin and its interacting partners.[226] Train machine learning models (e.g., random forests, support vector machines, or deep learning models) to predict whether a given pair of proteins interact based on their features.[226] Utilize positive and negative datasets derived from known calmodulin interactions and non-interacting protein pairs. Utilize information on individual protein-protein interactions, structural features, and dynamics to predict the stability and likelihood of complex formation. Combine machine learning predictions with experimental data to provide a comprehensive understanding of calmodulin and its interactions. Validate predictions with experiments such as X-ray crystallography, NMR spectroscopy, or other relevant techniques.

Apply the developed models to study novel interactions or predict the effects of mutations on calmodulin interactions. One can extract biological insights from the predicted outcomes and guide further experimental investigations. The success of science-informed machine learning in studying proteins relies heavily on the quality and diversity of data, feature engineering, and collaboration between computational biologists and experimental researchers.

ML models heavily depend on high-quality data. Incomplete or biased datasets can lead to inaccurate predictions. Some ML models, especially deep learning architectures, can be viewed as "black boxes," making it challenging to interpret their decisions. Efforts are made to enhance model interpretability in the context of protein research. Successful ML applications often involve integrating diverse data types, such as genomic, structural, and interaction data, to provide a comprehensive understanding of protein biology. Machine learning is revolutionizing the study of proteins by leveraging vast datasets and complex relationships within biological systems. These applications not only enhance our understanding of fundamental biological processes but also hold great promise for accelerating drug discovery and development.

BIBLIOGRAPHY

1. LaPelusa, A. and R. Kaushik, *Physiology, Proteins*, in *StatPearls*. 2024, StatPearls Publishing

Copyright © 2024, StatPearls Publishing LLC.: Treasure Island (FL).

2. Morris, R., Katrina A. Black, and Elliott J. Stollar, *Uncovering protein function: from classification to complexes.* Essays in Biochemistry, 2022. **66**(3): p. 255-285.
3. Brown, S.C.W., *DNA to mRNA to Protein.* Nature Education, 2008. **101**.
4. Paul E. Hertz, B.M.R.H.M., *Biology: The Dynamic Science*. 2016: Brooks/Cole Pub Co.
5. Reynaud, E., *Protein Misfolding and Degenerative Diseases.* Nature Education, 2010. **28**.
6. Jiang, S., et al., *Protein phase separation and its role in tumorigenesis.* eLife, 2020. **9**: p. e60264.
7. Qin, H., et al., *RNA-binding proteins in tumor progression.* Journal of Hematology & Oncology, 2020. **13**(1): p. 90.
8. Gromiha, M.M., *Chapter 6 - Protein Stability*, in *Protein Bioinformatics*, M.M. Gromiha, Editor. 2010, Academic Press: Singapore. p. 209-245.
9. Bolje, A. and S. Gobec, *Analytical Techniques for Structural Characterization of Proteins in Solid Pharmaceutical Forms: An Overview.* Pharmaceutics, 2021. **13**(4).
10. Alberts B, J.A., Lewis J, et al., *Analyzing Protein Structure and Function*. Molecular Biology of the Cell. 2002: Garland Science.
11. Clapham, D.E., *Calcium Signaling.* Cell, 2007. **131**(6): p. 1047-1058.
12. Gulbis, J.M., et al., *Structure of the C-Terminal Region of p21WAF1/CIP1 Complexed with Human PCNA.* Cell, 1996. **87**(2): p. 297-306.
13. Wang, Q., et al., *Assemblies of calcium/calmodulin-dependent kinase II with actin and their dynamic regulation by calmodulin in dendritic spines.* Proceedings of the National Academy of Sciences, 2019. **116**(38): p. 18937-18942.
14. Nde, J., et al., *Coarse-Grained Modeling and Molecular Dynamics Simulations of Ca2+-Calmodulin.* Frontiers in Molecular Biosciences, 2021. **8**(753).
15. Bayer, K.U. and H. Schulman, *CaM Kinase: Still Inspiring at 40.* Neuron, 2019. **103**(3): p. 380-394.
16. Her, C., et al., *Structural dynamics of calmodulin-ryanodine receptor interactions: electron paramagnetic resonance using stereospecific spin labels.* Scientific Reports, 2018. **8**(1): p. 10681.
17. Wang, Q., et al., *Protein recognition and selection through conformational and mutually induced fit.* Proceedings of the National Academy of Sciences, 2013. **110**(51): p. 20545-20550.
18. Wu, G., et al., *Calcium-induced changes in calmodulin structural dynamics and thermodynamics.* Int J Biol Macromol, 2012. **50**(4): p. 1011-7.
19. Hoffman, L., et al., *Conformational changes underlying calcium/calmodulin-dependent protein kinase II activation.* Embo j, 2011. **30**(7): p. 1251-62.

20. Park, H.Y., et al., *Conformational changes of calmodulin upon Ca2+ binding studied with a microfluidic mixer.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 542-7.
21. Berchtold, M.W. and A. Villalobo, *The many faces of calmodulin in cell proliferation, programmed cell death, autophagy, and cancer.* Biochim Biophys Acta, 2014. **1843**(2): p. 398-435.
22. Nde, J., et al., *Experiment and Simulation Reveal Residue Details for How Target Binding Tunes Calmodulin's Calcium-Binding Properties.* The Journal of Physical Chemistry B, 2023. **127**(13): p. 2900-2908.
23. Mruk, K., et al., *Structural insights into neuronal K⁺ channel–calmodulin complexes.* Proceedings of the National Academy of Sciences, 2012. **109**(34): p. 13579-13583.
24. Saimi, Y. and C. Kung, *Calmodulin as an Ion Channel Subunit.* Annual Review of Physiology, 2002. **64**(1): p. 289-311.
25. Lisman, J., R. Yasuda, and S. Raghavachari, *Mechanisms of CaMKII action in long-term potentiation.* Nature Reviews Neuroscience, 2012. **13**(3): p. 169-182.
26. Zhang, P., et al., *Opposing Intermolecular Tuning of Ca(2+) Affinity for Calmodulin by Neurogranin and CaMKII Peptides.* Biophysical journal, 2017. **112**(6): p. 1105-1119.
27. Waxham, M.N., A.L. Tsai, and J.A. Putkey, *A mechanism for calmodulin (CaM) trapping by CaM-kinase II defined by a family of CaM-binding peptides.* J Biol Chem, 1998. **273**(28): p. 17579-84.
28. Meyer, T., et al., *Calmodulin trapping by calcium-calmodulin-dependent protein kinase.* Science, 1992. **256**(5060): p. 1199-202.
29. Kühner, S., et al., *Proteome organization in a genome-reduced bacterium.* Science, 2009. **326**(5957): p. 1235-40.
30. Gingras, A.-C., et al., *Analysis of protein complexes using mass spectrometry.* Nature Reviews Molecular Cell Biology, 2007. **8**(8): p. 645-654.
31. Zahiri, J., et al., *Protein complex prediction: A survey.* Genomics, 2020. **112**(1): p. 174-183.
32. Calabrese, G., C. Molzahn, and T. Mayor, *Protein interaction networks in neurodegenerative diseases: From physiological function to aggregation.* Journal of Biological Chemistry, 2022. **298**(7): p. 102062.
33. Schmidt, M.F., et al., *Ubiquitin signalling in neurodegeneration: mechanisms and therapeutic opportunities.* Cell Death & Differentiation, 2021. **28**(2): p. 570-590.
34. Wu, H., et al., *Reorientation of INO80 on hexasomes reveals basis for mechanistic versatility.* Science, 2023. **381**(6655): p. 319-324.
35. Shi, X., et al., *Recent Advances in Investigating Functional Dynamics of Chromatin.* Front Genet, 2022. **13**: p. 870640.
36. Reyes, A.A., R.D. Marcum, and Y. He, *Structure and Function of Chromatin Remodelers.* J Mol Biol, 2021. **433**(14): p. 166929.
37. Clapier, C.R. and B.R. Cairns, *The biology of chromatin remodeling complexes.* Annu Rev Biochem, 2009. **78**: p. 273-304.
38. Rao, V.S., et al., *Protein-protein interaction detection: methods and analysis.* Int J Proteomics, 2014. **2014**: p. 147648.
39. Zhou, M., Q. Li, and R. Wang, *Current Experimental Methods for Characterizing Protein-Protein Interactions.* ChemMedChem, 2016. **11**(8): p. 738-56.

40. Prajapati, J.D., et al., *Computational Modeling of Ion Transport in Bulk and through a Nanopore Using the Drude Polarizable Force Field.* Journal of Chemical Information and Modeling, 2020. **60**(6): p. 3188-3203.
41. Ocello, R., et al., *Conduction and Gating Properties of the TRAAK Channel from Molecular Dynamics Simulations with Different Force Fields.* Journal of Chemical Information and Modeling, 2020. **60**(12): p. 6532-6543.
42. Allison, J.R., *Computational methods for exploring protein conformations.* Biochemical Society Transactions, 2020. **48**(4): p. 1707-1724.
43. Davtyan, A., et al., *AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing.* J Phys Chem B, 2012. **116**(29): p. 8494-503.
44. Durrant, J.D. and J.A. McCammon, *Molecular dynamics simulations and drug discovery.* BMC Biology, 2011. **9**(1): p. 71.
45. Robustelli, P., S. Piana, and D.E. Shaw, *Developing a molecular dynamics force field for both folded and disordered protein states.* Proceedings of the National Academy of Sciences, 2018. **115**(21): p. E4758-E4766.
46. Huang, J. and A.D. MacKerell, Jr., *Force field development and simulations of intrinsically disordered proteins.* Current opinion in structural biology, 2018. **48**: p. 40-48.
47. Kmiecik, S., et al., *Coarse-Grained Protein Models and Their Applications.* Chemical Reviews, 2016. **116**(14): p. 7898-7936.
48. Hospital, A., et al., *Molecular dynamics simulations: advances and applications.* Advances and applications in bioinformatics and chemistry : AABC, 2015. **8**: p. 37-47.
49. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**(5612): p. 585-90.
50. Karplus, M. and J. Kuriyan, *Molecular dynamics and protein function.* Proceedings of the National Academy of Sciences, 2005. **102**(19): p. 6679-6685.
51. Sinha, S., B. Tam, and S.M. Wang, *Applications of Molecular Dynamics Simulation in Protein Study.* Membranes (Basel), 2022. **12**(9).
52. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.
53. Hospital, A., et al., *Molecular dynamics simulations: advances and applications.* Adv Appl Bioinform Chem, 2015. **8**: p. 37-47.
54. Nguyen, D., L. Tao, and Y. Li, *Integration of Machine Learning and Coarse-Grained Molecular Simulations for Polymer Materials: Physical Understandings and Molecular Design.* Frontiers in Chemistry, 2022. **9**.
55. Beto, J.A., *The role of calcium in human aging.* Clin Nutr Res, 2015. **4**(1): p. 1-8.
56. Carafoli, E., *Calcium signaling: A tale for all seasons.* Proceedings of the National Academy of Sciences, 2002. **99**(3): p. 1115-1122.
57. Bootman, M.D., *Calcium Signaling.* Cold Spring Harbor Perspectives in Biology, 2012. **4**(7).
58. Marambaud, P., U. Dreses-Werringloer, and V. Vingtdeux, *Calcium signaling in neurodegeneration.* Molecular Neurodegeneration, 2009. **4**(1): p. 20.
59. Berridge, M.J., *Calcium microdomains: Organization and function.* Cell Calcium, 2006. **40**(5): p. 405-412.

60. Gifford, Jessica L., Michael P. Walsh, and Hans J. Vogel, *Structures and metal-ion-binding properties of the Ca2+-binding helix–loop–helix EF-hand motifs.* Biochemical Journal, 2007. **405**(2): p. 199-221.
61. Berridge, M.J., P. Lipp, and M.D. Bootman, *The versatility and universality of calcium signalling.* Nature Reviews Molecular Cell Biology, 2000. **1**(1): p. 11-21.
62. Chin, D. and A.R. Means, *Calmodulin: a prototypical calcium sensor.* Trends in Cell Biology, 2000. **10**(8): p. 322-328.
63. Chattopadhyaya, R., et al., *Calmodulin structure refined at 1.7 A resolution.* J Mol Biol, 1992. **228**(4): p. 1177-92.
64. Fernandes, I.P.G. and A.M. Oliveira-Brett, *Calcium-induced calmodulin conformational change. Electrochemical evaluation.* Bioelectrochemistry, 2017. **113**: p. 69-78.
65. Kuboniwa, H., et al., *Solution structure of calcium-free calmodulin.* Nat Struct Biol, 1995. **2**(9): p. 768-76.
66. Vetter, S.W. and E. Leclerc, *Novel aspects of calmodulin target recognition and activation.* Eur J Biochem, 2003. **270**(3): p. 404-14.
67. Barton, N.P., C.S. Verma, and L.S.D. Caves, *Inherent Flexibility of Calmodulin Domains: A Normal-Mode Analysis Study.* The Journal of Physical Chemistry B, 2002. **106**(42): p. 11036-11040.
68. Hoeflich, K.P. and M. Ikura, *Calmodulin in Action: Diversity in Target Recognition and Activation Mechanisms.* Cell, 2002. **108**(6): p. 739-742.
69. Villalobo, A., et al., *Calmodulin as a protein linker and a regulator of adaptor/scaffold proteins.* Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 2018. **1865**(3): p. 507-521.
70. Ikura, M., et al., *Solution structure of a calmodulin-target peptide complex by multidimensional NMR.* Science, 1992. **256**(5057): p. 632-638.
71. Chou, J.J., et al., *Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains.* Nat Struct Biol, 2001. **8**(11): p. 990-7.
72. Anthis, N.J., M. Doucleff, and G.M. Clore, *Transient, sparsely populated compact states of apo and calcium-loaded calmodulin probed by paramagnetic relaxation enhancement: interplay of conformational selection and induced fit.* J Am Chem Soc, 2011. **133**(46): p. 18966-74.
73. Wriggers, W., et al., *Structure and dynamics of calmodulin in solution.* Biophys J, 1998. **74**(4): p. 1622-39.
74. Shepherd, C.M. and H.J. Vogel, *A molecular dynamics study of Ca(2+)-calmodulin: evidence of interdomain coupling and structural collapse on the nanosecond timescale.* Biophysical journal, 2004. **87**(2): p. 780-791.
75. Qin, Z. and T.C. Squier, *Calcium-dependent stabilization of the central sequence between Met(76) and Ser(81) in vertebrate calmodulin.* Biophys J, 2001. **81**(5): p. 2908-18.
76. Fallon, J.L. and F.A. Quiocho, *A closed compact structure of native Ca(2+)-calmodulin.* Structure, 2003. **11**(10): p. 1303-7.
77. Mehler, E.L., J.L. Pascual-Ahuir, and H. Weinstein, *Structural dynamics of calmodulin and troponin C.* Protein Eng, 1991. **4**(6): p. 625-37.
78. Wang, Q., et al., *Protein recognition and selection through conformational and mutually induced fit.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(51): p. 20545-50.

79. Wall, M.E., J.B. Clarage, and G.N. Phillips, Jr., *Motions of calmodulin characterized using both Bragg and diffuse X-ray scattering.* Structure, 1997. **5**(12): p. 1599-1612.
80. Weinstein, H. and E.L. Mehler, *Ca(2+)-binding and structural dynamics in the functions of calmodulin.* Annu Rev Physiol, 1994. **56**: p. 213-36.
81. Meador, W.E., A.R. Means, and F.A. Quiocho, *Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures.* Science, 1993. **262**(5140): p. 1718-21.
82. Westerlund, A.M. and L. Delemotte, *Effect of Ca2+ on the promiscuous target-protein binding of calmodulin.* PLoS Comput Biol, 2018. **14**(4): p. e1006072.
83. van der Spoel, D., et al., *Bending of the calmodulin central helix: a theoretical study.* Protein science : a publication of the Protein Society, 1996. **5**(10): p. 2044-2053.
84. Yang, C., G.S. Jas, and K. Kuczera, *Structure and dynamics of calcium-activated calmodulin in solution.* J Biomol Struct Dyn, 2001. **19**(2): p. 247-71.
85. Vigil, D., et al., *Functional dynamics of the hydrophobic cleft in the N-domain of calmodulin.* Biophys J, 2001. **80**(5): p. 2082-92.
86. Komeiji, Y., Y. Ueno, and M. Uebayasi, *Molecular dynamics simulations revealed Ca2+-dependent conformational change of Calmodulin.* FEBS Letters, 2002. **521**(1): p. 133-139.
87. Yang, C., G.S. Jas, and K. Kuczera, *Structure, dynamics and interaction with kinase targets: computer simulations of calmodulin.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2004. **1697**(1): p. 289-300.
88. Liu, F., et al., *Molecular mechanism of multispecific recognition of Calmodulin through conformational changes.* Proceedings of the National Academy of Sciences, 2017: p. 201615949.
89. Delfino, F., et al., *Structural Transition States Explored With Minimalist Coarse Grained Models: Applications to Calmodulin.* Frontiers in Molecular Biosciences, 2019. **6**(104).
90. Nandigrami, P. and J.J. Portman, *Coarse-grained molecular simulations of allosteric cooperativity.* The Journal of Chemical Physics, 2016. **144**(10): p. 105101.
91. Gong, Z. and H. Sun, *A Coarse-Grained Force Field Parameterized for MgCl2 and CaCl2 Aqueous Solutions.* Journal of Chemical Information and Modeling, 2017. **57**(7): p. 1599-1608.
92. Monticelli, L., et al., *The MARTINI Coarse-Grained Force Field: Extension to Proteins.* Journal of Chemical Theory and Computation, 2008. **4**(5): p. 819-834.
93. Zuckerman, D.M., *Simulation of an Ensemble of Conformational Transitions in a United-Residue Model of Calmodulin.* The Journal of Physical Chemistry B, 2004. **108**(16): p. 5127-5137.
94. Tripathi, S., et al., *Conformational frustration in calmodulin-target recognition.* Journal of molecular recognition : JMR, 2015. **28**(2): p. 74-86.
95. Wang, Q., et al., *The Effect of Macromolecular Crowding, Ionic Strength and Calcium Binding on Calmodulin Dynamics.* PLOS Computational Biology, 2011. **7**(7): p. e1002114.
96. Wu, H., P.G. Wolynes, and G.A. Papoian, *AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins.* J Phys Chem B, 2018. **122**(49): p. 11115-11125.
97. Tsai, M.-Y., et al., *Electrostatics, structure prediction, and the energy landscapes for protein folding and binding.* Protein science : a publication of the Protein Society, 2016. **25**(1): p. 255-269.

98. Tripathi, S., et al., *Lessons in Protein Design from Combined Evolution and Conformational Dynamics.* Scientific Reports, 2015. **5**(1): p. 14259.
99. Yang, W., et al., *Structural analysis, identification, and design of calcium-binding sites in proteins.* Proteins: Structure, Function, and Bioinformatics, 2002. **47**(3): p. 344-356.
100. Drake, S.K., et al., *Molecular tuning of an EF-hand-like calcium binding loop. Contributions of the coordinating side chain at loop position 3.* The Journal of general physiology, 1997. **110**(2): p. 173-184.
101. Roux, B., *The Calculation of the Potential of Mean Force Using Computer-Simulations.* Computer Physics Communications, 1995. **91**(1-3): p. 275-282.
102. Shirts, M.R. and J.D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states.* J Chem Phys, 2008. **129**(12): p. 124105.
103. Heidorn, D.B. and J. Trewhella, *Comparison of the crystal and solution structures of calmodulin and troponin C.* Biochemistry, 1988. **27**(3): p. 909-915.
104. Seaton, B.A., et al., *Calcium-induced increase in the radius of gyration and maximum dimension of calmodulin measured by small-angle X-ray scattering.* Biochemistry, 1985. **24**(24): p. 6740-3.
105. Kataoka, M., et al., *Small-angle X-ray scattering study of calmodulin bound to two peptides corresponding to parts of the calmodulin-binding domain of the plasma membrane Ca2+ pump.* Biochemistry, 1991. **30**(25): p. 6247-51.
106. Homouz, D., et al., *Modulation of calmodulin plasticity by the effect of macromolecular crowding.* J Mol Biol, 2009. **391**(5): p. 933-43.
107. Probst, M.M., et al., *Molecular dynamics and x-ray investigation of an aqueous calcium chloride solution.* The Journal of Physical Chemistry, 1985. **89**(5): p. 753-759.
108. Liao, Q., et al., *Extending the Nonbonded Cationic Dummy Model to Account for Ion-Induced Dipole Interactions.* The journal of physical chemistry letters, 2017. **8**(21): p. 5408-5414.
109. Li, P. and K.M. Merz, Jr., *Metal Ion Modeling Using Classical Mechanics.* Chemical reviews, 2017. **117**(3): p. 1564-1686.
110. Fracchia, F., et al., *Force Field Parametrization of Metal Ions from Statistical Learning Techniques.* Journal of chemical theory and computation, 2018. **14**(1): p. 255-273.
111. Duarte, F., et al., *Force Field Independent Metal Parameters Using a Nonbonded Dummy Model.* The Journal of Physical Chemistry B, 2014. **118**(16): p. 4351-4362.
112. Gaertner, T.R., J.A. Putkey, and M.N. Waxham, *RC3/Neurogranin and Ca2+/calmodulin-dependent protein kinase II produce opposing effects on the affinity of calmodulin for calcium.* J Biol Chem, 2004. **279**(38): p. 39374-82.
113. Lepšík, M. and M.J. Field, *Binding of Calcium and Other Metal Ions to the EF-Hand Loops of Calmodulin Studied by Quantum Chemical Calculations and Molecular Dynamics Simulations.* The Journal of Physical Chemistry B, 2007. **111**(33): p. 10012-10022.
114. Capozzi, F., F. Casadei, and C. Luchinat, *EF-hand protein dynamics and evolution of calcium signal transduction: an NMR view.* JBIC Journal of Biological Inorganic Chemistry, 2006. **11**(8): p. 949-962.
115. Malmendal, A., et al., *Structural dynamics in the C-terminal domain of calmodulin at low calcium levels.* J Mol Biol, 1999. **293**(4): p. 883-99.

116. Shimoyama, H., *A structural comparison of 'real' and 'model' calmodulin clarified allosteric interactions regulating domain motion.* Journal of Biomolecular Structure and Dynamics, 2019. **37**(6): p. 1567-1581.
117. Jing, Z., et al., *Study of interactions between metal ions and protein model compounds by energy decomposition analyses and the AMOEBA force field.* The Journal of chemical physics, 2017. **147**(16): p. 161733-161733.
118. McCarthy, M.R., et al., *Impact of methionine oxidation on calmodulin structural dynamics.* Biochemical and Biophysical Research Communications, 2015. **456**(2): p. 567-572.
119. Project, E., et al., *A Molecular Dynamics Study of the Effect of Ca2+ Removal on Calmodulin Structure.* Biophysical Journal, 2006. **90**(11): p. 3842-3850.
120. Gilli, R., et al., *Thermodynamic Analysis of Calcium and Magnesium Binding to Calmodulin.* Biochemistry, 1998. **37**(16): p. 5450-5456.
121. Tidow, H. and P. Nissen, *Structural diversity of calmodulin binding to its target sites.* Febs j, 2013. **280**(21): p. 5551-65.
122. Johnson, C.K., *Calmodulin, conformational states, and calcium signaling. A single-molecule perspective.* Biochemistry, 2006. **45**(48): p. 14233-14246.
123. Mohanan, A.G., et al., *Role of Ca(2+)/Calmodulin-Dependent Protein Kinase Type II in Mediating Function and Dysfunction at Glutamatergic Synapses.* Front Mol Neurosci, 2022. **15**: p. 855752.
124. Buonarati, O.R., et al., *Conserved and divergent features of neuronal CaMKII holoenzyme structure, function, and high-order assembly.* Cell Reports, 2021. **37**(13): p. 110168.
125. Hell, J.W., *CaMKII: claiming center stage in postsynaptic function and organization.* Neuron, 2014. **81**(2): p. 249-65.
126. Myers, J.B., et al., *The CaMKII holoenzyme structure in activation-competent conformations.* Nature communications, 2017. **8**(1): p. 1-15.
127. Putkey, J.A. and M.N. Waxham, *A peptide model for calmodulin trapping by calcium/calmodulin-dependent protein kinase II.* The Journal of biological chemistry, 1996. **271**(47): p. 29619-23.
128. Meador, W.E., A.R. Means, and F.A. Quiocho, *Modulation of calmodulin plasticity in molecular recognition on the basis of X-ray structures.* Science, 1993. **262**: p. 1718-1721.
129. Waxham, M.N., A.L. Tsai, and J.A. Putkey, *A mechanism for calmodulin (CaM) trapping by CaM-kinase II defined by a family of CaM-binding peptides.* J. Biol. Chem., 1998. **273**: p. 17579-17584.
130. Meyer, T., et al., *Calmodulin trapping by calcium-calmodulin-dependent protein kinase.* Science 1992. **256**: p. 1199-1202.
131. Borodin, O. and G.D. Smith, *Development of quantum chemistry-based force fields for poly(ethylene oxide) with many-body polarization interactions.* Journal of Physical Chemistry B, 2003. **107**(28): p. 6801-6812.
132. Starovoytov, O.N., et al., *Induced Polarization Restricts Conformational Distribution of a Light-Harvesting Molecular Triad in the Ground State.* Phys. Chem. Chem. Phys., 2017. **19**: p. 22969-22980.
133. Ezerski, J.C., et al., *Molecular Dynamics Ensemble Refinement of Intrinsically Disordered Peptides According to Deconvoluted Spectra from Circular Dichroism.* Biophys J, 2020. **accepted**.

134. Payne, M.E., et al., *Calcium/calmodulin-dependent protein kinase II. Characterization of distinct calmodulin binding and inhibitory domains.* Journal of Biological Chemistry, 1988. **263**(15): p. 7190-7195.
135. Hanley, R.M., et al., *Mapping of calmodulin-binding domain of Ca2+/calmodulin-dependent protein kinase II from rat brain.* Biochemical and Biophysical Research Communications, 1988. **152**(1): p. 122-128.
136. Colbran, R.J., et al., *Regulatory Domain of Calcium/Calmodulin-dependent Protein Kinase II: Mechanism of Inhibition and Regulation by Phosphorylation.* Journal of Biological Chemistry, 1989. **264**(9): p. 4800-4804.
137. Clapham, D.E., *Calcium signaling.* Cell, 2007. **131**: p. 1047-1058.
138. Watterson, D.M., F. Sharief, and T.C. Vanaman, *The Complete Amino Acid Sequence of the Ca^{2+}-dependent Modulator Protein (Calmodulin) of Bovine Brain.* J Biol Chem, 1980. **255**: p. 962-975.
139. Yamniuk, A.P. and H. Vogel, *Structually homologous binding of plant calmodulin isoforms to the calmodulin-binding domain of vacuolar calcium-ATPase.* J. Biol. Chem., 2004. **279**.
140. Hoeflich, K.P. and M. Ikura, *Calmodulin in action: diversity in target recognition and activation mechanisms.* Cell, 2002. **108**: p. 739-742.
141. Kubota, Y., et al., *IQ-motif proteins influence intracellular free Ca2+ in hippocampal neurons through their interactions with calmodulin.* Journal of neurophysiology, 2008. **99**(1): p. 264-76.
142. Gaertner, T.R., J.A. Putkey, and M.N. Waxham, *RC3/Neurogranin and Ca2+/Calmodulin-dependent protein kinase II produce opposing effects on the affinity of calmodulin for calcium.* The Journal of Biological Chemistry, 2004. **279**: p. 39374-39382.
143. Zhang, P., et al., *Opposing Intermolecular Tuning of Ca(2+) Affinity for Calmodulin by Neurogranin and CaMKII Peptides.* Biophys J, 2017. **112**(6): p. 1105-1119.
144. Chattopadhyaya, R., et al., *Calmodulin structure refined at 1.7 A resolution.* J. Mol. Biol., 1992. **228**: p. 1177-1192.
145. Martin, S.R., et al., *Kinetics of calcium dissociation from calmodulin and its tryptic fragments. A stopped-flow fluorescence study using Quin 2 reveals a two-domain structure.* Eur J Biochem, 1985. **151**(3): p. 543-50.
146. Shepard, C.M. and H.J. Vogel, *A molecular dynamics study of Ca2+-calmodulin evidence of interdomain coupling and structural collapse on the nanosecond time scale.* Biophysical Journal, 2004. **87**: p. 780-791.
147. Kovesi, I., et al., *Interaction of antagonists with Calmodulin: Insights from molecular dynamics simulations.* J. Med. Chem., 2008. **51**: p. 3081-3093.
148. Shukla, D., A. Peck, and V.S. Pande, *Conformational heterogeneity of the calmodulin binding interface.* Nature Communications, 2016. **7**.
149. Westerlund, A.M. and L. Delemotte, *Effect of Ca2+ on the promiscuous target-protein binding of calmodulin.* PloS Computational Biology, 2018.
150. Zuckerman, D.M., *Simulation of an ensemble of conformational transitions in a united-residue model of calmodulin.* J Phys Chem B, 2004. **108**(16): p. 5127-5137.
151. Chen, Y.-G. and G. Hummer, *Slow conformational dynamics and unfolding of the calmodulin C-terminal domain.* J Am Chem Soc, 2007. **129**: p. 2414-2415.

152. Nandigrami, P. and J.J. Portman, *Coarse-grained molecular simulations of allosteric copperativity.* Journal of Chemical Physics, 2016. **144**: p. 105101.
153. Li, W., W. Wang, and S. Takada, *Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains.* Proceedings of the National Academy of Sciences, 2014. **111**(29): p. 10550-10555.
154. Humphreys, I.R., et al., *Computed structures of core eukaryotic protein complexes.* Science, 2021. **374**(6573): p. 1340-+.
155. Davtyan, A., et al., *AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing.* Journal of Physical Chemistry B, 2012. **116**: p. 8494-8503.
156. Tsai, M.Y., et al., *Electrostatics, structure prediction and the energy landscapes for protein folding and binding.* Protein Science, 2016 **25**: p. 255-269.
157. Zheng, W., et al., *Exploring the aggregation free energy landscape of the amyloid-beta protein (1-40).* Proc Nat Acad Sci USA., 2016. **113**: p. 11835-11840.
158. Ezerski, J.C., et al., *Molecular Dynamics Ensemble Refinement of Intrinsically Disordered Peptides According to Deconvoluted Spectra from Circular Dichroism.* Biophys J, 2020. **118**: p. 1665-1678.
159. Zhang, P., et al., *Determining the atomic charge of calcium ion requires the information of its coordination geometry in an EF-hand motif.* Journal of Chemical Physics, 2021. **154**: p. 124104.
160. Forest, A., et al., *Role of the N- and C-lobes of calmodulin in the activation of Ca(2+)/calmodulin-dependent protein kinase II.* Biochemistry, 2008. **47**(40): p. 10587-99.
161. Tripathi, S., et al., *Conformational frustration in calmodulin-target recognition.* Journal of Molecular Recognition 2015. **28**: p. 74-86.
162. Liwo, A., et al., *Theory and Practice of Coarse-Grained Molecular Dynamics of Biologically Important Systems.* Biomolecules, 2021. **11**(9).
163. Taketomi, H., Y. Ueda, and N. Gō, *Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions.* Int J Pept Protein Res, 1975. **7**(6): p. 445-59.
164. de Jong, D.H., et al., *Improved Parameters for the Martini Coarse-Grained Protein Force Field.* Journal of Chemical Theory and Computation, 2013. **9**(1): p. 687-697.
165. Kolinski, A., *Protein modeling and structure prediction with a reduced representation.* Acta Biochim Pol, 2004. **51**(2): p. 349-71.
166. Plimpton, S., *Fast parallel algorithms for short-range molecular dynamics.* Journal of Computationl Physics, 1995. **117**: p. 1-19.
167. Zuckerman, D.M., *Equilibrium sampling in biomolecular simulations.* Annu Rev Biophys, 2011. **40**: p. 41-62.
168. Sanbonmatsu, K.Y. and A.E. García, *Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics.* Proteins, 2002. **46**(2): p. 225-34.
169. Torrie, G.M. and J.P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* Journal of Computational Physics, 1977. **23**(2): p. 187-199.
170. Evans, R., et al., *Combining Machine Learning and Enhanced Sampling Techniques for Efficient and Accurate Calculation of Absolute Binding Free Energies.* Journal of Chemical Theory and Computation, 2020. **16**(7): p. 4641-4654.

171. Shirts, M.R. and A.L. Ferguson, *Statistically Optimal Continuous Free Energy Surfaces from Biased Simulations and Multistate Reweighting.* Journal of Chemical Theory and Computation, 2020. **16**(7): p. 4107-4125.
172. Shirts, M.R. and J.D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states.* The Journal of Chemical Physics, 2008. **129**(12): p. 124105.
173. Robustelli, P., S. Piana, and D.E. Shaw, *Developing a molecular dynamics force field for both folded and disordered protein states.* Proc Natl Acad Sci U S A, 2018. **115**: p. E4758-E4766.
174. Wu, Y., et al., *Effect of Ca²⁺ binding states of calmodulin on the conformational dynamics and force responses of myosin lever arm.* The Journal of Chemical Physics, 2022. **157**(3): p. 035101.
175. Aykut, a.o., A. Atilgan, and C. Atilgan, *Designing Molecular Dynamics Simulations to Shift Populations of the Conformational States of Calmodulin.* PLoS computational biology, 2013. **9**: p. e1003366.
176. Wriggers, W., et al., *Structure and dynamics of calmodulin in solution.* Biophysical Journal, 1998. **74**: p. 1622-1639.
177. Greene, D.A., et al., *Computational Analysis of Binding Interactions between the Ryanodine Receptor Type 2 and Calmodulin.* The Journal of Physical Chemistry B, 2021. **125**(38): p. 10720-10735.
178. Basit, A., A.K. Yadav, and P. Bandyopadhyay, *Calcium Ion Binding to the Mutants of Calmodulin: A Structure-Based Computational Predictive Model of Binding Affinity Using a Charge Scaling Approach in Molecular Dynamics Simulation.* Journal of Chemical Information and Modeling, 2022. **62**(11): p. 2821-2834.
179. Kohagen, M., M. Lepsik, and P. Jungwirth, *Calcium Binding to Calmodulin by Molecular Dynamics with Effective Polarization.* The journal of physical chemistry letters, 2014. **5**(22): p. 3964-9.
180. Guan, X., et al., *Role of water-bridged interactions in metal ion coupled protein allostery.* PLOS Computational Biology, 2022. **18**(6): p. e1010195.
181. Nodelman, I.M. and G.D. Bowman, *Biophysics of Chromatin Remodeling.* Annu Rev Biophys, 2021. **50**: p. 73-93.
182. Otsuka, S., et al., *A quantitative map of nuclear pore assembly reveals two distinct mechanisms.* Nature, 2023. **613**(7944): p. 575-581.
183. Poli, J., S.M. Gasser, and M. Papamichos-Chronakis, *The INO80 remodeller in transcription, replication and repair.* Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1731).
184. Brahma, S., et al., *The Arp8 and Arp4 module acts as a DNA sensor controlling INO80 chromatin remodeling.* Nat Commun, 2018. **9**(1): p. 3309.
185. Liu, J., et al., *Enhancing alphafold-multimer-based protein complex structure prediction with MULTICOM in CASP15.* Communications Biology, 2023. **6**(1): p. 1140.
186. Kim, S.J., et al., *Integrative structure and functional anatomy of a nuclear pore complex.* Nature, 2018. **555**(7697): p. 475-482.
187. Karaca, E. and A.M. Bonvin, *Advances in integrative modeling of biomolecular complexes.* Methods, 2013. **59**(3): p. 372-81.
188. Magaña-Acosta, M. and V. Valadez-Graham, *Chromatin Remodelers in the 3D Nuclear Compartment.* Front Genet, 2020. **11**: p. 600615.

189. Eustermann, S., et al., *Structural basis for ATP-dependent chromatin remodelling by the INO80 complex.* Nature, 2018. **556**(7701): p. 386-390.
190. Willhoft, O. and D.B. Wigley, *INO80 and SWR1 complexes: the non-identical twins of chromatin remodelling.* Curr Opin Struct Biol, 2020. **61**: p. 50-58.
191. Ricketts, M.D., et al., *Molecular basis for chromatin assembly and modification by multiprotein complexes.* Protein Sci, 2019. **28**(2): p. 329-343.
192. Baker, R.W., et al., *Structural insights into assembly and function of the RSC chromatin remodeling complex.* Nat Struct Mol Biol, 2021. **28**(1): p. 71-80.
193. Oberbeckmann, E., et al., *Ruler elements in chromatin remodelers set nucleosome array spacing and phasing.* Nat Commun, 2021. **12**(1): p. 3232.
194. Oberbeckmann, E., et al., *Genome information processing by the INO80 chromatin remodeler positions nucleosomes.* Nat Commun, 2021. **12**(1): p. 3231.
195. Wang, J., et al., *Roles of the INO80 and SWR1 Chromatin Remodeling Complexes in Plants.* Int J Mol Sci, 2019. **20**(18).
196. Aramayo, R.J., et al., *Cryo-EM structures of the human INO80 chromatin-remodeling complex.* Nat Struct Mol Biol, 2018. **25**(1): p. 37-44.
197. Willhoft, O., et al., *Crosstalk within a functional INO80 complex dimer regulates nucleosome sliding.* eLife, 2017. **6**: p. e25782.
198. Shen, X., et al., *A chromatin remodelling complex involved in transcription and DNA processing.* Nature, 2000. **406**(6795): p. 541-4.
199. Lin, C.L., et al., *Functional characterization and architecture of recombinant yeast SWR1 histone exchange complex.* Nucleic Acids Res, 2017. **45**(12): p. 7249-7260.
200. Watanabe, S., et al., *Structural analyses of the chromatin remodelling enzymes INO80-C and SWR-C.* Nat Commun, 2015. **6**: p. 7108.
201. Tosi, A., et al., *Structure and subunit topology of the INO80 chromatin remodeler and its nucleosome complex.* Cell, 2013. **154**(6): p. 1207-19.
202. Chen, L., et al., *Subunit organization of the human INO80 chromatin remodeling complex: an evolutionarily conserved core complex catalyzes ATP-dependent nucleosome remodeling.* J Biol Chem, 2011. **286**(13): p. 11283-9.
203. Sardiu, M.E., et al., *Conserved abundance and topological features in chromatin-remodeling protein interaction networks.* EMBO Rep, 2015. **16**(1): p. 116-26.
204. Sundaramoorthy, R. and T. Owen-Hughes, *Chromatin remodelling comes into focus.* F1000Res, 2020. **9**.
205. Jin, J., et al., *A mammalian chromatin remodeling complex with similarities to the yeast INO80 complex.* J Biol Chem, 2005. **280**(50): p. 41207-12.
206. Knoll, K.R., et al., *The nuclear actin-containing Arp8 module is a linker DNA sensor driving INO80 chromatin remodeling.* Nat Struct Mol Biol, 2018. **25**(9): p. 823-832.
207. Chen, L., R.C. Conaway, and J.W. Conaway, *Multiple modes of regulation of the human Ino80 SNF2 ATPase by subunits of the INO80 chromatin-remodeling complex.* Proc Natl Acad Sci U S A, 2013. **110**(51): p. 20497-502.
208. Ayala, R., et al., *Structure and regulation of the human INO80-nucleosome complex.* Nature, 2018. **556**(7701): p. 391-395.
209. Sardiu, M.E., et al., *Topological scoring of protein interaction networks.* Nat Commun, 2019. **10**(1): p. 1118.
210. Zhang, X., et al., *Structure and functional interactions of INO80 actin/Arp module.* J Mol Cell Biol, 2019. **11**(5): p. 345-355.

211. Kunert, F., et al., *Structural mechanism of extranucleosomal DNA readout by the INO80 complex.* Sci Adv, 2022. **8**(49): p. eadd3189.
212. Wu, J., Y. Lao, and B. Li, *Nuclear actin switch of the INO80 remodeler.* J Mol Cell Biol, 2019. **11**(5): p. 343-344.
213. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
214. Kozakov, D., et al., *The ClusPro web server for protein–protein docking.* Nature Protocols, 2017. **12**(2): p. 255-278.
215. Dominguez, C., R. Boelens, and A.M.J.J. Bonvin, *HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information.* Journal of the American Chemical Society, 2003. **125**(7): p. 1731-1737.
216. Boehr, D.D. and P.E. Wright, *How Do Proteins Interact?* Science, 2008. **320**(5882): p. 1429-1430.
217. Lange, O.F., et al., *Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution.* Science, 2008. **320**(5882): p. 1471-1475.
218. Ringe, D. and G.A. Petsko, *How Enzymes Work.* Science, 2008. **320**(5882): p. 1428-1429.
219. Stigler, J., et al., *The Complex Folding Network of Single Calmodulin Molecules.* Science, 2011. **334**(6055): p. 512-516.
220. Zheng, L.E., et al., *Machine Learning Generation of Dynamic Protein Conformational Ensembles.* Molecules, 2023. **28**(10).
221. Shi, Y.-F., et al., *Machine Learning for Chemistry: Basics and Applications.* Engineering, 2023.
222. Bradford, G., et al., *Chemistry-Informed Machine Learning for Polymer Electrolyte Discovery.* ACS Central Science, 2023. **9**(2): p. 206-216.
223. Zhang, L.V., A. Marani, and M.L. Nehdi, *Chemistry-informed machine learning prediction of compressive strength for alkali-activated materials.* Construction and Building Materials, 2022. **316**: p. 126103.
224. Gupta, A., et al., *Artificial intelligence guided conformational mining of intrinsically disordered proteins.* Communications Biology, 2022. **5**(1): p. 610.
225. Audagnotto, M., et al., *Machine learning/molecular dynamic protein structure prediction approach to investigate the protein conformational ensemble.* Sci Rep, 2022. **12**(1): p. 10018.
226. AlQuraishi, M., *Machine learning in protein structure prediction.* Curr Opin Chem Biol, 2021. **65**: p. 1-8.
227. Xu, Y., et al., *Deep Dive into Machine Learning Models for Protein Engineering.* J Chem Inf Model, 2020. **60**(6): p. 2773-2790.

APPENDIX A

Table A.1. Experimental techniques to study proteins, protein-protein interactions, and protein complexes.

Approach	Purpose	Principle	Application
X-ray Crystallography	Determines the three-dimensional structure of proteins	X-rays are directed at a crystallized protein, and the resulting diffraction pattern is analyzed to determine the arrangement of atoms in the protein	Reveals atomic-level details of protein structures
Nuclear Magnetic Resonance (NMR) Spectroscopy	Provides information about the structure and dynamics of proteins in solution	NMR detects signals from atomic nuclei in a magnetic field, and the data is used to deduce structural information	Useful for studying smaller proteins and investigating dynamic changes in protein structures
Mass Spectrometry	Identifies and quantifies proteins based on their mass and charge	Proteins are ionized and separated based on their mass-to-charge ratio, allowing for the determination of their mass and identification	Protein identification, characterization, and determination of post-translational modifications
Cryo-electron microscopy (cryo-EM)	Investigate large biological complexes including macromolecular assemblies, cellular organelles; imaging of samples in their native, hydrated state (crucial for preserving the structural and functional integrity of biological specimens)	It uses a transmission electron microscope to generate high-resolution images of frozen-hydrated specimens. The samples are rapidly frozen to cryogenic temperatures, typically in liquid ethane, to capture them in a vitrified (glass-like) state. This minimizes artifacts introduced by dehydration or chemical fixation. Images are recorded at various tilt angles, allowing the reconstruction of a three-dimensional structure using computational techniques. It captures information about	Three-dimensional architecture of cellular organelles, such as the nucleus, endoplasmic reticulum, and mitochondria, providing insights into their structure and organization; structure and organization of large protein complexes, such as molecular machines; structure of large viruses and their interactions with host cells (which aids in understanding viral

		the entire sample, providing a holistic view of large complexes	entry, replication, and assembly processes)
Cryo-electron tomography (cryo-ET)	Visualize the three-dimensional structures of biological specimens, such as cells or organelles, in their native, hydrated state; study the spatial organization of proteins and other macromolecules within the context of the cellular environment	capture a series of two-dimensional images of a frozen-hydrated sample at different tilt angles (the specimen is typically flash-frozen to preserve its native state); The sample is tilted at various angles, and images are acquired at each tilt. This tilt series of images provides information about the three-dimensional structure of the specimen; Computational methods are then used to reconstruct a three-dimensional volume from the tilt series of images	Study the three-dimensional architecture of subcellular structures, such as organelles, membranes, and cellular compartments; investigate the three-dimensional structures of proteins and other macromolecules within their native cellular context
Fluorescence Resonance Energy Transfer (FRET)	Investigate molecular interactions between proteins or between different regions of the same protein (useful for studying dynamic processes in real-time)	It relies on the energy transfer between two fluorophores (donor and acceptor) when they are in close proximity (typically within 1-10 nm). The donor fluorophore is excited by an external light source, and part of the energy is transferred to the acceptor fluorophore if they are close enough	Study protein-protein interactions in living cells; study conformational changes within a single protein
Electrophoresis	Separates proteins based on their size and charge	Proteins are subjected to an electric field, causing them to migrate through a gel matrix. The separation is based on size and charge	SDS-PAGE (Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis) for size separation; Native PAGE for native protein complexes
Western Blotting	Detects and quantifies specific proteins in a sample	Proteins are separated by electrophoresis, transferred to a membrane, and probed	Protein detection, quantification, and analysis of post-

		with antibodies specific to the protein of interest	translational modifications
Enzyme-Linked Immunosorbent Assay (ELISA)	Quantifies the concentration of a specific protein in a sample	Uses antibodies to bind to the target protein, and the amount of bound antibody is measured	Biomarker detection, protein quantification in biological samples
Surface Plasmon Resonance (SPR)	Measures real-time interactions between proteins	Utilizes changes in the refractive index on a sensor surface due to binding events, allowing for the monitoring of protein-protein or protein-ligand interactions	Kinetics and affinity studies, drug discovery
Protein Crystallization	Prepares proteins for X-ray crystallography	Proteins are crystallized to obtain well-ordered structures suitable for X-ray analysis	Structural determination of proteins
Protein Microarrays	High-throughput analysis of protein interactions	Proteins are immobilized on a surface, and their interactions with other proteins or molecules are studied	Profiling protein-protein interactions, identifying binding partners
Circular Dichroism (CD)	It provides information about the secondary structure of proteins. It can distinguish between different types of secondary structures, such as α-helices, β-sheets, and random coils	refers to the differential absorption of left-handed circularly polarized light (L-CPL) and right-handed circularly polarized light (R-CPL). This phenomenon arises due to the interaction of circularly polarized light with chiral molecules, such as optically active amino acids in proteins; the CD spectrum provides information about the secondary structure composition of the protein (alpha-helical structures typically exhibit two minima around 208 and 222 nm, while beta-sheet structures have a minimum near 218 nm)	Used to estimate the secondary structure content of proteins; monitors changes in protein conformation during processes like folding and unfolding; suitable for studying conformational changes induced by environmental factors such as temperature, pH, or the presence of denaturants. As well to monitor the impact of mutations on protein structure
stopped-flow spectroscopy	Investigates the kinetics of fast chemical or	It involves the rapid mixing of two or more reactants in a reaction chamber. After	Use to investigates the kinetics of ligand binding to proteins. It

	biochemical reactions, such as protein folding, ligand binding, enzyme catalysis, and conformational changes (captures the rapid time course of these reactions with millisecond to microsecond time resolution)	mixing, the reaction progress is monitored over time using a spectroscopic technique such as absorbance, fluorescence, or circular dichroism; the mixing is initiated by rapidly displacing a barrier that separates the reactants in the reaction chamber. This triggers the reaction, and the resulting changes in spectroscopic signals are recorded over time	provides information about association and dissociation rates, as well as the identification of intermediate states in the binding process; It is also used to study the kinetics of enzymatic reactions, including substrate binding, catalysis, and product release
Proteomics Techniques	Comprehensive analysis of the entire set of proteins (proteome) in a cell or organism	Involves techniques such as mass spectrometry, 2D gel electrophoresis, and protein microarrays	Identification of proteins, quantification, and characterization of post-translational modifications

Table A.2. Limitations and implications of the experimental techniques used to characterize proteins.

Sources	**Limitation**	**Implications**
Protein Crystallization Challenges	Not all proteins can be easily crystallized	Some proteins may resist crystallization, hindering the determination of their three-dimensional structures using X-ray crystallography
Dynamic Nature of Proteins	Many experimental techniques provide static snapshots and may not capture the dynamic nature of proteins	Important conformational changes, dynamics, and interactions may be overlooked, impacting the understanding of functional mechanisms

Size and Complexity	Large and complex proteins or protein complexes can be challenging to study	Techniques such as NMR and X-ray crystallography may face difficulties in resolving intricate structures, leading to incomplete information
Protein Folding and Misfolding	Understanding the process of protein folding and misfolding is challenging	Protein folding is a complex process, and misfolding can lead to diseases. Experimental techniques may not fully capture these intricate events
Technical Constraints in Imaging	Imaging techniques may have limitations in resolution and sensitivity	Fine details at the molecular level may not be discernible, affecting the accuracy of structural information
Low Abundance Proteins	Low-abundance proteins may be challenging to detect and quantify	Important regulatory or signaling proteins with low expression levels may be overlooked in proteomic analyses
Artifacts and Interference	Sample preparation and experimental conditions may introduce artifacts or interference	Obtained results might be influenced by factors such as sample purification methods or the choice of experimental buffers
Quantification Accuracy	Quantifying protein levels may be subject to inaccuracies	Obtained results might be influenced by factors such as sample purification methods or the choice of experimental buffers
Overexpression Artifacts	Overexpression of proteins for study may introduce artifacts	Protein folding, localization, and interactions may differ from

		natural conditions, affecting the relevance of the obtained data
Limited Functional Insights	Experimental techniques may not always provide direct insights into protein function	Understanding the functional roles of proteins often requires a combination of experimental and computational approaches
Ethical and Technical Constraints in Human Studies	Studying human proteins directly can be challenging ethically and technically	Insights gained from model organisms may not fully represent human biology
Data Integration Challenges	Integrating data from different experimental techniques can be complex	A comprehensive understanding of a protein's behavior may require integrating information from various sources, and discrepancies can arise

Table A.3. Computational techniques to study proteins, protein-protein interactions, and protein complexes.

Approach	**Purpose**	**Principle**	**Application**
Homology Modeling	Predicts the three-dimensional structure of a protein based on the known structure of a homologous protein	Aligns the sequence of the target protein with a template protein of known structure, and the model is built based on the template's structure	Provides structural insights when experimental structures are unavailable
Molecular Dynamics (MD) Simulations	Simulates the movement and interactions of atoms and molecules over time to study protein dynamics	Solves Newton's equations of motion numerically, allowing the simulation of protein behavior at an atomic level	Investigates protein folding, conformational changes, and interactions with ligands
Protein-Ligand Docking	Predicts the binding mode and affinity of a small molecule (ligand) to a protein target	Computes energetically favorable conformations of the ligand within the protein's binding site	Drug discovery, predicting protein-ligand interactions

Quantum Mechanics/Molecular Mechanics (QM/MM)	Combines quantum mechanics for the active site with molecular mechanics for the rest of the system to model enzymatic reactions	Allows for accurate description of the electronic structure in the active site while considering the rest of the system at a classical level	Studies enzymatic mechanisms, reaction pathways
Protein Structure Prediction	Predicts protein structure from its amino acid sequence when experimental structures are unavailable	Integrates information from homology modeling, threading, and ab initio methods to generate structural models	Aids in understanding protein function and designing experiments
Bioinformatics and Sequence Analysis	Analyzes protein sequences to identify patterns, motifs, and potential functional domains	Involves algorithms for sequence alignment, motif discovery, and phylogenetic analysis	Functional annotation, predicting post-translational modifications
Network Analysis	Studies protein-protein interaction networks and signaling pathways	Analyzes large-scale interaction data to identify hubs, pathways, and functional modules	Understanding cellular processes, disease mechanisms
Protein Folding Prediction	Predicts the three-dimensional structure of a protein based on its amino acid sequence	Combines physics-based force fields, statistical potentials, and machine learning methods	Understanding protein folding pathways, studying misfolding diseases
Artificial Intelligence	Predicts protein function, structure, and interactions using data-driven approaches	Trains algorithms on large datasets to recognize patterns and make predictions	Predicting protein functions, identifying drug targets
Proteomics Data Analysis	Analyzes large-scale proteomics datasets	Involves statistical methods, pathway analysis, and integration with other omics data	Identifying biomarkers, understanding global changes in protein expression

Table A.4. Limitations and the associated implications linked to the computational approaches used in the study of protein.

Sources	**Limitation**	**Implications**

Accuracy of Protein Structure Prediction	Predicting the three-dimensional structure of proteins with high accuracy remains challenging	Inaccurate structural predictions can impact downstream analyses and applications, such as drug design
Homology Modeling Reliance on Templates	Homology modeling heavily relies on the availability of homologous protein structures	In the absence of close homologs, accuracy diminishes, and models may not accurately represent the target protein
Molecular Dynamics Simulation Time Scales	Molecular dynamics simulations often cover limited time scales compared to biological processes	Long-timescale events, such as large conformational changes, may not be adequately captured
Scalability and Computational Resources	Some calculations are computationally intensive and require significant resources	Accessibility and affordability of resources can limit the scope and scale of computational studies
Complexity of Protein-Ligand Docking	Protein-ligand docking predictions can be sensitive to the choice of algorithms and parameters	False positives or false negatives in binding predictions may affect drug discovery efforts
Inadequate Representation of Solvent Effects	Some computational models may oversimplify or neglect solvent effects	Inaccuracies in predicting interactions with water molecules can affect the realism of simulations
Data-Driven Approach Dependency	Machine learning and data-driven approaches heavily depend on the quality and quantity of training data	Biases or limitations in training data can impact the generalization of models
Prediction of Protein Flexibility	Predicting the flexibility of proteins accurately is challenging	Incorrect assessment of protein flexibility can affect predictions related to binding and function

Limited Understanding of Protein Folding	Despite progress, a complete understanding of the protein folding process is lacking	Inaccuracies in predicting folding pathways may limit insights into protein misfolding diseases
Lack of Experimental Validation	Computational predictions often need experimental validation	The lack of validation can introduce uncertainties, and experimental verification may be challenging for certain predictions
Biomolecular Interaction Predictions	Predicting complex biomolecular interactions accurately is complex	The accuracy of interaction predictions in large-scale networks may be limited
Biological Context Consideration	Some computational models may oversimplify the biological context	The complexity of cellular environments may not be fully captured in certain simulations or predictions
Insufficient Treatment of Post-Translational Modifications	Some computational methods may struggle to accurately model post-translational modifications	The impact of modifications on protein function may be underestimated
Challenges in Ab Initio Structure Prediction	Ab initio methods for structure prediction face challenges for larger proteins	Accurate prediction for proteins without homologous structures remains a formidable task
Ethical Considerations	In some cases, computational studies may lack ethical considerations relevant to biological experiments	Ethical implications of certain predictions or applications may be overlooked

APPENDIX B

1-) List of parameters used in the energy function to conduct the coarse-grained MD simulations.

```
[Chain]
10.0 10.0 30.0
2.45798 2.50665 2.44973
[Shake]
10.0 3.83 2.36 2.71
[Chi]
20.0 -0.83
[Excluded]
5.0 3.0
5.0 4.0
[Excluded_P]-
4
1.0 3.0
1.0 4.0
[Excluded_R6]-
39.5935 4.0
39.5935 4.0
[Epsilon]
1.0
[Rama]
2.0
3
1.3149 15.398 0.15 1.74 0.65 -2.138
1.32016 49.0521 0.25 1.265 0.45 0.318
1.0264 49.0954 0.65 -1.041 0.25 -0.78
2.0 419.0 1.0 0.995 1.0 0.820
2.0 15.398 1.0 2.25 1.0 -2.16
[Rama_P]
3
0.0 0.0 1.0 0.0 1.0 0.0
2.17 105.52 1.0 1.153 0.15 -2.4
2.15 109.09 1.0 0.95 0.15 0.218
0.0 0.0 1.0 0.0 2.0 0.0
0.0 0.0 1.0 0.0 2.0 0.0
[SSWeight]
0 0 0 1 1 0
0 0 0 0 0 0
[ABC]
0.483 0.703 -0.186
0.444 0.235 0.321
0.841 0.893 -0.734
[Dssp_Hdrgn]
1.0
0.0 0.0
1.37 0.0 3.49 1.30 1.32 1.22 0.0
1.36 0.0 3.50 1.30 1.32 1.22 3.47 0.33 1.01
```

1.17 0.0 3.52 1.30 1.32 1.22 3.62 0.33 1.01
0.76 0.68
2.06 2.98
7.0
1.0 0.5
12.0
[P_AP]-
1.0
1.5
1.0 0.4 0.4
8.0
7.0
5 8
4
[Water]
1.0
5.0 7.0
2.6
10
2
4.5 6.5 1
6.5 9.5 1
[Burial]
1.0
4.0
0.0 3.0
3.0 6.0
6.0 9.0
[Helix]
1.2
2.0 -1.0
7.0 7.0
3.0
4
15.0
4.5 6.5
0.77 0.68 0.07 0.15 0.23 0.33 0.27 0.0 0.06 0.23 0.62 0.65 0.50 0.41
-3.0 0.35 0.11 0.45 0.17 0.14
0 -3.0
0.76 0.68
2.06 2.98
#[Fragment_Memory_Table]
scaling_factor
mem_file
gamma_file
rmin rmax dr
frag_table_well_width
fm_use_pre-computed_table_flag
fm_sigma_exp
[Fragment_Memory_Table]
0.1

```
multi_mem_2.mem
uniform.gamma
0 50 0.1
0.1
0
0.15
# The Debye-Hückel parameters
pair_style hybrid/overlay coul/debye 0.0127 3.5 vexcluded 2 3.5 3.5
pair_coeff * * vexcluded 0.0
pair_coeff 1 1 vexcluded 20.0 3.5 4.5
pair_coeff 1 4 vexcluded 20.0 3.5 4.5
pair_coeff 4 4 vexcluded 20.0 3.5 4.5
pair_coeff 3 3 vexcluded 20.0 3.5 3.5
pair_coeff 4 4 coul/debye 80
```

2-) Tutorial to conduct the coarse-grained MD simulations on CaM and CaM/CaMKIIp interactions.

Calmodulin Project tutorial

This mini-tutorial shows step-by-step how to simulate the dynamics of calmodulin (CaM) using the Associative memory, Water mediated, Structure, and Energy Model (AWSEM) force field, implemented in LAMMPS software packages. It was written by Jules Nde, a graduate student from Dr. Cheung's research group.

1- Get the PDB structures of CaM

http://www.rcsb.org/pdb/explore/litView.do?structureId=1CLL

https://www.rcsb.org/structure/1prw

2- Check the availability of VMD and LAMMPS-AWSEM

From the Linux command line, type the following commands

```
$vmd        #to check VMD for visualization
$lmp_serial    #to check lammps-awsem
```

If VMD is not installed, please follow this link to install it

https://www.ks.uiuc.edu/Research/vmd/

If LAMMPS-AWSEM is not installed, please follow these links to install it

http://lammps.sandia.gov/download.html

https://github.com/adavtyan/awsemmd

3- Build lammps-awsem

Unzip both files awsem-master and lammps-master

```
$unzip awsem-master.zip
$unzip lammps-master.zip
```

Move to LAMMPS directory

```
$cd lammps-master/src
```

Copy the dependencies from awsem-master to the current (src) directory

```
$cp ~/awsem-master/src/*.h .
$cp ~/awsem-master/src/*.cpp
```

Type the following commands to build lammps-awsem

```
$make yes-molecul
$make clean-all
$make serial
```

If everything works well, you should get the executable file called “lmp_serial”

Troubleshooting: Make sure you have the last version of python and biopython installed as they will be required to prepare the input file

4- Prepare the input file for the simulations

Create a directory called CaM-AWSEM:

```
$mkdir CaM-AWSEM
```

Copy the structures to the new directory:

```
$cp 1cll.pdb CaM-AWSEM
$cp 1prw.pdb CaM-AWSEM
```

Move to the new directory:

```
$cd CaM-AWSEM
```

Copy all the necessary files form awsem-master to the current directory

```
$cp ~/awsem-master/parameters/* .
```

Create the input files

```
$~/awsem-master/tools/create_project_tools/PdbCoords2Lammps.sh 1cll CaM
```

After this step, you should get the following files: `data.CaM, CaM.seq, CaM.coord, and CaM.in`

Open CaM.in to add the electrostatics term

```
$vi CaM.in
```

Steps to add electrostatics term and assign the charges to the charged residues

`“pair_style hybrid/overlay coul/debye 0.1 3.50 vexcluded 2 3.5 3.5”`

- Identify the positively (Lys and Arg) and negatively (Asp and Glu) charged residues and group them. For example, positive_resid and negative_resid
    - “`group positive_resid id 27 51 78` …”
    - “`group negative_resid id 6 9 21` …”
- Assign the (+1) and (-1) charge to the positively and negatively charged residues respectively using the following commands:
    - “`set group positive_resid charge +1.0`”
    - “`set group negative_resid charge -1.0`”
- Create other groups with the charged residues in the calcium binding loops
    - “`group ef1 id 48 54` …” for ef1, ef2, ef3, and ef4
- Assign charges to the charged residues in the calcium binding loops according to their chemistry coordination
    - “`set group ef1 charge -0.5`”, similar for ef2, ef3, and ef4

5- Create the memory file

```
$python ~/awsem-master/tools/frag_mem_tools/Pdb2Gro.py 1cll 1cll.gro
$python ~/awsem-master/tools/frag_mem_tools/Pdb2Gro.py 1prw 1prw.gro
```

Make a directory called memory and copy the *.gro in that directory

```
$mkdir memory
$cp *.gro memory
```

Make a memory file and copy the following lines inside

```
$vi double_mem.mem
```

"memories"
"CaM"
"memory/1cll.gro 1 4 71 1"
"memory/1cll.gro 72 76 5 1"
"memory/1cll.gro 77 81 66 1"
"memory/1prw.gro 1 4 71 1"
"memory/1prw.gro 72 76 5 1"
"memory/1prw.gro 77 81 66 1"

Open the fix_backbone_coeff.data and provide the appropriate name of the memory file "double_mem.mem" in the memory table section

6- Run the simulation

```
$lmp_serial < CaM.in
```

This simulation should produce the following files: `log.lammps, energy.log, dump.lammpstrj`

7- Analyze the data

Convert the dump.lampstrj back to pdb format for visualization on VMD.

```
$python ~/awsem-master/tools/results_analysis_tools/BuildAllAtomsFromLammps.py
dump.lampstrj dump.pdb CaM.seq
```

Compute the pairwise comparison

```
$python ~/awsem-master/tools/CalcQValue.py 1cll dump.pdb Q_output_file
```

Plot the “Q_output_put” using R, or Matlab, or gnuplot, etc.

Compute the RMSD

```
$python ~/awsem-master/tools/CalcRMSD.py 1cll dump.pdb RMSD_output_file
```

8- Repeat step 4 in case of multiple memories

3-) Tutorial to conduct the coarse-grained MD simulations on the NHP10 module of the INO80 complex.

Coarse-grained MD simulations of the NHP10 module of the INO80 complex

This mini tutorial shows step-by-step how to conduct the MD simulation of the NHP10 module of the chromatin remodeling complex INO80 using the Associative memory, Water mediated, Structure, and Energy Model (AWSEM) force field, implemented in LAMMPS software package. It was written by Jules Nde, a graduate student from Cheung’s research group at the University of Washington.

1- Get the PDB structure, initial template, from the molecular docking

We use the software package “ClusPro” to get the initial template. The initial template is guided by the crosslinking data from Tosi et al., which shows the connectivity among subunits of the INO80 complex.

https://cluspro.bu.edu/publications.php

https://doi.org/10.1016/j.cell.2013.08.016

2- Check the availability of VMD and LAMMPS-AWSEM

From the Linux command line, type the following commands:

`$vmd` #to check VMD for visualization (VMD is not a requirement, but it will be useful to visualize the structures, although software can do the same job).

`$lmp_serial` #to check LAMMPS-AWSEM (It is now preferable to install the parallel version of the LAMMPS-AWSEM)

If VMD is not installed, please follow this link to install it,

https://www.ks.uiuc.edu/Research/vmd/

If LAMMPS-AWSEM is not installed, please follow these links to install it

http://lammps.sandia.gov/download.html

https://github.com/adavtyan/awsemmd

3- Build lammps-awsem

Unzip both files awsem-master and lammps-master

```
$unzip awsem-master.zip
$unzip lammps-master.zip
```

Move to LAMMPS directory,

```
$cd lammps-master/src
```

Copy the dependencies from awsem-master to the current (src) directory,

```
$cp ~/awsem-master/src/*.h .
$cp ~/awsem-master/src/*.cpp .
```

Type the following commands to build lammps-awsem

```
$make yes-molecul
$make clean-all
$make serial (use $make mpi instead to install the parallel version of the
LAMMPS-AWSEM)
```

If everything works well, you should get the executable file called “lmp_serial for serial and lmp_mpi for parallel”. This installation may require other dependencies to be installed first, please follow the error message(s) in case of any.

Troubleshooting: Make sure you have the last version of python and biopython installed as they will be required to prepare the input file (use the 2to3 in Linux terminal to convert the python scripts, which were written in python2 to python3).

4- Prepare the input file for the simulations

Create a directory called AWSEM_NHP10:

```
$mkdir AWSEM_NHP10
```

Copy the structures to the new directory:

```
$cp NHP10_M.pdb AWSEM_NHP10
```

Move to the new directory:

```
$cd AWSEM_NHP10
```

Copy all the necessary files form awsem-master to the current directory,

```
$cp ~/awsem-master/parameters/* .
```

Create the input files

```
$~/awsem-master/tools/create_project_tools/PdbCoords2Lammps.sh NHP10_M NHP10_M
```

After this step, you should get the following files: `data.NHP10_M`, `NHP10_M.seq`, `NHP10_M.cood`, and `NHP10_M.in`

Open CaM.in to add the electrostatics term,

```
$vi CaM.in
```

Steps to add electrostatics term and assign the charges to the charged residues,

```
“pair_style hybrid/overlay coul/debye 0.1 10.0 vexcluded 2 3.5 3.5”
```

- Identify the positively (Lys and Arg) and negatively (Asp and Glu) charged residues and group them. For example, positive_resid and negative_resid
    - “`group positive_resid id 27 51 78` …”
    - “`group negative_resid id 6 9 21` …”
- Assign the (+1) and (-1) charge to the positively and negatively charged residues respectively using the following commands:
    - “`set group positive_resid charge +1.0`”
    - “`set group negative_resid charge -1.0`”

5- Create the memory file

This part requires more cares as we must download all the available PDBs either in the protein databank or AlphaFold databank, that align with at least one part of the structure, and use them to guide the simulations (please see the memory directory, which is provided together with this tutorial, in the GitHub).

```
$python ~/awsem-master/tools/frag_mem_tools/Pdb2Gro.py NHP10_M NHP10_M.gro
```

Make a directory called memory and copy the *.gro in that directory,

```
$mkdir memory
$cp *.gro memory
```

Make a memory file and copy the following lines inside,

```
$vi NHP10_M.mem
```

Please see the example file, which is provided in the same directory with this tutorial.

Open the fix_backbone_coeff.data and provide the appropriate name of the memory file "`NHP10_M.mem`" in the memory table section.

6- Run the simulation

`$lmp_serial -in NHP10_M.in` for serial and `$lmp_mpi -n 4 -in NHP10_M.in` for serial. I provided a .slurm file to show how to run the same simulations on the HPC

This simulation should produce the following files: `log.lammps, energy.log, dump.lammpstrj`

7- Analyze the data

Convert the dump.lampstrj back to pdb format for visualization on VMD.

```
$python ~/awsem-master/tools/results_analysis_tools/BuildAllAtomsFromLammps.py
dump.lampstrj dump.pdb CaM.seq
```

Compute the pairwise comparison,

```
$python ~/awsem-master/tools/CalcQValue.py 1cll dump.pdb Q_output_file
```

Plot the “Q_output_put” using R, Matlab, gnuplot, python, or your desired plotting software.

Compute the RMSD

```
$python ~/awsem-master/tools/CalcRMSD.py 1cll dump.pdb RMSD_output_file
```

Compute the contact maps.

Use the Jupiter notebook, which is provided with this tutorial, to compute the contact maps. The contact maps verify is the different interactions match the ones observed experimentally, with crosslinking mass spectrometry.

3-) Python script to calculate the contact map.

```
from Bio.PDB import *
import os
import sys
import random
import time
from random import seed, randint
import argparse
import platform
from datetime import datetime
import imp
import numpy as np
import fileinput
from itertools import product
import pandas as pd
from scipy.interpolate import griddata
from scipy.interpolate import interp2d
import seaborn as sns
from os import listdir
import scipy
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.interpolate import griddata
import matplotlib as mpl
from brokenaxes import brokenaxes
# from .. import notebookFunctions

%matplotlib inline
plt.rcParams['figure.figsize'] = (10,6.180)    #golden ratio
# %matplotlib notebook
```

```
%load_ext autoreload
%autoreload 2

# one protein
data = np.zeros((161,161))
parser = PDBParser()
target = "/Users/jndekeng/Desktop/CaM_P_Umb_Samp/contact_aaa/20.s.pdb"
structure = parser.get_structure('target', target)
# Assume only one chain
model = structure[0]
# chain = model['A']
#chain = model.child_list[0]
# with open("contact.csv", "w") as out:
#     out.write("i, j, Distance\n")
for chain in model:
    for i_residue in chain:
        is_regular_res = i_residue.has_id('CA')
        if i_residue.get_resname() == "GLY":
            res_i = i_residue['CA']
        else:
            res_i = i_residue['CB']
        for j_residue in chain:
            if j_residue.get_resname() == "GLY":
                res_j = j_residue['CA']
            else:
                res_j = j_residue['CB']
            distance = (res_i.get_coord()[0] - res_j.get_coord()[0])**2
            distance += (res_i.get_coord()[1] - res_j.get_coord()[1])**2
            distance += (res_i.get_coord()[2] - res_j.get_coord()[2])**2
            distance = distance**0.5
            i = i_residue.id[1]
            j = j_residue.id[1]
            data[i-1,j-1] = distance
            out_line = str(i) + ", " + str(j) + ", " + str(distance) + "\n"

plt.imshow(data<7, cmap="Greys", clim=(0,1))
plt.colorbar()
```

VITA

Jules Berlin NDE KENGNE

2011 B.S. in Physics, University of Yaoundé 1, Cameroon

2013 M.S. in Physics option Materials Science, University of Yaoundé 1, Cameroon

2014 – 2018 Research Assistant at the University of Texas Medical Branch, Galveston, Texas, USA

2018 Graduate student in a Master of Engineering program, Materials Science and Engineering, University of Houston, Texas, USA (transferred)

2019 Graduate student in Physics – Ph.D., University of Houston; Center for Theoretical Biological Physics at Rice University, Texas, USA (transferred)

2019 – 2021 Research Assistant in Biological Physics, University of Houston, Texas, USA

2019 – 2021 Teaching Assistant in Physics, University of Houston, Texas, USA

2021 – 2024 Research Assistant in Biological Physics, University of Washington, Washington, USA

2021 – 2024 Teaching Assistant in Physics, University of Washington, Washington, USA

2024 Ph.D. in Physics, University of Washington, Washington, USA

FIELD OF STUDY

Physics option Biological Physics

PUBLICATIONS

- **Jules Nde**, Pengzhi Zhang, Neal Waxham, Margaret S. Cheung: Experiment and Simulation Reveal Residue Details for How Target Binding Tunes Calmodulin's Calcium-binding Properties JPCB (2023)
- **Jules Nde**, Pengzhi Zhang, Jacob C. Ezerski, Wei Lu, Kaitlin Knapp, Peter G. Wolynes, Margaret S. Cheung, Coarse-grained Modeling and Molecular Dynamics Simulations of Ca^{2+}-calmodulin, Front. Mol. Biosci. (2021) doi:10.3389/fmolb.2021.661322
- Karim Nchourupouo, **Jules Nde**, Yannick Wadop, Serge Zekeng, Bernard Fongang: Evolutionary Couplings and Molecular Dynamic Simulations Highlight Details of GPCRs Heterodimers' Interfaces Molecules (2023)
- Abhishek Mitra, Norbert Dojer, Bernard Fongang, **Jules Nde**, Yingjie Zhu, Maga Rowicka: Analyzing and interpreting DNA double-strand break sequencing data bioRxiv, (2020) doi: doi.org/10.1101/2020.03.05.977801
- Yingjie Zhu, Anna Biernacka, Benjamin Pardo, Norbert Dojer, Romain Forey, Magdalena Skrzypczak, Bernard Fongang, **Jules Nde,** Razie Yousefi, Philippe Pasero, Krzysztof Ginalski, and Maga Rowicka: qDSB-Seq is a general method for genome-wide quantification of DNA double-strand breaks using sequencing. Nature Communication, (2019) 10:2313
- A. Bianchi-Smiraglia, D. W. Wolff, D. J. Marston, Z. Deng, Z. Han, S. Moparthy, R. M. Wombacher, A. L. Mussell, S. Shen, J. Chen, D. Yun, A. Cox, C. M. Furdui, E. Hurley, M. L. Feltri, J. Qu, T. Hollis, **Jules Nde**, B. Fongang, R. Sousa, M. E. Kandel, E. S. Kandel, K. M. Hahn, and M. A. Nikiforov: Regulation of local GTP availability controls RAC1 activity and cell invasion. Nature Communication, (2021) 12:6091

- A Biernacka, Y. Zhu, M. Skrzypczak, R. Forey, B. Pardo, M. Grzelak**, Jules Nde**, A. Mitra, A. Kudlicki, N. Crosetto, P. Pasero, M. Rowicka, and K. Ginalski: i-BLESS: Ultra-sensitive detection of DNA double-strand breaks using agarose beads. Communication Biology, (2018) 1:181
- **Jules Berlin Nde Kengne**, Bernard Fongang, and Serge Zekeng, Structural properties of Fe/Cu magnetic multilayers: a Monte Carlo Approach. SPIN/2018/1850012
- Wei Shi, Therese Vu, Didier Boucher, Anna Biernacka, **Jules Nde** Pandita RK, Straube J, Boyle GM, Al-Ejeh F, Nag P, Jeffery J, Harris JL, Bain AL, Grzelak M, Skrzypczak M, Mitra A, Dojer N, Crosetto N, Cloonan N, Becherel OJ, Finnie J, Skaar JR, Walkley CR, Pandita TK, Rowicka M, Ginalski K, Lane SW, Khanna KK, Ssb1 and Ssb2 cooperate to regulate mouse haematopoietic stem and progenitor cell function by resolving replicative stress in vivo BLOOD/2016/725093.
- Y. Zhu, N. Dojer, A. Biernacka, B. Pardo, R. Forey, M. Skrzypczak, B. Fongang, **J. Nde**, R. Yousefi, G. Legube, P. Pasero, K. Ginalski, and M. Rowicka: Quantitative DSB sequencing (qDSB-Seq): a method for genome-wide accurate estimation of absolute DNA double strand break frequencies per cell, **bioRxiv**, 171405, doi:https://doi.org/10.1101/171405 (2017).
- Y. Zhu, A. Biernacka, B. Pardo, R. Forey, N. Dojer, **J. Nde**, B. Fongang, R. Yousefi, A. Mitra, J. Li, M. Skrzypczak, A. Kudlicki, P. Pasero, K. Ginalski, and M. Rowicka* (corresponding author): Integrated analysis of DSB patterns precisely reveals DSB mechanisms following replication fork collapse, **bioRxiv**, 171439, doi:https://doi.org/10.1101/171439 (2017).

Manuscripts in preparation

- **Jules Nde**, Jiayi Wang, Andrei Gasic, Jesse Fichman, Mihaela E. Sardiu, and Margaret S. Cheung: Integrated study of the subunits' organization of the INO80 complex: experimentally guided molecular simulations
- Pengzhi Zhang, Yossi Eliaz, Nathaniel Jennings, **Jules Nde**, Piotr Cieplak, Margaret S. Cheung: Chemistry-informed Machine Learning Enhances Explanation of Atomic Charges of Ca^{2+} in Proteins.