Title
Conformational Rheostats in Protein Folding and Binding: A Computational Study

Author
Nagpal, Suhani

Publication Date
2021

Peer reviewed|Thesis/dissertation

University of California
Merced

Conformational Rheostats in Protein Folding and Binding: A Computational Study

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Suhani Nagpal

2021

Chapter 2 is reproduced with permission from Nagpal S, Luong TDN, Sadqi M, Muñoz V. Downhill (Un)Folding Coupled to Binding as a Mechanism for Engineering Broadband Protein Conformational Transducers. ACS Synth Biol. 2020 Sep;9(9) 2427-2439. doi:10.1021/acssynbio.0c00190.

Chapter 4 (first half) is reproduced with permission from He Y, Nagpal S, Sadqi M, Muñoz V. Glutton: a tool for generating structural ensembles of partly disordered proteins from chemical shifts, Bioinformatics, Volume 35, Issue 7, 01 April 2019, Pages 1234–1236, doi.org/10.1093/bioinformatics/bty755.

The dissertation of Suhani Nagpal is approved.

Victor Muñoz

Ajay Gopinathan

Angel E. Garcia

Michael Colvin, Committee Chair

University of California, Merced

2021

To my parents, Neera and S.L. Nagpal

To Parika, my little sister

To Ashish, my husband, for there is music in our home always

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my PhD advisor, Prof. Victor Muñoz, for his invaluable supervision and continuous support over the past five years. His immense knowledge, vision, and expertise have encouraged me in my academic research. Thank you, Victor, for inviting me to be a part of your lab at UC Merced when I applied to join in Spain and allowing me to be part of multiple collaborative research projects in the lab. It has been an incredible learning experience and has prepared me well for the next transition in my scientific trajectory.

I am grateful to my committee members for their guidance, constructive comments, and suggestions across my various stages of research. I express my sincere appreciation to Prof. Michael Colvin for teaching the most engaging course on biomolecular simulations. It enhanced my fundamentals and aided my research. I would like to thank Prof. Ajay Gopinathan for his support. I have greatly benefited from his substantial efforts in graduate training through CCBM. I express my utmost gratitude to Dr. Angel Garcia for the insightful discussions about my research. I initially contacted Dr. Garica in 2014 while learning enhanced sampling techniques in India. His quick and detailed response to my technical queries has implanted in me what it means to be a scientist ever since.

Much of the work described in my dissertation was made possible only due to the perseverance and hard work of Thinh, a fellow graduate student, Dr. Yi He, a former postdoctoral scientist, and Dr. Mourad Sadqi, CCBM project scientist in Muñoz lab. Thinh, who leads the experimental aspect of our integrated research projects, contributed equally to Chapters 2 and 3, and we share equal authorship on the publication (s). Yi's work is highlighted at the beginning of Chapter 4, and I am a co-author on the publication. A great scientist and a friend, Mourad's upbeat demeanor has a positive impact in the lab. It has been a delight to work on these intriguing research projects, thank you all! I am grateful to my lab mates, Nivin, Rama, Ameed, Abhigyan and Thinh, it was a pleasure working

alongside them for 5 years. Being a member of such a skilled and intelligent group was both motivating and inspiring. Working at the Muñoz lab has made an indelible impression on me, and none of this would have been possible without their support.

I am thankful to the NSF-CREST CCBM center for providing opportunities for my growth as a researcher and professional development. In addition, I would like to express my appreciation to the Bioengineering department for building a thriving research environment, as well as Tomiko Hale and Becky Mirza for their continued support and assistance during my PhD program. My sincere thanks to Prof. Shahar Sukenik for letting me be the TA for the graduate course on computer simulations. I appreciate and acknowledge the custodian staff; their cheerful conversations were a constant source of encouragement. Special thanks to my Merced friends and colleagues, especially Nivin, Ameed, Rama, Thinh, Nargis, Suryabhan, Donglei, and Farnaz.

I am forever grateful to Dr. Kausik Chakraborty, Dr. Lipi Thukral and Dr. Koyeli Mapa at CSIR-IGIB, who motivated me to work on my first computational research project on protein folding. I grew tremendously as a researcher despite having little expertise with computation, and I would like to express my gratitude to Lipi for that. In addition, I am thankful to my friends and lab members at IGIB for their support and camaraderie. Also, I would like to convey my sincerest thanks to the faculty, research staff, and the friends I made at JUIT. I still think fondly of my time as an undergraduate at JUIT.

Finally, I acknowledge my wonderful parents and my sister, for their unconditional love and care. Their unwavering support and encouragement have been the constant pillar during my PhD. From building a model of the Hubble telescope with me and enrolling me in a S.P.A.C.E program because of my childhood passion for astronomy to instilling in me the most crucial life lesson of continuing to read books; my parents (Neera and S.L. Nagpal) have propelled my scientific journey, their talents and intellectual interest have shaped who I am today. Parika, my younger sister, is the most beautiful person inside-out I know. I can't

thank her enough for everything she does for our family; she inspires me to be a better person every day.

I owe thanks to an extraordinary person, my husband Ashish, for a decade of joyful adventures. His love, affection, music, infinite support, and understanding during my pursuit of PhD degree have made the completion of my dissertation possible. He brings out the best in me while also assisting me in keeping things in perspective. I greatly value his contribution in teaching me how to get started with Linux, Python and helping me troubleshoot codes. I also appreciate our dog that we are yet to adopt.

My sincerest gratitude to my beloved mother-in-law, and father-in-law (Neelam and Shiv Ram Yadav), for their love and moral support. Dr. Manish Yadav, my brother-in-law, for all the fascinating scientific and music chats.

My heartfelt thanks to my aunt (Geeta Nagpal), who helped me build an electromagnetic torch during my final year of school, which was a rewarding experience. My late grandparents, who believed in me and encouraged me to follow my dreams, their treasured memories are ingrained within me. Much love to Bhawna, my cousin, who has always inspired me after a long day with her witty nature and insight. Much love and appreciation to my friends, Tanmmay, Tripti, Surbhi, Adi, Pradeepika, Ankush, Gurpreet, Raghvi, Monty, Kanishk, Ritika and Ish, I will forever cherish their support and encouragement during my research journey.

Lastly, I would like to thank the GROMACS developer community for their technical support and their humorous remarks, it made learning enjoyable! In addition, I express my appreciation to the technical staff at the San Deigo Supercomputing facility for their assistance.

Vita

2016–2021	Graduate Student Researcher, University of California, Merced
2016–2021	Teaching Assistant, University of California, Merced
2013–2015	Senior Research Fellow, CSIR-Institute of Genomics and Integrative Biology , India
2007–2012	Integrated Bachelor and Master of Technology in Biotechnology, Jaypee University of Information Technology, India

Publications

T. D. Luong*, **S. Nagpal***, M. Sadqi, V. Muñoz "Molecular LEGO: An Approach to Map Out the Conformational Landscapes of Unbound Intrinsically Disordered Proteins", *In review*

S. Nagpal*, T. D. Luong*, M. Sadqi, V. Muñoz "Downhill (Un)Folding Coupled to Binding as a Mechanism for Engineering Broadband Protein Conformational Transducers", *ACS Synth. Biol. 2020*

Y. He, **S. Nagpal**, M. Sadqi, V. Muñoz "Glutton: a tool for generating structural ensembles of partly disordered proteins from chemical shifts", *Bioinformatics.* 2019.

S. Nagpal, S. Tiwari, K. Mapa, L. Thukral "Decoding Structural Properties of a Partially Unfolded Protein Substrate: En Route to Chaperone Binding", *PLoS Comput Biol.* 2015

Abstract of the Dissertation

Conformational Rheostats in Protein Folding and Binding: A Computational Study

by

Suhani Nagpal

Doctor of Philosophy in Bioengineering

University of California, Merced, 2021

Professor Victor Muñoz, Graduate Advisor

In order to execute their biological activities, most proteins fold into their unique, three-dimensional structure. The discovery of intrinsically disordered proteins (IDPs) about two decades ago, which are now widely found in eukaryotes, has since challenged the structure-function paradigm. IDPs, which in isolation exist as broad, non-random, conformational ensembles of interconverting states, are centrally involved in many biological processes. The key to their functioning is the ability to fold when bound to ligand partner(s), thus operating as morphing proteins. Despite booming interest in morphing behavior, investigating their structural transitions and mechanism remains extremely difficult because of their distinct characteristics.

Previously, we observed a close connection between intrinsically partially disordered proteins (IPDPS) and gradual (un)folding transitions of downhill folders, leading to the hypothesis that many IPDPs work as a conformational rheostat. The scope of this dissertation is to investigate the biological and technological implications of gradual conformational transitions. We first demonstrate the design principles of protein-based scaffolds by utilizing gradual (un)folding coupled to binding for developing rheostatic conformational transducers using computational modeling and experiments. Our engineered transducers showcase >6 orders of magnitude change in analyte concentration (broadband sensitivity) and have practical advantages over extant ones, which conventionally operate as con-

formational switches.

Next, inspired by the LEGO toy, we devised a novel modular approach to dissect the folding cooperativity and the energetic contributions of native interactions in defining the conformational ensemble and binding properties of IPDPs. Using an integrated strategy of computation and experiments, we perform an ensemble-based conformational analysis and find that the approach provides an exciting new tool for analyzing morphing transitions that should generally apply to any IPDP, thereby addressing a fundamental gap in the field.

One particularly interesting IPDP is NCBD that binds to multiple structurally diverse ligand partners and recruits the basal transcription machinery. We then explore the concept of NCBD functioning as a conformational rheostat, which allows its promiscuous binding. Finally, using extensive all-atom Molecular Dynamics simulations of NCBD and its biological partners in their free and bound forms, we decipher the hidden conformational biases in the dynamics of the heterogeneous ensemble of NCBD, undergoing gradual morphing transitions hinting at a working conformational rheostat in transcription.

CHAPTER 1

Introduction

I contain multitudes

Bob Dylan

Every protein has its own story, how it folds, its interactions, its biological function, and how it sometimes misfolds and causes disease. These tasks are mostly based on the protein's shape. Among proteins are a challenging set of shape-shifters under physiological conditions known as Intrinsically Disordered Proteins (IDPs). IDPs have a ton of character. Since their physicochemical characteristics are more undeveloped, they exist as ensembles of components that constantly change configurations behaving as morphing proteins. Nonetheless, as opposed to rigid proteins, IDPs are well advanced in their functional roles as they can form interactions with multiple binding partners in cells. Biology seems to find ways to leverage various aspects of physically feasible scenarios to attain desired outcomes. The higher prevalence of these IDPs in eukaryotes than in prokaryotes has led to an emerging hypothesis that as biological complexity grows, so does the content of intrinsic disorder, allowing for multitudes of crucial biological functions. Also, numerous IDPs are implicated in human diseases, including cancer, cardiovascular disease, and neurodegenerative diseases, and are largely considered undruggable. Despite booming interest in IDP behavior, investigating their conformational properties and mechanisms remain challenging because of their unique properties.

This dissertation research aims to (1) provide a new approach to dissect the folding landscapes of IDPS, (2) gain high-resolution mechanistic insights into their morphing coupled to binding behavior, and (3) develop molecular scaffolds

for protein engineering applications (biosensing). Overall, the research emphasis is on understanding the fundamentals of protein metamorphosis in the context of disordered proteins and their multiple binding modes. One particular area of interest within this topic is the functional consequence of the conformational rheostat (CR) mechanism in controlling eukaryotic transcription.

1.1 Background

One of the open fundamental questions in Molecular and Structural Biology is how molecular processes control protein folding. Also, a thorough understanding of this issue would have enormous practical implications for protein engineering and design. Different areas of protein science studies have uncovered protein behaviors that defy the classical structure-function paradigm over the last two decades [1]. In this regard, multiple comprehensive computational studies have shown that a significant fraction of the eukaryotic proteome is now believed to contain naturally unstructured domains in their functional states [2],[3]. Instead, they tend to fluctuate between an ensemble of short-lived conformations rather than adopting a unique structure, enabling them to perform essential functions in cells [4],[5],[6]. The key to their functioning is their promiscuous coupled folding and binding to structurally diverse ligand partners [6],[7],[8], [9] thus operating as morphing proteins. Also, mutations of disordered proteins are often implicated in a range of human diseases [10]. Due to their intrinsic disorder and conformational flexibility, it's a challenge to study morphing proteins, and currently, no effective drugs exist that target these proteins. Mechanistic understanding of morphing protein dynamics can aid in understanding the molecular bases of human diseases and in designing rational strategies to modulate IDP functions for therapeutic purposes.

1.1.1 Intrinsically Disordered Proteins

IDPs are a class of proteins that, in the native state, possess no well-defined structure, existing as broad, non-random, dynamic ensembles of conformations [11]. Extensive bioinformatics studies have concluded that naturally flexible proteins, instead of just being rare exceptions, are abundant in eukaryotes, estimating that over 40% of any eukaryotic proteome contains such disordered regions [2],[10]. IDPs fail to fold autonomously, primarily attributed to their amino acid compositions [12]. High contents in polar and charged amino acids, together with proline, are the typical sequence signatures of IDPs [13]. These segments are natively unfolded (>50 residues) and unable to fold cooperatively due to a lack of hydrophobic amino acids [14]. Intrinsic disorder has been implicated in various regulatory functions that require IDPs to interact with other ligand partners [7],[8]. Many of these interactions promote disorder to order transitions within IDPs or preserve their disordered state by forming "fuzzy" complexes [15], which have a high degree of structural heterogeneity. A recent study of two highly charged IDPs (H1 and Pro-Tα) using single-molecule Foster resonance energy transfer (sm-FRET) showed that very high, opposing net charges between the two result in increased binding affinity. Still, the complex remains devoid of specific interactions and intrinsically disordered with minimal structural changes [16]. This type of complex where the two IDPs remain unstructured even though tightly bound together widens the known spectrum of protein-protein interactions. Many IDPs are promiscuous binders, interacting with multiple ligand partners [17],[18] and studies have shown that such a complex network of interactions necessitates allosteric activity [19], in which disordered regions exhibit allosteric coupling between binding sites [20]. Many disordered proteins adopt different conformations upon binding to different target proteins, which allows the protein to fulfill more than one unrelated function, a property known as moonlighting [21]. For example, the nuclear coactivator-binding domain (NCBD) of the CREB-binding protein (CBP) adopts two distinct conformations when it binds to the activation domain of p160 nuclear receptor coactivators (ACTR) [17] or

to interferon regulatory factor 3 (IRF3) [22]. Although the NCBD is stabilized as a three-helix bundle in all of its complexes, the length of the helices and the packing topology vary significantly [23].

1.1.2 Gradual Protein (un)Folding

Downhill folding proteins are fast-folding proteins that lose or gain structure without crossing any significant free energy barrier, unlike two-state folding proteins that require a cooperative transition between the unfolded and native states [24]. This gradual structural disorder is a characteristic feature of downhill folding proteins as it dramatically decreases the magnitude of their folding time [25],[26]. These proteins also make attractive scaffolds for developing protein engineering applications due to their gradual (un)folding behavior, including the implementation of conformational transducers, which can be an essential component of an advanced biosensor [27]. The features of a conformational rheostat have been effectively demonstrated in a downhill folding module, BBL protein which is naturally sensitive to change in pH (at least three sites with titratable sidechains), working as a pH transducer, undergoing a gradual one-state unfolding accompanied by proportionate changes in proton binding affinity in microsecond folding times (folding rate of BBL) [28] as illustrated in Figure 1.1 [1]. In essence, conformational rheostats are protein domains that naturally populate a large, non-random conformational ensemble that gradually morphs into different structures in response to cues, such as ligand binding, which also remarkably applies to intrinsically partially disordered proteins (IPDPs). It has been widely known that IPDPs are unable to fold cooperatively unlike two-state folding proteins, which puts them at the extreme end of the cooperativity scale [29]. Therefore, a direct connection between gradual (un)folding and intrinsic disorder has been proposed and requires in-depth exploration [29],[30],[31].

[1] Adapted with permission from (Michele Cerminara, Tanay M. Desai, Mourad Sadqi, and Victor Muñoz Journal of the American Chemical Society 2012 134 (19), 8010-8013 DOI: 10.1021/ja301092z). Copyright © 2012, American Chemical Society

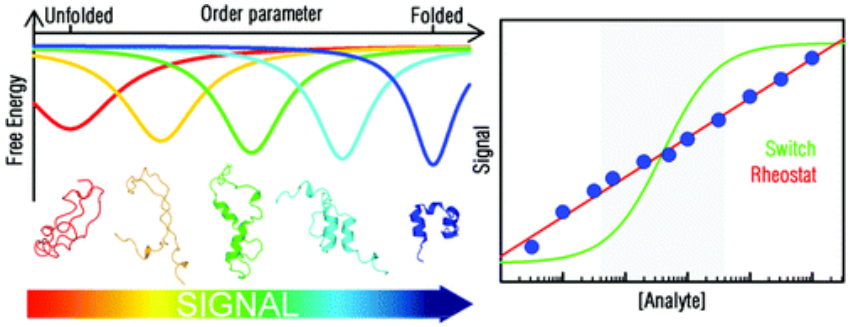

Figure 1.1: Left. Schematic of a molecular rheostat based on the coupling of a signal to the folding ensemble of a downhill folding protein. Right. Signal vs. analyte concentration for rheostat and switch mechanism. Figure reproduced with permission.

1.1.3 Coupled Folding and Binding

The interaction between folding and binding is one of the fascinating aspects of IDP: Is binding followed by folding or vice versa? Induced fit and conformational selection are two extreme mechanisms that have been proposed [32]. The protein binds to its binding partner in a largely disordered state, then folds to form the well-structured complex in the first mechanism. In the conformational selection mechanism, the ligand partner molecule protein 'selects' a preformed binding competent state that contains the structural features of the protein in the complex from the ensemble of conformations sampled by the IDP (when free in solution). Although it is expected that IDPs undergo some degree of conformational change even with a preformed structural element upon binding to the ligand partner [7]. The binding free energy of these complexes is used in part to promote folding of the IDPs. One of the most studied IDPs, pKID, has been shown to undergo induced folding coupled to binding in which the KIX domain forms a weak encounter complex that is stabilized by hydrophobic contacts between the two [8]. Given IDPs' dynamic nature and wide range of conformational properties, ranging from the lack of any structure to the formation of compact molten globules, it's likely that they use a variety of mechanisms or a combination of mechanisms [33],[34]. An extreme form of coupled folding and

binding occurs when both partners are more or less disordered in their free form but become well-ordered upon association. Such a type of mutual synergistic folding is well documented for NCBD:ACTR complex and NCBD binding to the transactivation domain of p53 (NCBD:p53-TAD) [9]. Furthermore, it has been shown that NCBD undergoes a gradual (un)folding behavior in its free form, i.e., the domain unfolds in a continuous manner without crossing a free-energy barrier, displaying characteristic unimodal probability distributions as a function of various structural properties [30]. As a result, in recent times a conformational rheostat mechanism has been proposed as a mechanism underlying the dynamics of partially ordered IDPs, as it describes the allosteric effects of coupled folding and binding [29].

1.1.4 Functional Advantages to IDPs

Intrinsic flexibility is advantageous. IDPs showcase the importance of conformational plasticity and heterogeneity in protein function [35]. It is mainly attributed to high specificity for multiple targets and low-affinity binding [36]. As a result, they frequently function as molecular hubs in protein interaction networks via modulation of allosteric interactions [37],[19]. Weak binding (and a fast dissociation rate) appeared to be especially important for signaling transduction, as it enables signals to switch quickly in response to ligand binding, post-translational modification, or changes in the cellular environment [38]. Due to their weak and multiple binding interactions, they can undergo phase transitions to form membrane-less organelles [39]. Furthermore, they help in the ordered assembly of macromolecules machines such as ribosomes and play key regulatory roles during the transcription process [6],[37],[40]. In fact, IPDPs are also suggested to function as molecular rheostats to support a continuum of conformational states and transitions tuned by diverse binding modes and post translational modifications that regulate dynamics between subpopulations and subsequent ligand binding [29],[41].

1.1.5 A Challenge to Study IDPs

Given their high degree of disorder, studying IDPs requires a combination of experimental, computational, and bioinformatics analyses to classify and characterize disordered regions and their mechanisms, resulting in a better understanding of their vital functional role in biological processes. Still, we do not have a complete understanding of IDP conformational propensities and their various binding modes, as it's a challenge to study experimentally. While small-angle X-ray scattering [42] and single-molecule FRET [43] can provide useful information on intra/inter molecular distance distributions, and NMR can provide both local and long-range structural information on free and bound states [44], IDPs exist as complex ensembles, making characterizing their structural properties extremely difficult. Moreover, current methods rely on conventional structural biology, because any observable will be averaged over a heterogeneous ensemble of structures. Many of these seminal experimental studies are discussed in detail in the following Chapters. Multiple methods and structural probes are needed due to the various time and length scales applicable to the roles performed by IDPs. Biomolecular simulations play an important complementary role in this scenario: sufficiently accurate molecular simulations can help predict experimental outcomes, allowing for better structural interpretation and mechanistic insights.

1.1.6 Computer Simulations

Molecular dynamics simulations (MD) have become a ubiquitous tool in modern life science. It leverages the laws of physics to understand the motions and behavior of biomolecules. Understanding biomolecular dynamics requires probing the system at biologically relevant timescales such as protein folding, ligand binding, etc. MD simulations use a classical Newtonian representation of atoms, molecules, and the forces between them are encoded in a classical force field that contains all the chemical specificity. Therefore, the quality of MD simulation results depends on the accuracy of the force field employed. All-atom simulations are increasingly used to obtain IDP conformational ensembles. Recent advancements in the force

field development allow MD to describe atomic-level properties of IDPs more accurately [45],[46]. These simulations can capture all the high-resolution microscopic details, such as local structure formation and protein motions, considering adequate sampling of the IDP heterogeneous ensembles [47],[48]. Furthermore, with the advent of GPU accelerated computing and the implementation of Particle Mesh Ewald (PME) to compute long-range electrostatic interactions on GPU nodes[49], one can employ multiple sets of MD trajectories of such morphing proteins to achieve better sampling statistics. Overall, MD simulations complement experiments in studying the structure, dynamics, and functions of IDPs [46],[50]. But simulating these systems also presents significant technical challenges.

1.1.7 Challenges of Molecular Dynamics Simulations

- **IDPs display high sensitivity to force field inaccuracies**

  The recent development of force fields with a primary focus on modifying the backbone and side-chain dihedral-angle potentials and critical evaluations of their performance when applied to IDPs show an improved description of IDP conformational propensities. In a simulation study of the disordered 24-residue arginine/serine peptide using seven different force fields, the conformational ensemble obtained using CHARMM22* (c22*) agreed best with all available experimental data [51]. The c22* ensemble performs best in this force field comparison: it has the lowest error in chemical shifts and J-couplings and agrees well with the SAXS data [52],[53]. Other studies have shown that the sampled disordered states are more compact than estimated from experiments for proteins >60 residues [54]. Furthermore, the effect of the solvent model has also been found to be important in the sampling of IDPs, and the water models are, in general, concurrently improved with progress in force field development. The four-site water models such as the TIP4P-D [55] with modified dispersion interactions have been shown to reproduce well the hydrophobic effect and water density in a wide temperature range that allows for more extended conformations. Further

improved force fields such as c36m have been shown to improve the structural properties of short disordered peptides [54], however it does not solve the problem of disordered proteins being too compact. The latest a99SB-disp has been reported to provide accurate descriptions of both ordered and disordered proteins [56].

- **Studying coupled folding and binding of IDPs**

  Simulating coupled folding and binding is challenging, but significant progress has been made, particularly through the use of native-centric models. Structure-based models (SBM) successfully conjugate the essence of the energy landscape theory of protein folding with computationally very efficient implementations [57]. These models have a coarse-grained representation, typically with a single interaction site per residue, located at the Cα position. A short-range attractive pair potential is used to describe interactions between residues in contact in the starting/native structure. All interactions between residues that are not in contact in the native state are then described using an excluded-volume repulsion term. The use of topology-based models allows us to circumvent the problems associated with sampling capability to reach the biological timescale of such processes (i.e., microseconds to milliseconds) and obtain statistically meaningful observations by sampling multiple reversible binding/folding events. These models have been successfully used to investigate many aspects of protein folding (complex conformational transitions) [41],[58],[59]. Many of these studies are discussed in the introduction of Chapter 5. However, coarse-grained methods lose the atomistic details of IDP structures. The SBM models, in particular, are unsuitable for describing the heterogeneous structures of unbound IDPs, making it difficult to investigate the role of non-native interactions in IDP dynamics. Deciphering the binding process using all-atom MD simulations is challenging due to the large number of degrees of freedom and the extensive conformational transitions involved. In this regard, the recent unbiased MD simulations produced by the Anton specialized

hardware, Robustelli et al. observed over 70 binding and unbinding transitions between the α-helical molecular recognition element (α-MoRE) of the intrinsically disordered C-terminal domain of the measles virus nucleoprotein (NTAIL) and the X domain (XD) of the measles virus phosphoprotein complex governed by induced folding pathways [60]. This promising study broadens the range of biological systems amenable to MD simulations.

1.2 NCBD

IDPs are known for their structural adaptability, which is the basis for their promiscuous binding. One particularly interesting IDP is the Nuclear Coactivator Binding Domain (NCBD) from the CREB Binding Protein (CBP), which has been the subject of many folding studies [22],[23], [30],[61],[62],[63],[64],[65]. NCBD is responsible for the interaction of the CBP with many other proteins to recruit the transcription machinery. It has more than ten known ligand partners, including the transactivation domain of p53 (p53-TAD) [23], ACTR [17], and IRF3 [22]. Their rapid dissociation facilitates a fast response of the co-activator machinery to this wide variety of regulators.

The NMR-determined structure of the free NCBD forms a compact, three-helix bundle structure but does not exhibit cooperative thermal unfolding shown in (Figure 1.2.A) [62],[66]. Recent multivariate analysis of experiments and computer simulations indicated that NCBD is a one-state downhill folder (i.e., a protein that unfolds gradually rather than cooperatively) [30]. In the IRF3 and ACTR or p53-TAD bound states (Figure 1.2.B), structures of NCBD show major topological differences, with significantly different configurations of the helices (Figure 1.2.C). ACTR and p53-TAD are also IDPs as demonstrated in (Figure 1.2.D). Several studies have characterized NCBD structure and its binding interactions. However, the in-depth mechanistic details of how it's proposed gradual (un)folding enables various binding modes remain mostly unexplored.

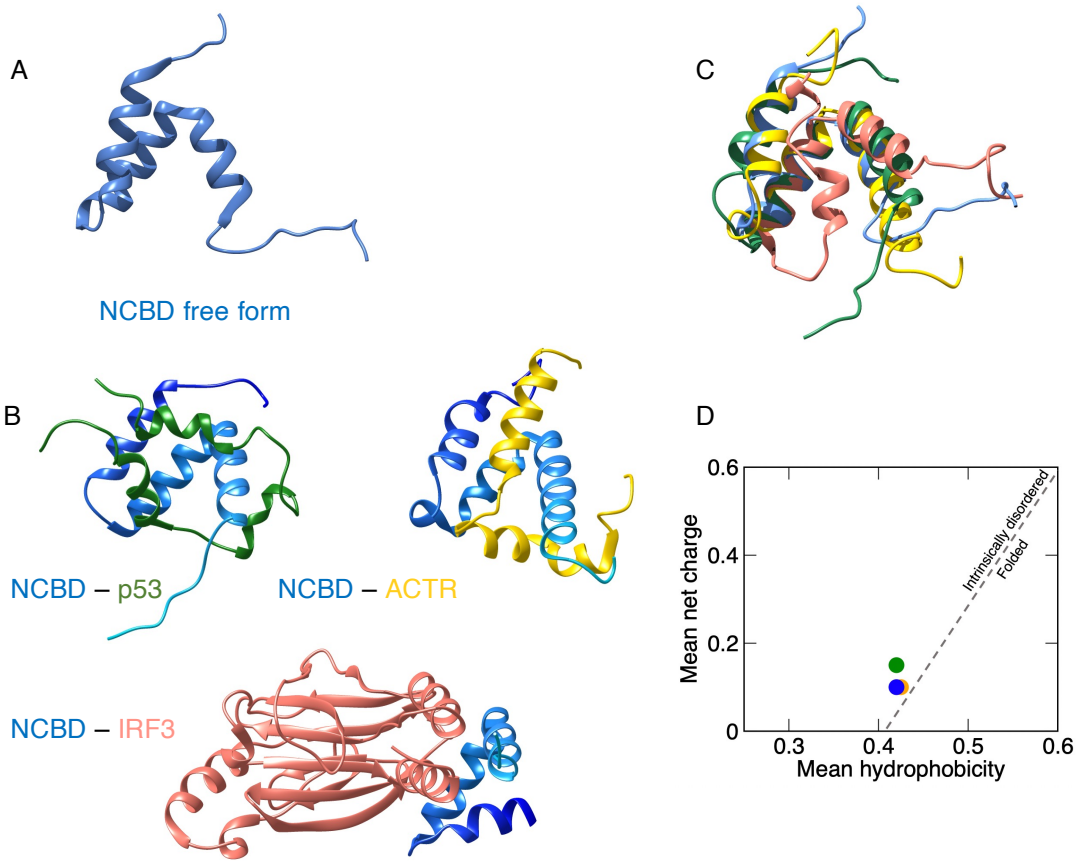

Figure 1.2: Experimentally determined A. NCBD free form structure (PDB ID: 2KKJ). B. NCBD bound to ligand partners: p53, ACTR, and IRF3 shown in cartoon representation. Dark to light blue represents N- to C-terminal. C. Structural superimposition of NCBD free form with bound conformers color-coded according to its respective ligand partner. D. Mean net charge versus mean hydrophobicity plot for three proteins; NCBD (blue), ACTR (yellow), and P53-TAD (green), showcasing distinction between IDPs and folded proteins based on their amino acid compositions.

1.3 Proposed Contributions

The primary contribution of my dissertation research are as follows:

- Downhill (Un)Folding Coupled to Binding as Mechanism for Engineering Broadband Protein Conformational Transducers. Here, we engineer pH transducers into the gradual (un)folding protein gpW, which is naturally pH insensitive. Particularly, we engineer histidine grafts into the gpW hydrophobic core to induce unfolding via histidine ionization. We design and test the effects of ionization via computational modeling and all-atom MD simulations, followed by experiments of the four promising scaffolds in collaboration. Our results demonstrate that the pH-dependent unfolding occurs in a rheostatic fashion and sense up to 6 orders of magnitude in [H+]. Notably, these properties make downhill (un)folding coupled to binding a powerful mechanism to engineer protein-based transducers that can be an important component of an advanced biosensor.

- Molecular LEGO: An Approach to Map Out the Conformational Landscapes of Unbound Intrinsically Disordered Proteins. Here, we devised a novel approach to measure the energetics driving the conformational-(un)folding landscape and the patterns of native interactions of partially disordered proteins. First, we dissect the morphing protein, NCBD, into all of its elementary components. Then, using an integrated strategy of computation and experiments, we perform an ensemble-based conformational analysis of all the components and establish the interactions between them by direct comparison of relevant components. This approach provides an exciting tool for analyzing morphing transitions that should generally apply to any disordered proteins, thus filling a fundamental gap in the field.

- Decoding Conformational Rheostats in Transcription: Morphing Coupled to NCBD Binding. Here, we explore dynamical properties of a morphing protein conformational ensemble, the nuclear coactivator binding domain (NCBD) that plays a vital role in organizing the eukaryotic transcription

complex by interacting with other morphing partners such as the transactivation domain of p53 (p53-TAD) and ACTR, or well-folded proteins like IRF3. Our analysis of $\approx$ 60 microseconds long, all-atom MD data on NCBD in the absence of partners shows a highly dynamic ensemble that populates sub-ensembles that resemble the various bound complexes in terms of topology and secondary structure hinting at a relatively unexplored conformational rheostatic behavior. We then decipher the interconversion timescales between its various sub-ensembles and find that the protein context modulates the interconversion between sub-ensembles and that NCBD samples pre-formed binding competent structures that bind to other morphing proteins (p53-TAD and ACTR) and folded protein (IRF3), providing high-resolution mechanistic insights into a working conformational rheostat in transcription.

- Dissecting the Interplay between NCBD Folding and its Ligand Partners. Finally, we assess the conformational properties of NCBD when bound to its various, structurally diverse ligand partners. We first study NCBD's morphing partners, p53-TAD and ACTR, and decipher their structural properties in their unbound forms. Next, we elucidate the interplay of NCBD bound to these morphing proteins and the well-folded protein, IRF3, by employing all-atom MD simulations of NCBD in the presence of its partners. This study shows that the NCBD ensemble retains a great deal of flexibility when bound (even to the stable, folded IRF3) and undergoes interesting transitions that reveal a subtle interplay between the conformational landscape of NCBD and partner binding. The most intriguing behavior emerges from the simulations in complex with other morphing proteins where the conformational modes of both partners appear intimately intertwined in nontrivial ways. This study provides the missing clues to interpret mechanistically the functioning of a conformational rheostat in recruiting the eukaryotic transcription complex.

CHAPTER 2

Downhill (Un)Folding Coupled to Binding as a Mechanism for Engineering Broadband Protein Conformational Transducers

2.1 Abstract

Canonical proteins fold and function as conformational switches that toggle between their folded (on) and unfolded (off) states, a mechanism that also provides the basis for engineering transducers for biosensor applications. One of the limitations of such transducers, however, is their relatively narrow operational range, limited to ligand concentrations 20-fold below or above their C50. Previously, our lab discovered that certain fast-folding proteins lose/gain structure gradually (downhill folding), which led us to postulate their operation as conformational rheostats capable of processing inputs/outputs in analog fashion. Conformational rheostats could make transducers with extended sensitivity.

In this chapter[1], we examine the extensibility of downhill (un)folding coupled to binding mechanism for engineering transducers with sensitivity over many orders of magnitude in ligand concentration (broadband). We investigate this hypothesis by engineering pH transducing into the naturally pH insensitive, downhill folding protein gpW. Particularly, we engineered histidine grafts into its hydrophobic core to induce unfolding via histidine ionization. We designed and tested the effects of ionization via computational modeling and studied experi-

[1]Reproduced with permission from Nagpal S, Luong TDN, Sadqi M, Muñoz V. Downhill (Un)Folding Coupled to Binding as a Mechanism for Engineering Broadband Protein Conformational Transducers. ACS Synth Biol. 2020 Sep;9(9) 2427-2439. doi:10.1021/acssynbio.0c00190. PMID: 32822536.

mentally the four most promising single grafts and two double grafts. All tested mutants become reversible pH transducers in the 4-9 range and their response increases proportionally to how buried the histidine graft is. Importantly, the pH-dependent reversible (un)folding occurs in rheostatic fashion, so the engineered transducers can detect up to 6 orders of magnitude in [H+] for single grafts, and even more for double grafts. Our results demonstrate that downhill (un)folding coupled to binding produces the gradual, analog responses to the ligand (here H+) that are expected of conformational rheostats, and which make them a powerful mechanism for engineering transducers with a broadband response.

2.2 Introduction

The engineering of protein folding/unfolding equilibria coupled to binding to a suitable ligand offers a generalizable strategy for developing biosensors that exploits the unparalleled specificity and selectivity of protein-mediated biomolecular recognition [67]. The strategy entails engineering the protein to be intrinsically unstable in the absence of ligand, and use the free energy provided by binding to the analyte in question (which only binds to the native structure) to trigger refolding, thus transducing the binding event into a monitorable output [68]. These transducers toggle between the unfolded-free and the folded-bound states given that their folding mechanism is usually two-state and the native structure is only formed upon binding the ligand, thus exemplifying the operation of conformational switches [69]. The result are binary signals and typically sigmoidal saturation curves that provide sensitivity to ligand concentrations within 20-fold below and above the apparent K_d or IC_{50} [70]. Another characteristic of such transducers is that their time response is ultimately determined by the rate of folding into their native state, which can take up to minutes for two-state folding proteins [71]. These characteristics make it challenging to produce protein transducers capable of broadband and/or real-time sensing, features that are often desired in biosensing applications.

In that regard, one exciting possibility is to use downhill folding proteins

as scaffolds for building transducers based on (un)folding coupled to binding. Downhill proteins fold and unfold in very short times (microseconds) [72], and change their structural properties gradually upon (de)stabilization [73], which results in broad, structurally heterogeneous (un)folding transitions [74],[75]. In fact, it has been proposed that the thermodynamic coupling between a biological signal and the gradual (un)folding of a downhill protein can result in conformational rheostats, a mechanism by which the protein produces analog responses to the input strength, e.g. ligand concentration [76]. The non-canonical features of downhill protein folding present a unique opportunity for building conformational transducers with broadband sensitivity and real time response.

Initially, our lab explored the merit of this idea on the BBL domain, a showcase of the most extreme, one-state downhill (un)folding behavior [73] and microsecond folding kinetics [77]. BBL folding is naturally pH sensitive due to two histidine residues that are partially buried within the protein core. A detailed study of the pH response of BBL has shown that this protein changes its structure gradually over four orders of magnitude in proton concentration and can record changes in pH with response times of a few microseconds [28].

2.2.1 Potential Protein Candidate for Engineering a de novo Conformational Transducer

The BBL study did not address the issue of whether such remarkable broadband response is extensible to other downhill (un)folding coupled to binding processes, or rather a unique result of natural selection on BBL. To determine the extensibility of the broadband response of transducers based on downhill (un)folding coupled to binding, and rationalize its structural/energetic determinants, we decided to de novo engineer pH transducing into gpW(W protein of bacteriophage lambda). GpW is a protein that folds into an $\alpha+\beta$ topology [78] and is stable and unaffected by pH in the 4-9 range [79]. Moreover, gpW folds and unfolds in microseconds, and exhibits the thermodynamic features of a downhill folding domain, including a minimally cooperative unfolding transition and multiprobe de-

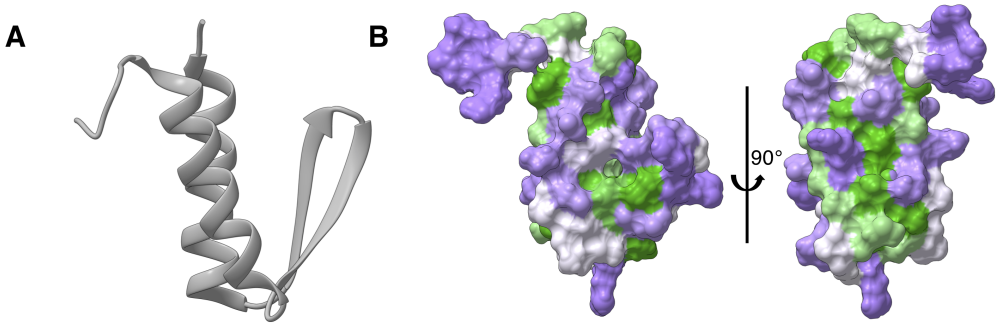

Figure 2.1: Structural features of gpW. (A) The protein is composed of 62 residues forming an all-antiparallel $\alpha+\beta$ topology consisting of one β-hairpin and two α-helices. (B) Molecular surface representation of the gpW native structure with color coding signifying the degree of hydrophobicity from polar (purple), intermediate (white) to hydrophobic (green). The two projections highlight the cores between helix-2 and the β-hairpin (left) and between the two helices (right).

pendence [80], as well as an atomically heterogeneous thermal unfolding process as probed both by NMR and long-time-scale molecular dynamics (MD) simulations [79]. Relative to BBL, gpW is considerably larger (62 vs. 45 residues), is a full gene product rather than an excised domain [80], and features an all-antiparallel fold with two distinct hydrophobic cores (Figure 2.1) that offers various structural loci for engineering pH transducing.

2.2.2 Strategy to Design Potential Scaffolds

To de novo engineer pH transducing into gpW we resorted to a histidine grafting strategy by which we introduce histidine residues into structurally targeted protein locations. Several groups have reported that the dual aromatic/ionic character of histidine and its pK_a value close to physiological pH (i.e. 6-6.5) can induce pH-dependent conformational changes in the native ensemble of a variety of proteins [81]-[82]. The role of histidine ionization as trigger of conformational transitions on proteins is also exploited functionally by Nature, like in the transmembrane protein OmpG, which opens and closes its central pore in response to

the (de)ionization of two histidine residues [83]. The molecular mechanism behind these conformational changes hinges on the ability of the histidine residue to interact favorably with a surrounding hydrophobic environment in nonionic form, and strongly destabilize the same environment when ionized (i.e. due to charge desolvation). Namely, a histidine that is buried into the core of a protein destabilizes the folded state upon ionization by an amount proportional to its drop in pK$_a$($\Delta\Delta G^{UN}_{+0} = 2.3026RT\Delta pK_a$).

Mutational analyses have shown that the pK_a shifts of buried histidine residues vary widely, being ultimately determined by the local environment in the native structure, including the degree of burial and the interactions with surrounding residues[84]. Practically, we implemented a computational/ experimental histidine grafting strategy in four steps: i) identification of structural loci in gpW suitable for accommodating a partially buried histidine graft; ii) mutation design in silico, followed by computational assessment of mutant stability through all-atom MD simulations in explicit solvent; iii) selection of mutations that do not drastically perturb the stability of the native fold in non-ionic form, and production in the lab; iv) computational and experimental analysis of the pH response of each select mutation.

2.2.3 The Protein gpW and pH Sensing as Model to Engineer Broadband Transducing

There are several reasons for using pH sensing for this study: 1) Proton binding-release is a relatively straightforward process to engineer into proteins using histidine grafts; 2) Histidine grafting allows for the introduction of multiple proton binding sites onto the protein scaffold as strategy for amplification or modulation of the transducer response; 3) Ionization-deionization processes are the fastest reactions in aqueous solution because proton transfer is orders of magnitude faster than conventional diffusion controlled processes [85]. Such an ultrafast binding process makes the transducer response time to be solely determined by the conformational transition of the protein, which takes place in microseconds for downhill

folders [86]. GpW does indeed fold and unfold in microseconds [80], but to be a suitable scaffold for this study it also needs to be naturally insensitive to pH in the neutral to mildly acidic range that is most biologically relevant (between 4 and 9). From an aminoacid composition viewpoint, gpW has a sole histidine (H15) that is located on the exterior of helix-1 and fully solvent exposed, and therefore unlikely to experience significant pK_a shifts.

2.3 Computational Methods

2.3.1 Design Strategy

To design mutations to histidine in core positions with varying degree of solvent exposure of the protein gpW (PDB ID: 2L6Q), we used the Chimera tool and DUET algorithm [87]. A combination of structural analysis to identify target locations and stereochemical criteria were used to identify conservative replacements to histidine (e.g. with enough room to accommodate the imidazole ring). Target mutation sites were ranked according to the predicted change in stability upon mutation calculated with DUET:

$$\Delta\Delta G^{UN}_{(M-WT)} = \Delta G^{UN}_{M} - \Delta G^{UN}_{WT} \tag{2.1}$$

The six single point mutations to histidine were designed with the Chimera tool and refined via energy minimization. The fully solvent accessible histidine 15 in gpW was also replaced to Ala to investigate the effects of ionization of the natural histidine. As further computational test of the intrinsic native stability, all-atom MD trajectories in explicit solvent were run for each mutant in deprotonated form as well as for wildtype gpW (see below).

Electrostatic Potential Calculations The electrostatic potential maps of gpW and its mutants were calculated at physiological conditions (pH-7) using the adaptive Poisson–Boltzmann solver (APBS) [88]. The input PQR files were generated by the PDB2PQR server using the PARSE force-field and the protonation states were assigned by PROPKA. The grid dimensions were automatically

set by APBS based on the input structure. Electrostatic potentials were calculated by solving the linear Poisson– Boltzmann equation with a single DH sphere boundary condition. The solvent accessible surface area was calculated using a solvent radius of 1.4 Å.

2.3.2 All-Atom MD Simulations

MD simulations were performed in the GROMACS suite [89] using the OPLS all-atom force field [90]. Water molecules were modeled with the TIP4P representation [91]. Periodic boundary conditions were used, and long-range electrostatic interactions were treated with the Particle Mesh Ewald (PME) summation using a grid spacing of 0.16 nm combined with a fourth-order cubic interpolation to derive the potential and forces in-between grid points [92]. The real space cutoff distance was set to 1.0 nm and the van der Waals cutoff to 1.0 nm. The bond lengths were fixed [93] and a time step of 2 fs was used for numerical integration of the equations of motion. Coordinates were recorded every 10 ps. The simulations were performed at 310 K starting from the coordinates of the lowest energy conformer in the gpW NMR structural ensemble modified to carry the mutations to histidine and the H15A pseudowildtype. The protein was placed in a dodecahedral water box large enough to contain protein and at least 1.0 nm of solvent on all sides (volume $\approx$ 233 nm^3). The structure was solvated with 7,300 water molecules, and 4-5 Cl- ions were added to neutralize the system. The starting structure was subjected to energy minimization using the steepest descent method. All systems were equilibrated at a constant temperature of 310 K utilizing the two-step ensemble process (NVT and NPT). First, the system was subjected to NVT (constant number of particles, volume and temperature) equilibration for 100 ps with the position of the protein restrained, followed by NPT (constant number of particles, pressure and temperature) equilibration for another 100 ps. The simulations were subjected to the modified Berendsen thermostat with 0.1 ps relaxation time to maintain the exact temperature [94], followed by Parrinello-Rahman [95] with 0.2 ps relaxation time for pressure coupling

at 1 bar before the production run was started. All the simulations were run on the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputing center (SDSC). The total simulation time per variant was 1 μs for the wildtype, 0.75 s for the mutants L7H, M18H, F35H and V40H, and 0.4 μs for A10H and A13H. MD Simulations of Mildly Acidic Conditions. All the protonated mutant MD trajectories (L7H+, M18H+, F35H+ and V40H+) were run for 2.2 μs. In absence of well-established methodology to define and control pH during MD simulations (simulations do not contain free hydronium ions and protons cannot be exchanged in classical MD) we altered the protonation state of all titratable residues before the MD run based on their estimated pK_a values and relative to a nominal pH of 5. The ionization states were kept constant for the entire MD run. Histidine protonation was carried out using the pdb2gmx tool (HISE type) by first performing a standard pK_a calculation of the starting structure using DelphipK_a.

MD Simulations of Double His-Grafts MD simulations were carried out in deprotonated and protonated form for two double histidine grafts: M18H-F35H and F35H-V40H. Each double mutant was simulated for 1 and 2.2 μs in the deprotonated and protonated forms, respectively.

2.3.3 Analysis of MD Trajectories

The root-mean-square fluctuations (RMSF) per residue were calculated for each MD trajectory (protonated, deprotonated and wildtype) using the gmx rmsf tool. The RMSF of each mutant trajectory was expressed in reference to the wildtype as ΔRMSF (RMSF$_{mut} - RMSF_{wt}$).

Conformational Landscapes of gpW Mutants Maps were obtained from the normalized probability distribution as a function of the relevant set of order parameters. The probability distribution was converted into an energy scale using the following expression:

$$\Delta A_{ref \to i} = -RTln(\frac{P_i}{P_{ref}}) \tag{2.2}$$

where the probability of going from a reference state (ref) of the system to any state i (e.g., from folded to unfolded) at constant temperature and constant volume is evaluated. R is the ideal gas constant, T is the temperature and p_i and p_{ref} are the probabilities of finding the system in state i and state ref, respectively. We project the conformational space onto two order parameters: the radius of gyration (Rg) and the fraction of native backbone contacts (Q). In this calculation a contact is considered formed when the minimum pairwise distance between atoms of the interacting residues is <=0.55 nm and the residue pair is >3 residues apart in the protein sequence. Conformations collected at 100 ps intervals were projected onto the Q-Rg plane using a 40x40 grid (1600 cells) and sampling statistics were compiled to evaluate Equation 2.2 . The grid cell with the largest population was used as reference state.

2.4 Computational Results

2.4.1 Design Principles of Conformational pH Transducers Based on Histidine Grafting

The imidazole ring of histidine is ionizable with a standard pK_a around 6.5. When the imidazole is deprotonated its aromatic character predominates, and thus it can form stabilizing interactions with neighboring hydrophobic residues within the protein core. On the other hand, a protonated histidine located in a buried position destabilizes the native structure due to the large energetic penalty involved in desolvating the charge. This net destabilization shifts the effective pK_a to lower values (i.e. needing higher proton concentrations to become ionized). The larger the pK_a drop, the stronger the destabilization of the native state that is induced by histidine ionization, which can eventually drive protein unfolding once the destabilization is comparable to the intrinsic stability of the native state [96]. In this regard, as most downhill folding domains [76], gpW's native

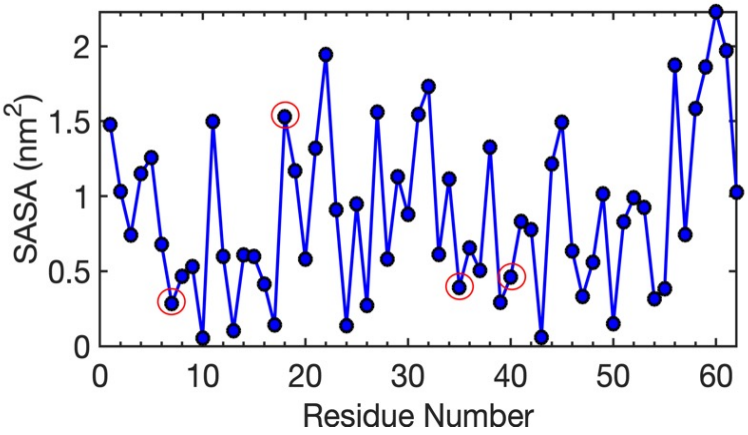

Figure 2.2: Solvent accessible surface area per residue of the gpW native structure calculated using UCSF Chimera tool with 0.14 nm solvent probe. Red circles indicate each of the four selected mutations.

state has relatively low intrinsic stability, i.e. about 14 kJ/mol [80]. There are two important implications that emerge from these considerations: 1) Histidine grafts should be structurally conservative to avoid excessive destabilization of the gpW scaffold that could result on a protein that is unfolded over the entire pH range; 2) The grafts should still be sufficiently buried (and experience pK_a downshifts) to be able to trigger unfolding upon protonation at mildly acidic pH. With these principles in mind, we used structural analysis to select the locations on gpW for histidine grafting. The recipient sites are residues that participate in one of the two gpW hydrophobic cores (Figure 2.1) and have enough space to accommodate the imidazole ring into the cavity without introducing significant steric clashes. We identified six such positions: A10, A13, L7, M18, F35 and V40 (see Methods for more details). The six locations have varying degrees of solvent exposure (Figure 2.2), hence, providing us with flexibility to engineer different pH responses and explore how to maximize the transducer dynamic range.

We then designed the histidine mutations in silico and evaluated the effect on the stability of gpW using the DUET algorithm (Table 2.1). The calculations with DUET indicated that histidine substitutions into the L7, A10, and A13 positions could reduce the native stability of gpW at room temperature by more

Mutant	ΔΔG
L7H	-6.02
M18H	2.26
F35H	-0.17
V40H	-2.01
A10H	-6.65
A13H	-6.40

Table 2.1: Change in stability (Equation 2.1) in kJ/mol of gpW single histidine mutants as predicted by the DUET algorithm.

than half, potentially placing these grafts at the brink of stability even at neutral pH.

2.4.2 Molecular Dynamics Analysis of Histidine Graft Stability

We then used atomistic MD simulations to investigate the intrinsic destabilization induced by the histidine grafts in deprotonated form. Particularly we simulated each of the six mutants and the wildtype for 400 ns. Figure 2.3 shows the time trajectories of the root mean square deviation (RMSD).

Given the marginal stability and ultrafast folding of gpW, we expected these relatively short trajectories to display significant structural fluctuations and possibly even global unfolding. The control trajectory on wildtype gpW does show distinct structural transitions that take place within the first 100 ns, followed by stabilization onto a relatively low RMSD ensemble. These fluctuations correspond mostly to the b-hairpin flapping in and out from its interaction with the two helices, which remain closely in contact throughout the simulation. The lower stability and enhanced structural dynamics of the gpW hairpin have been reported before from NMR analysis and long timescale MD simulations [79]. Simulations of M18H, F35H and V40H showed minimal structural fluctuations throughout the entire trajectory with RMSD below 0.4 nm throughout (Figure 2.3). L7H displays larger structural fluctuations than the other three grafts, and signifi-

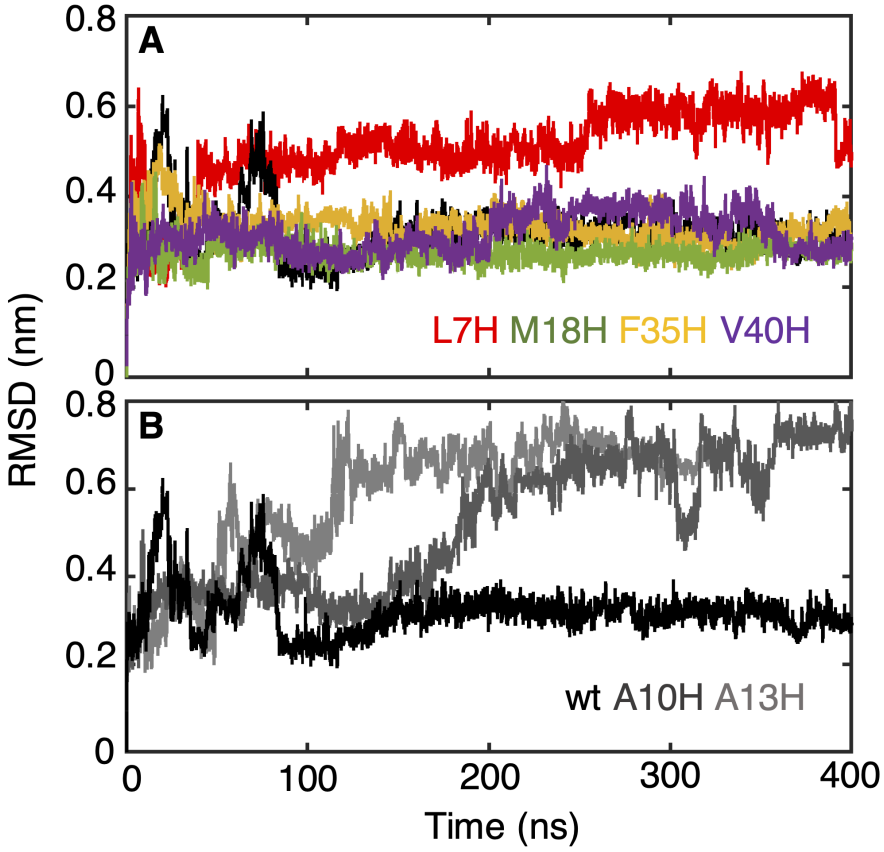

Figure 2.3: Time evolution of the root-mean-square deviation (RMSD) relative to the lowest energy conformer of the gpW NMR ensemble from MD simulations of gpW (black) and the six designed single (deprotonated) histidine substitutions. (A) Trajectories of the four mutants that show structural fluctuations below the threshold (0.65 nm). (B) Trajectories of the mutants that exceeded the 0.65 nm RMSD threshold.

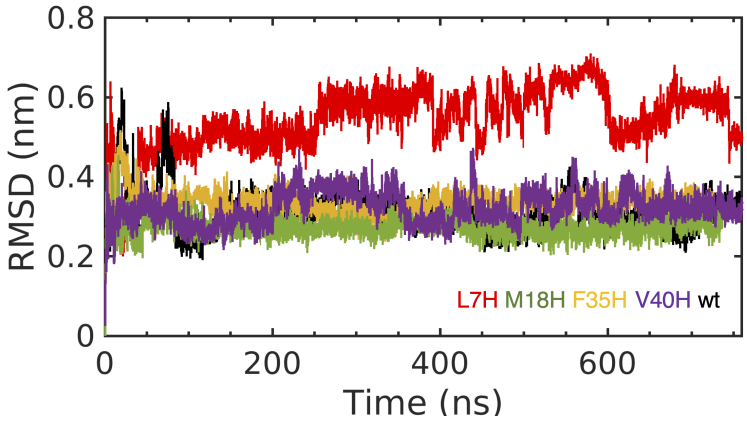

Figure 2.4: Time evolution of RMSD for the MD trajectories of the wild-type and four selected mutants.

cantly higher mean RMSD (about 0.55 nm), which is consistent with the large destabilization predicted by DUET (Table 2.1) and its fully buried location in gpW (Figure 2.2).

The structural fluctuations of L7H are more frequent, but they still are comparable in magnitude (< 0.65 nm) to those experienced by the wildtype. Moreover, as in the wildtype, the structural fluctuations of L7H concentrate on the β-hairpin. These results suggest that L7H is probably a viable graft. In contrast, A10H and A13H show even larger structural fluctuations that get close to 0.8 nm and which may not be fully equilibrated after 400 ns (Figure 2.3B). Moreover, the conformational fluctuations of A10H and A13H involve the entire structure, which is again consistent with the large destabilization predicted by DUET (Table 2.1) and the more aggressive design of these mutations (introducing a bulky imidazole at a core location where there was only a methyl group). Therefore, we ruled out the A10H and A13H grafts for further study. For the other four grafts we extended the simulation time up to 750 ns (Figure 2.4) to ascertain whether the structural fluctuations (especially on L7H) would stabilize or continue evolving towards more unfolding. The longer trajectories showed small-scale reversible transitions with signs of stabilization around their characteristic mean values. From these combined results we decided to focus on L7H, M18H, F35H and V40H as select grafts.

2.4.3 The Structural Environment of Select Grafts

Figure 2.5 shows the structural environment of the four gpW sites selected for histidine grafting. This figure highlights that L7 is buried and surrounded by non-polar residues, but because leucine is bulky, mutation to histidine results on minor structural changes. M18 is solvent accessible, but it is surrounded by non-polar residues, and the grafted histidine sits at the exposed/buried interface. F35 is embedded into the protein core formed between helix-2 and the β-hairpin. However, the designed histidine has plenty of room to fit the imidazole into the cavity left by the removed phenol. The mutation does affect the local electrostatic potential reflecting the slightly more polar nature of the imidazole ring. V40 is partially buried, and accordingly, its replacement by histidine results on a conservative mutation on a semi-exposed core position with a slight increase in neighboring interactions of the bulkier sidechain.

2.4.4 Enhanced Structural Fluctuations upon Histidine Ionization

We then examined the structural effects induced by histidine protonation via MD simulations in which the grafted histidine was kept protonated throughout the entire trajectory (see Methods). Particularly, we produced 2 μs of MD simulation for each of the protonated mutants. Relative to trajectories where the histidine is deprotonated, the protonated trajectories reveal generalized conformational rearrangements that take place in the sub-microsecond timescale. As tool to examine the structural flexibility of the protonated and deprotonated trajectories, we calculated the time averaged root mean square fluctuations (RMSF) of each protein residue along the trajectory. As reference, we performed the same analysis on the wildtype trajectory and calculate the RMSF difference relative to the wildtype (hence positive values imply that the residue is more flexible/unstructured than in the wildtype). The results of these calculations are shown in Figure 2.6. The deprotonated simulations indicate that M18H, F35H and V40H are as conformationally stable in the sub-μs timescale as the wildtype. In contrast, MD simulations in protonated form show marked increases in structural fluctuations

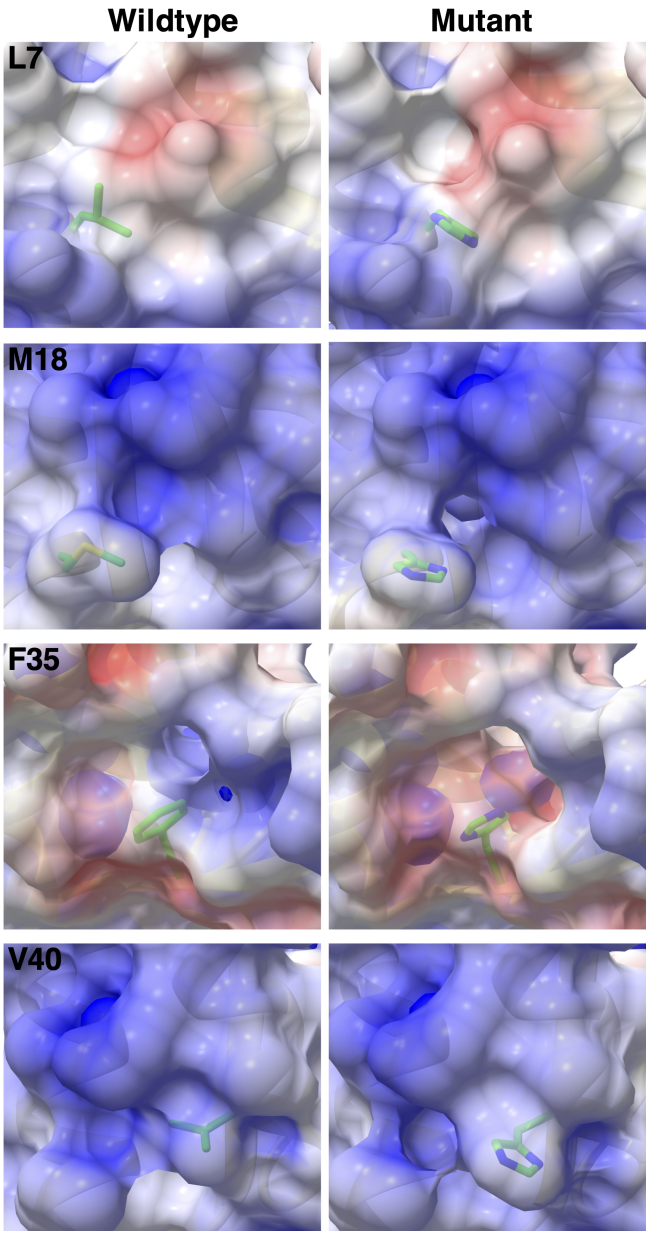

Figure 2.5: Variability in structural environment of selected histidine grafts. The left panels show the structural environment surrounding the four sites in gpW selected for histidine mutation, and the right panels the modeled environment after introducing the mutation in silico. The environment is depicted using a molecular surface representation colored according to the electrostatic potential and the sidechain of the specific residue (wildtype or histidine) is shown in stick representation color coded according to atom type.

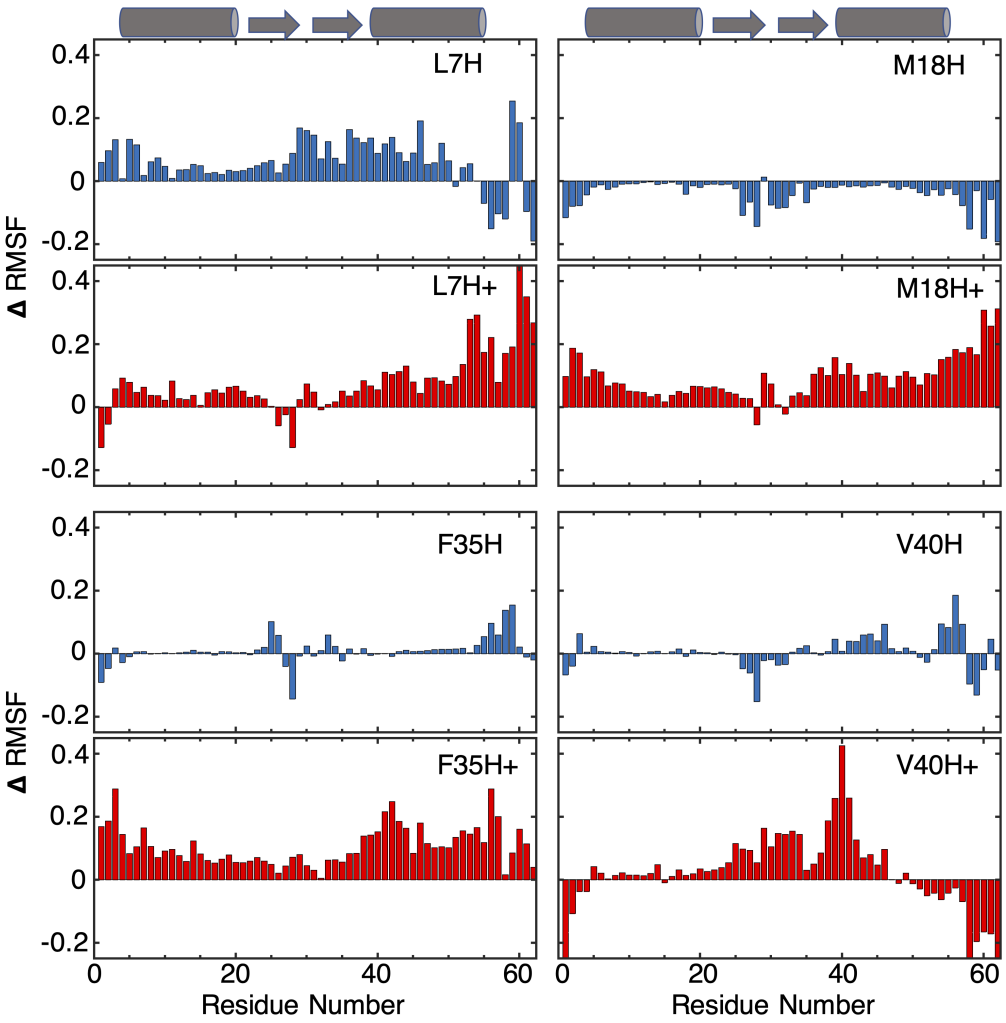

Figure 2.6: MD analysis of the structural fluctuations of gpW histidine grafts. The plots provide the difference in root-mean square fluctuations (RMSF) per residue for each histidine graft. Blue represents the deprotonated simulations and red the protonated simulations relative to the wildtype trajectory.

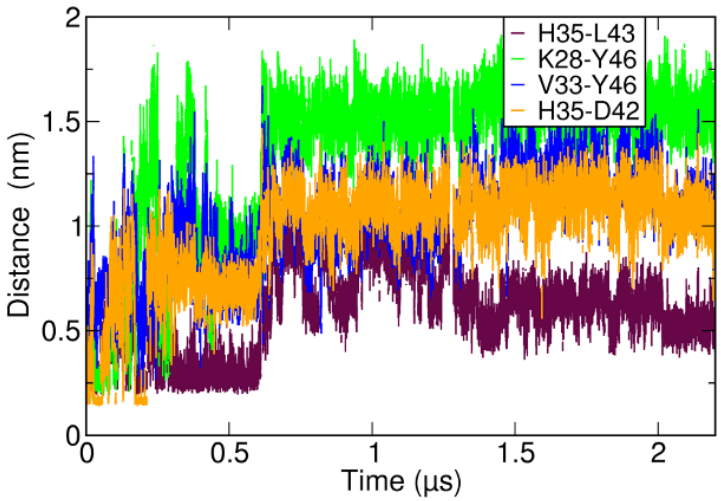

Figure 2.7: Evolution of residue-residue interacting pairs between β-hairpin and α-helix-2 along the F35H+ trajectory.

on the four grafts. For instance, M18H+ loses tertiary interactions between the two α-helices, which triggers global unfolding. F35H+ show similarly enhanced overall dynamics with increased flexibility of the residues in close contact close contact with H35 (helix-2 residues 40-50). The F35H+ trajectory shows the rupture of all tertiary interactions between the β-hairpin and helix-2 (the second gpW mini-core, Figure 2.1) via a sharp transition at $\approx$ 600 ns (Figure 2.7).

The V40H+ trajectory reveals a localized pattern of enhanced structural fluctuations around the mutated site that propagates into the hairpin, but the interactions between the protein termini appear stabilized. Finally, the L7H+ ensemble shows enhanced fluctuations and partial unfolding, mostly of the end of helix-2 (residues 50-60), which forms tertiary interactions with helix-1 (where L7H is) that are weakened by histidine protonation. The most notable feature in L7H+ is the opening of the hydrophobic core formed between the two helices in the native structure (see Figure 2.8).

Overall these results provide computational evidence that there is effective thermodynamic coupling between histidine ionization and gpW unfolding in the four cases, which suggests that the grafts should be able to conformationally transduce changes in pH.

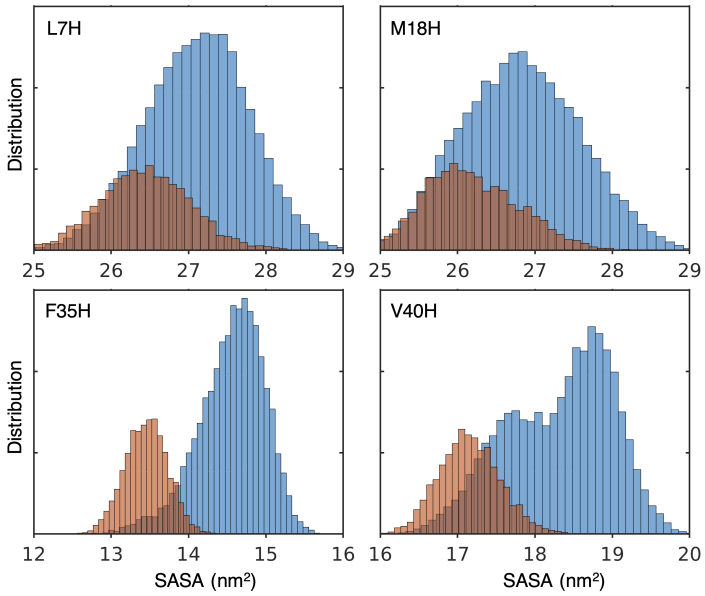

Figure 2.8: Solvent accessible surface area distribution. Deprotonated and protonated ensembles are in orange and blue respectively. For each respective mutant, only the local environment, defined as the neighboring secondary structure elements is used for the calculation: L7H and M18H consider the two α-helices; F35H and V40H, consider the β-hairpin and α-helix-2.

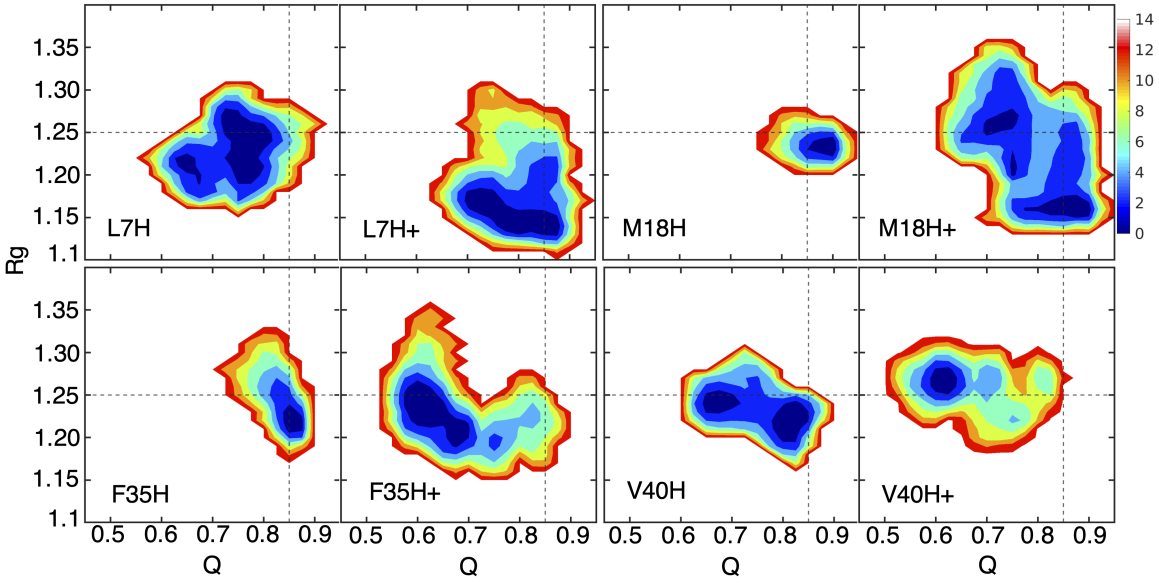

Figure 2.9: Conformational landscapes projected onto Q and Rg of the deprotonated and protonated trajectories at 310 K for each gpW histidine graft. The crossed dashed lines signal the wildtype gpW minimum for reference. The color bar is in kJ/mol.

2.4.5 Conformational Landscapes of Neutral and Ionized Histidine Grafts

One characteristic of downhill folding is a gradual, minimally cooperative, unfolding behavior, which by face value seems consistent with the patterns that we see in MD simulations. To further investigate the nuanced effects of histidine grafting/ionization on gpW we analyzed the trajectories in terms of projections onto two widely used order parameters: the fraction of native backbone contacts (Q) and the radius of gyration (Rg). Q informs on the overall degree of native structure that is made, and Rg informs on the degree of overall compaction of the polypeptide. The resulting conformational landscapes are shown in Figure 2.9.

Even though conformational sampling might be somewhat limited, the projected landscapes are consistent with our previous conclusions from the RMSF analysis and provide further mechanistic insights. For instance, L7H significantly destabilizes the native ensemble, resulting on a decrease in Q relative to the

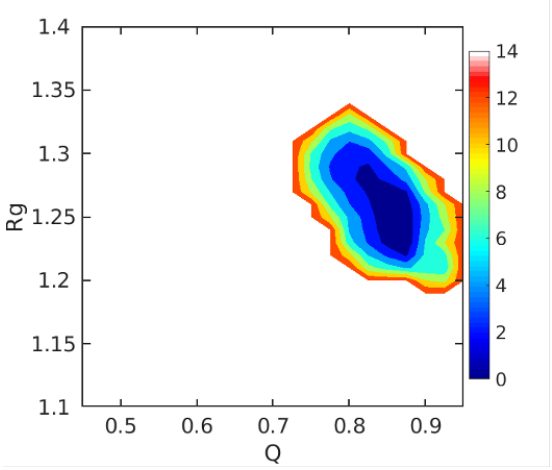

Figure 2.10: Free energy contour map of wildtype gpW as a function of Q and Rg. The color bar denotes the Gibbs free energy in kJ/mol.

wildtype's behavior (Figure 2.10). Interestingly, protonation (L7H+) shifts the ensemble to slightly more native Q, but at the same time it makes it much more compact (lower Rg). This is so because breaking the helix-helix core exposes hydrophobic surface, disorganizes the local helical conformation, and favors the collapse of the hairpin onto the newly exposed surface. M18H behaves quite differently. In neutral form the ensemble is very similar to that of the wildtype. However, protonation produces a substantial loss in native contacts that points to extensive unfolding.

The conformational ensemble of F35H is native-like in neutral form, but it experiences the most dramatic unfolding induced by protonation with Q $\approx$ 0.55. We also noticed an increase in solvent-accessible surface area in the surrounding structural elements, which highlights the breakage of the two protein cores (Figure 2.8). The V40H graft has enhanced dynamics in neutral form, and extensive unfolding upon protonation, manifested by the loss of about 40% of the native contacts and a slightly more expanded ensemble (larger Rg). These observations further confirm the coupling between histidine ionization and native unfolding in these grafts. They also reveal that gpW unfolding starts locally and propagates from the graft to neighboring areas. Therefore, our computational results suggest that the engineered pH induced unfolding of gpW is gradual, as expected for the

downhill (un)folding scenario.

2.5 Experimental Result

Here we use the CD signal at 222 nm at room temperature as indicator of the transducer response. The data for the four grafts is given in Figure 2.11. This figure demonstrates that all single grafts are conformational pH transducers with sensitivity at pH > 4, whereas the wildtype is insensitive in that range. Therefore, the histidine grafting approach works as general strategy to engineer conformational pH transducers into proteins. Figure 2.11 also highlights the broadband behavior of these transducers [2].

[2]Performed by Thinh D.N. Luong in Muñoz Group

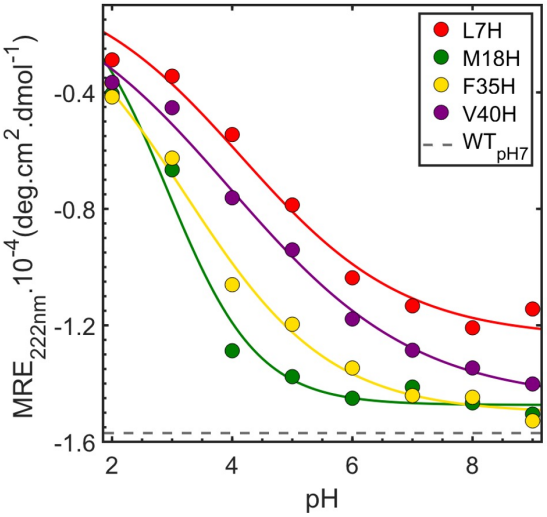

Figure 2.11: CD signal at 222 nm as a function of pH of all the single mutants at 298 K (circles) and their corresponding colored lines to guide the eye. The dashed line (–) represents the CD signal of wildtype gpW in its folded state (pH 7) as reference.

2.6 Modulating the Transducer Mechanism via Multiple Grafts

Our results indicate that the response of the pH transducer depends on the structural environment of the histidine graft: more buried positions lead to broader dynamic range. The question remaining is whether the effects of more than one graft are additive or exhibit positive or negative cooperativity. We decided to explore this issue by producing double grafts.However, we had to be careful as the single gpW grafts are already marginally stable. We ruled out L7H since it is already at the brink of native stability, and targeted M18H-F35H as a conservative double graft and F35H-V40H as a more aggressive one (Figure 2.11). As we did for the single grafts, we analyzed the conformational behavior of the double grafts. The increased perturbation without ionization of the double mutants is evident in the RMSF analysis (Figure 2.12 top). Ionization of the double grafts produces enhanced RMSF, particularly for F35HV40H, in line with what we expected from our experimental data on the pH response of the single mu-

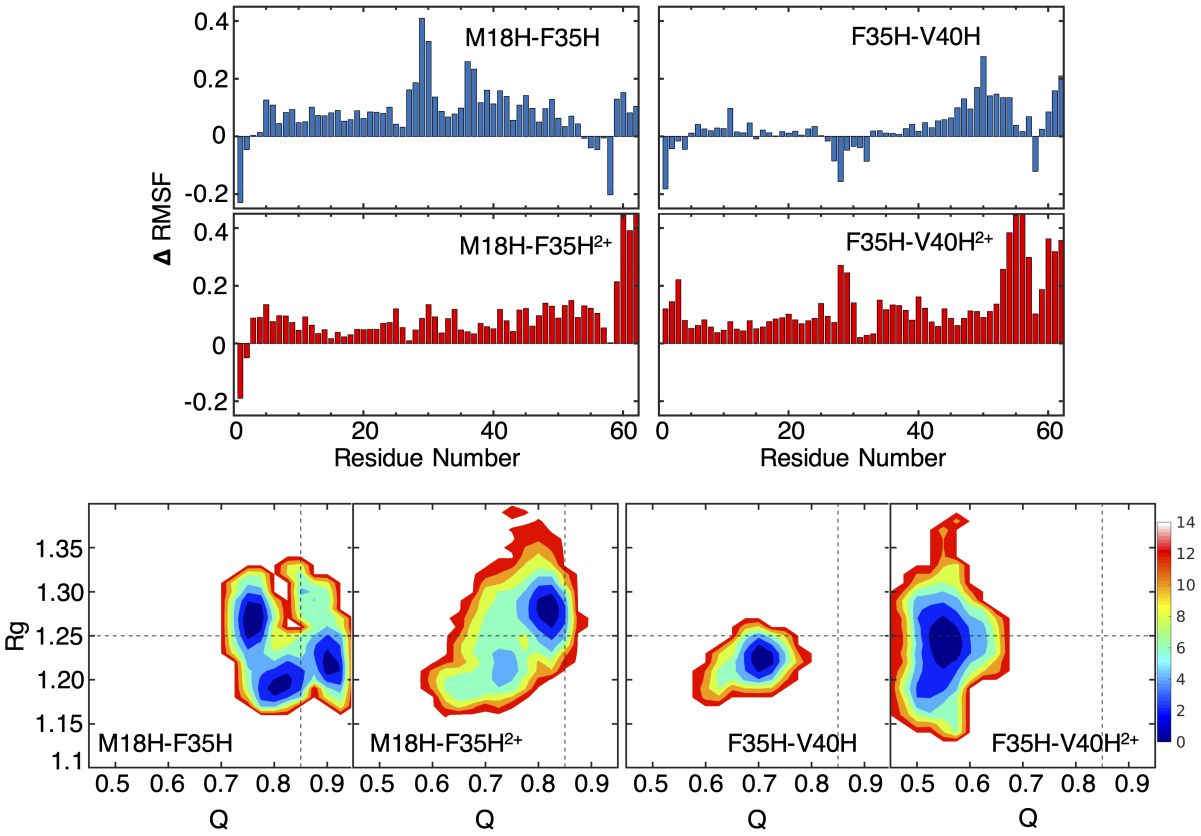

Figure 2.12: Computational analysis of double histidine grafts. (Top) delta RMSF per residue for M18H-F35H and F35HV40H in deprotonated (blue) and protonated (red) form. (Bottom) conformational landscapes as a function of Q and Rg for each double mutant at 310 K. The color bar is in kJ/mol. The crossing lines signal the minimum if the wildtype gpW simulation as reference.

tants (Figure 2.11). The conformational landscapes of the double grafts confirm that both are marginally stable in their deprotonated forms and undergo large unfolding upon double histidine ionization (Figure 2.12 bottom). The degree of disordering is particularly evident for F35H+ - V40H+, amounting to $\approx 50\%$ loss in native contacts.

The experimental analysis of the double grafts highlights a slightly different pH response, which provides further insight into the coupling between downhill (un)folding and proton binding (Figure 2.13). The CD signal at 222 nm has a convex dependence with pH, rather than the sigmoidal-concave dependence of the

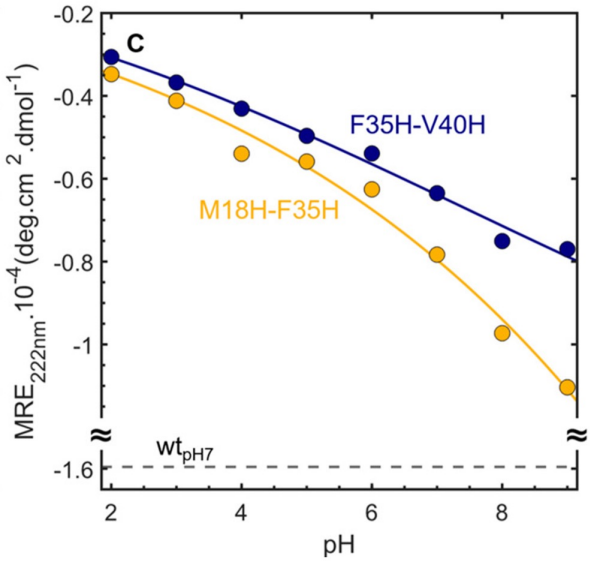

Figure 2.13: CD signal at 222 nm as a function of pH for the double grafts (circles). The corresponding colored lines are to guide the eye. The dashed line (–) represents the CD signal at 222 nm of wildtype gpW at pH 7.

single grafts. The reason behind these differences is that the double mutants are already partially unfolded at room temperature, even when the histidine residues are deionized. Therefore, histidine ionization can only tilt the already partially unfolded ensemble toward more disorder, hence resulting on the convex pH dependence. The response is fairly linear over the full pH range (2–9), which indicates that the double grafts in conjunction with a marginally stable nonionized downhill folding scaffold result in lower sensitivity, but also on ultrabroadband pH transducers.

2.7 Discussion

Lessons for Engineering Protein-Based Conformational Transducers We have explored a strategy for engineering conformational transducers for biosensor applications based on thermodynamically coupling the (un)folding process of a downhill folding protein domain to the binding of an analyte of interest. The underlying hypothesis was that such coupling might give rise to rheostatic (analog) rather than switching (binary) conformational transducers [76]. As protein scaf-

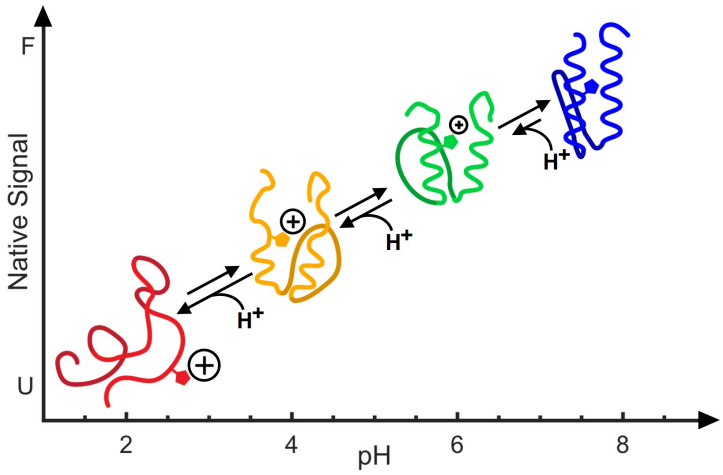

Figure 2.14: GpW rheostatic conformational transducer schematic.

fold we chose the fast, downhill folder gpW, pH as analyte and histidine grafting as approach to engineer the conformational transducer. A first observation is that the histidine grafting strategy we used introduces sensitivity to pH in the mildly acidic to neutral range in a systematic way. pH sensitivity results from the thermodynamic coupling between imidazole ionization and protein unfolding. We find that the strategy is customizable by selecting the degree of histidine burial and the number of grafts. The key for a successful pH transducer is to balance out the repository of folding free energy existing in the scaffold with the perturbation induced by histidine(s) ionization (pK_a shifts) so that the protein undergoes reversible unfolding at the desired pH range. This delicate tradeoff is a key parameter to define the transducer's performance. When the folding free energy repository is larger the transducer's response becomes sharper (e.g. M18H). On the other hand, when the repository is very small the response is broadest, but the amplitude of the signal inevitably drops (e.g. F35H-V40H). For the gpW scaffold the optimal pH transducing response appears to be for grafts with a folding free energy repository of about 5 kJ/mol (e.g. V40H). The right balance can be attained by combining the (des)stabilization of the scaffold's native structure via site-directed mutagenesis and the structure-based design of the histidine grafts.

These lessons from the histidine grafting approach should be extrapolatable to

the engineering of conformational transducers for analytes that require a structurally defined binding site. In such case, the most straightforward approach could be to engineer the folding properties of a protein domain that already contains the binding site for the target ligand. Using protein engineering, one could then tune its sequence to make it both inherently unstable and downhill-like (i.e. by enhancing secondary structure propensity and weakening the hydrophobic core [71],[76]), and balancing the overall perturbation so that it folds gradually upon binding to the ligand (folding upon binding transducer). In that regard, the combination of computational and experimental methods that we use here can prove quite useful. The structural stereochemical rules that we have used for the design of histidine grafts ar straightforward and extensible. MD simulations in the few microseconds timescale seem to recapitulate the conformational effects induced by local perturbations (mutation and ionization), at least for microsecond folding domains such as gpW. We also find good agreement between the results from the simulations and the experimental characterization of the histidine grafts, which opens the possibility of using computational methods to screen larger collections of mutants before producing and characterizing them in the lab. A general approach to engineer rheostatic conformational transducers for other analytes could then involve: 1) identification of a protein scaffold that naturally binds to the analyte; 2) structure-based design of mutations to decrease folding cooperativity and reduce native stability; 3) mutational screening via molecular simulations; 4) production and experimental characterization of most promising candidates.

Downhill (Un)Folding Coupled to Binding: Implications for Broadband Transducing

Our results shed light into the interplay between downhill (un)folding and binding and how it can give rise to conformational transducers with special properties. We can extract several practical lessons. First, the coupling between histidine ionization and downhill (un)folding converts the destabilization directly into structural changes. This feature is inherent to downhill (un)folding domains, which have flexible native ensembles, and gradual (continuous) unfolding

transitions. Therefore, even relatively minor free energy perturbations result in structural changes that can be detected. In other words, the structural changes are not limited to the interconversion between a native and an unfolded state, but also involve the gradual (dis)ordering of these ensembles.

we can conclude that rheostatic conformational transducers add a new, exciting tool to the biosensor engineering toolbox. The sharp response of switching transducers will be preferable for applications where the range in ligand concentration is narrow and minor changes in ligand levels must be detected. An example of ultrasharp response are physiological temperature sensors, which need to detect changes of even a fraction of a degree. On the other hand, rheostatic transducers could provide improved performance to monitor signals that vary widely. For instance, pH changes between 8 and 4 inside living cells statically and dynamically, depending on the cellular compartment and/or the cell's metabolic status. Currently, there are intracellular pH sensors (fluorophore-based and fluorescent protein-based) for either the neutral or the acidic (lysosomal) ranges [97]. However, there are no pH sensors available that can simultaneously operate in all intracellular locations and/or all metabolic stages, even though this capability is widely recognized as essential to understand the role of pH homeostasis in cell biology and physiology [98].

Acknowledgements. This work was supported by the W. M. Keck Foundation. V.M. acknowledges additional support from the European Research Council (grant ERC-2012-ADG- 323059), the National Science foundation (NSF-MCB-1616759), and the CREST Center for Cellular and Biomolecular Machines (grant NSF-CREST-1547848).

CHAPTER 3

Molecular LEGO: An Approach to Map Out the Conformational Landscapes of Unbound Intrinsically Disordered Proteins

> Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.
>
> Richard Feynman

In the previous chapter, we demonstrated that downhill (un)folding coupled to binding is a powerful approach for engineering broadband conformational rheostatic response. Notably, we find that our strategy to engineer a pH transducer is customizable by selecting the degree of histidine burial and the number of mutation grafts providing insights into their morphing behavior. This chapter is devoted to our effort in building a novel approach to investigate morphing transitions of naturally occurring intrinsically partially disordered proteins in the context of the energetic contributions of their native interactions.

3.1 Abstract

Intrinsically disordered proteins (IDPs) fold upon binding to select/recruit various partners, morph around the partner's structure, and exhibit allostery. However, we do not know whether these properties emerge passively from disorder,

or rather are encoded into the IDP's folding mechanisms. A major reason for this gap is the lack of suitable methods to dissect the energetics of IDP conformational landscapes without partners. Here we introduce such an approach that we term molecular LEGO and apply it to the helical, molten-globule-like protein NCBD as proof of concept. The approach entails the experimental and computational characterization of the protein, its separate secondary structure elements (LEGO building blocks), and their super-secondary combinations. Comparative analysis reveals distinct energetic biases in the conformational/folding landscape of NCBD, including: 1) strong local signals that define the three native helices and their limits; 2) stabilization of helix-helix interfaces via mild pairwise tertiary interactions; 3) cooperative stabilization of an overall 3-helix bundle fold; 4) the formation of sets of tertiary interactions that are not present in the NMR ensemble (non-native), but recapitulate the different structures NCBD adopts in complex with various partners. Crucially, the competition between conflicting tertiary interactions makes NCBD gradually shift between structural sub-ensembles. Such conformational behavior provides a built-in mechanism to modulate binding and switch/recruit partners that is likely at the core of NCBD's function as a transcriptional coactivator. Hence, the molecular LEGO approach emerges as a powerful new tool to dissect the conformational landscapes of unbound IDPs and rationalize their functional mechanisms.

3.2 Introduction

The traditional biochemical paradigm states that protein sequences are encoded to fold into unique 3D structures that are thermodynamically stable and define their biologically functional states [99]. However, recent estimates indicate that about 40% of the human proteome is composed of protein domains and/or regions that are intrinsically disordered (IDPs or IDRs) [10]. IDPs are paradigm challengers because they are disordered in their resting state [1],[10], fold, completely or partially, upon binding to their biological effectors [8], can bind structurally diverse partners [17],[79], and exhibit allostery [18],[19],[100] without having qua-

ternary or even a defined tertiary structure. IDPs are more abundant in higher-order organisms, and participate in many fundamental biological processes by playing specialized regulatory roles [101],[6]. From a physical/mechanistic viewpoint, we know that IDPs have distinct sequence patterns [14], including high net charge, low hydrophobicity, and enriched proline content [12],[102]. Some IDPs are devoid of any structure, even after binding to partners [16], but many are partially disordered (IPDP) and morph to accommodate the structural patterns of partners [7]. Hence, research efforts over the last two decades have focused on their folding upon binding. These studies have shown that IPDPs bind partners following either a conformational selection (fold first and then bind) or induced-fit (bind first and fold while bound) process [7],[103]. However, what remains a mystery is the role (if any) that the folding mechanism of the IPDP plays in defining its binding/functional properties. For instance, structural disorder is generally considered necessary and sufficient to enable IPDPs to morph into any required shape on cue. But the question is then: how does an IPDP manage to bind specifically, select among partners, and exhibit allostery? In addition, folding upon binding is often interpreted as a two-state transition (conformational switch). However, such transitions require simultaneous folding and binding [41], which contradicts findings of IPDP binding via induced-fit [104],[105], or the fact that IPDPs can also alternate between conformational selection and induced-fit depending on the partner [106],[107]. Moreover, to fold upon binding as a conformational switch, IPDPs would need sequences that fully encode for the multiple structures that they form in complex with various partners.

As a solution to these puzzles, it has been proposed that IPDPs fold upon binding as conformational rheostats (CR) [29],[108], a functional mechanism associated with the marginally cooperative transitions of downhill folding [101]. Downhill domains have IDP-like sequences and are largely stabilized by local inter-actions, which makes them fold fast but also marginally unstable, and hence partially disordered [29]. The key to CR function is a flexible conformational ensemble that contains built-in energetic biases towards specific (potentially multiple) sub-ensembles. Such biases provide the driving force for selecting

partners and allostery. The gradual transitions of CRs, can also explain how IPDPs morph in response to diverse partners and integrate conformational selection and induced-fit binding [29],[108]. However, investigating the role that the folding mechanism plays in how IPDPs bind and function requires approaches that resolve their conformational landscapes and energetics in the absence of partners. For conventional folded structural domains, this is simply achieved via equilibrium denaturation experiments that are interpreted with a two-state model (unfolded and native) to determine the free energy of folding (ΔG^{UN}) . When performed on many point mutations (designed based on the native structure), this analysis informs on the energetic contributions associated with the structural perturbation caused by each mutation [109],[110]. This analysis interprets the unfolding transition as a binary interconversion between two conformationally invariant states throughout the transition. It also requires that the pre- and post-transition baselines are well defined to infer and extrapolate the properties of the end states [111]. However, none of these requirements holds for IPDP denaturation, which shows extremely broad transitions without baselines, and hence without suitable references to estimate the degree of native structure (or disorder) present at any given condition. It also seems unrealistic to interpret their noticeably uncooperative transitions as a two-state process. In response to this challenge, we introduce here a modular approach that we term molecular LEGO. The approach starts by decomposing an IPDP into its basic secondary structural elements, or LEGO building blocks, and their combinations. The combined elements recapitulate subsets of tertiary interactions, in analogy to the complementary indentations between bricks in the LEGO toy. The conformational analysis of building blocks probes the contributions from local interactions and provides reference ensembles for interpreting the results on the higher-order elements. Such reference ensembles are essential to detect any subtle biases that might occur in the protein and to convert them into energetic contributions using simple statistical thermodynamic analysis. The approach is inspired by work in the early 90s that searched for local folding nuclei in series of peptides spanning the protein's sequence [112]. Those studies, which were performed on two-state

folding domains, revealed weak native-like biases in the fragments [112] and the need for almost the entire protein to elicit any detectable folding [113]. However, we reason that the high contributions from local interactions and minimal folding cooperativity expected for IPDPs [29] make them more suited for these types of studies. A similar modular approach has been, in fact, recently used to investigate the folding landscape of IDPs via molecular simulations, in which the much faster dynamics of the small protein fragments greatly enhance conformational sampling [114]. Another key advantage of a modular approach is that it facilitates the direct quantitative comparison between experiments and simulations.

To demonstrate the molecular LEGO approach, we focused on the protein NCBD. NCBD is categorized as IPDP, and there is a wealth of biophysical data available on its folding and binding, including NMR [17],[62],[23], molecular simulations [63],[115],[30] and SM-FRET [18],[65],[116]. NCBD binds to multiple, structurally diverse partners, including IDPs (e.g., p53-TAD [23] and ACTR [17] and globular proteins such as IRF [22], by adapting its ensemble to the partner. In its free form, NCBD exhibits high α-helical content without defined tertiary structure, but it adopts a three-helix bundle fold driven by a few mid-range contacts [62]. However, the (dis)ordering transitions of NCBD are broad and featureless, including its thermal unfolding and stabilization via the cosolvent trifluoroethanol. All these properties make NCBD ideal for the molecular LEGO proof of concept. We designed the LEGO elements based on the existing NMR structural ensemble (Figure 3.1) and analyzed their behavior experimentally[1] and computationally. Particularly, we studied all the LEGO elements and NCBD using circular dichroism and the structure-promoting cosolvent trifluoroethanol as thermodynamic variable. We also performed all-atom Molecular Dynamics (MD) simulations in explicit solvent, taking advantage of the shorter timescales (μs) involved in the conformational changes of IDPs and the recent availability of IDP-improved force fields [45], [117]. Experiments and simulations were interpreted and compared using an elementary statistical thermodynamic treatment of the helix/coil transition. Our results on NCBD demonstrate that the LEGO

[1]Experiments performed by Thinh D.N. Luong in Muñoz Group

approach is a powerful tool to map out the folding landscapes of unbound IPDPs and rationalize their folding upon binding, and hence their function.

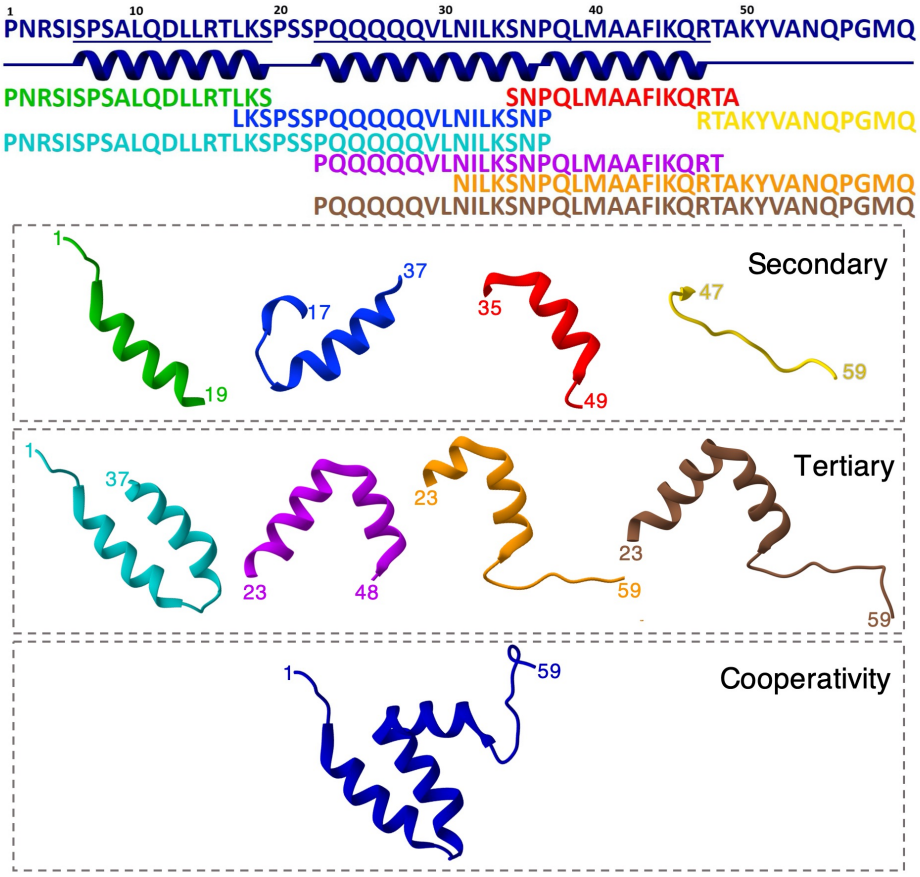

Figure 3.1: Molecular LEGO design. (From top to bottom) The complete NCBD sequence (ID: 2KKJ) and a diagram showing the 3 α-helices found in the NMR ensemble are shown in navy blue. The sequences of the 8 fragments, designed according to the sequence and structural patterns of NCBD, are shown color-coded: building blocks in primary colors (H1 green, H2 blue, H3 red, T yellow), and the combined elements in corresponding secondary colors (H1-H2 cyan, H2-H3 magenta, H3-T orange, and H2-H3-T brown). Diagram showing the structure of each fragment and full protein in the NCBD NMR structure in cartoon representation. The color coding is maintained. The building blocks report on secondary structure propensities. Building block combinations report on pairwise (element to element) tertiary interactions: e.g., H1-H2 reports on the tertiary interactions between helices 1 and 2. Comparison of the fragments with the behavior of the full protein reports on the degree of folding cooperativity.

3.3 Methods

All-atom MD simulations. We carried MD simulations in explicit solvent using the GROMACS package [118],[119], and the Charmm22* force field [52]. Water molecules were described using the TIP3P model. Periodic boundary conditions were used, and long-range electrostatic interactions were treated with the Particle Mesh Ewald (PME) [120] summation using a grid spacing of 0.16 nm combined with a fourth-order cubic interpolation to derive the potential and forces in-between grid points. The real space cutoff distance was set to 1.2 nm, and the van der Waals cutoff to 1.2 nm. The bond lengths were fixed[121] , and a time step of 2 fs was used for the numerical integration of the equations of motion. Coordinates were recorded every 10 ps. For NCBD, we performed two separate 12 μs trajectories starting from the lowest energy structure of the NCBD NMR ensemble (PDB ID: 2KKJ). The protein was placed in a dodecahedral water box (volume = 262.38 nm^3) large enough to contain the protein and at least a 1.0 nm layer of solvent on all sides. The structure was solvated with 8,216 water molecules, and six Cl- ions were added to neutralize the system. The starting coordinates for the 8 NCBD fragments (as defined in Figure 3.1) were extricated from the protein's PDB file. The fragments were acetylated and/or amidated as needed to replicate the chemically synthesized peptides (H1, H12 free and amidated; H2, H3, H23 acetylated and amidated; T, H3T, H23T acetylated and free). The CHARMM22* force field was then adjusted to include the parameters for N-acetylation and C-amidation. Box dimensions were kept sufficiently large to account for the high flexibility and large-scale motions expected on these peptides. Two 2 μs trajectories were performed for each fragment (three 2 μs trajectories for the larger fragments H12 and H23T. In all cases, the starting structure was subjected to energy minimization using the steepest descent method. All systems were equilibrated at a constant temperature of 310 K utilizing the two-step ensemble procedure (NVT and NPT). First, the system was subjected to NVT (constant number of particles, volume, and temperature) equilibration for 100 ps with the position of the protein restrained, followed by NPT

(constant number of particles, pressure, and temperature) equilibration for 2 ns each. The simulations were subjected to the modified Berendsen thermostat with a 0.1 ps relaxation time [122] to maintain the temperature. The structures were then subjected to Parrinello-Rahman with 0.2 ps relaxation time for pressure coupling [123] at 1 bar before the production run was started. All the simulations were run on the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputing center (SDSC).

Analysis of MD simulations. The number of native contacts per residue was calculated from each MD trajectory with a 1 ns time step and using the NMR structure as the reference of native contacts. Contacts were defined using a 0.5 nm cutoff between any two pairs of heavy atoms that are at least 3 residues apart in the sequence. The number of native contacts trajectory was then converted into the fraction of native contacts (Q). We used the peptide bonds as basic conformational unit to compare with experimental data analyzed with the Zimm-Bragg model. Each trajectory was then analyzed to assign each peptide bond of the simulated molecule to either helix or coil state at each time frame. The helical state (H) was defined according to the local conformation (dihedral angles) and backbone hydrogen bonding status. These processed trajectories were finally used to calculate the number of helical residues per time frame, and the average fraction helix per residue for each molecule.

Analysis of dihedral angles. We classified the conformation of a peptide bond unit based on its flanking ψ and ϕ angles. Particularly, we defined a helical peptide bond (h) when its dihedral angles are -50° $< \psi <$ -17° and -80° $< \phi <$ -50°, and coil peptide bond (c) as everything else.

Analysis of hydrogen bonds. A hydrogen bond between residues i and i+4 was considered formed when the donor-acceptor distance was $<$ 0.35 nm and the donor-hydrogen-acceptor angle >160°. We computed every hydrogen bond formed at each time frame using the MD Analysis python toolkit: we first evaluated all possible hydrogen bonds per time frame, and then every time a i,i+4 hydrogen bond was formed according to our criteria, we assigned a hydrogen-

bonded state (h_{HB}) to peptide bonds i+1, i+2, and i+3.

Helix-Coil treatment. We describe the formation of helical structure using the Zimm-Bragg helix-coil theory. In the Zimm-Bragg model, each peptide bond can be in either helical conformation (h) or coil (c), and helix formation occurs by process of nucleation (cost of forming the first helical hydrogen bond, defined by the parameter σ) and elongation (defined by the parameter s). With this definition and using the coil as reference state, the statistical weight matrix is defined as:

$$M = \begin{pmatrix} 1 & \sigma s \\ 1 & s \end{pmatrix} \tag{3.1}$$

for which the partition function is

q=(1,0)M^n $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ where n is the number of peptide bonds in the molecule.

Computing helix nucleation and elongation. We define the elongation parameter (s) as the equilibrium constant between the helix and coil states of the central peptide bond in a triplet. For a given time frame, the helix state of the central peptide bond is any of the following: [c h h], [h h h], [h h c] or [c h c]; and the coil state is either [h c c], [h c h], [c c h] or [c c c]. For a predefined helical segment, s is simply the average of the elongation for all the peptide bonds within it. The nucleation parameter (σ) is defined as the equilibrium constant for the formation of the first helical (i, i+4) hydrogen bond (flanked by coil peptide bonds). To calculate σ, we used a rolling window of 7 peptide bonds and defined nucleation on the third peptide bond ($\tilde{h}$) as:

$$\sigma = \frac{1}{t}\sum(cc\tilde{h}hhcc) \tag{3.2}$$

where t is the number of time frames in the trajectory. The final parameters for one molecule were determined as the average over all the available MD trajectories.

Time-averaged contact map. To calculate the time-averaged contact maps we considered that a contact is formed at any given 10 ns interval when at least one atom of residue i is within a cutoff distance of 0.5 nm of at least one atom

of residue j (where j >= i+3) with a probability > 0.7 during such time interval. Native contacts were defined on the basis of the atomic coordinates of the NMR structure.

Estimating pairwise tertiary interactions and cooperativity. In the Zimm-Bragg model, the statistical weight (w) of a given helical conformation is given by w=σs^i, where i is the number of helical peptide bonds. We can calculate the statistical weight expected for a fully folded molecule containing two helical elements (molecular LEGO's building blocks) that are not interacting with one another, as the product of the statistical weights of the fully formed helical elements. Hence, the contributions from tertiary interactions between the two elements can be obtained from the ratio between the statistical weights of the entire molecule divided by the product of the weights of its separated elements as:

$$\Delta G_{mn} = -RT \ln \left(w_{mn} / \left(w_m w_n \right) \right) \tag{3.3}$$

where $w_m = \sigma_m s_m^{\mathrm{k}_m}$ and $w_n = \sigma_n s_n^{\mathrm{k}_n}$ are the statistical weights of the fully-induced helical conformation of building blocks m and n, and $w_{mn} = \sigma_{mn}^2 s_{mn}^{\mathrm{k}_{mn}}$ is the statistical weight of a molecule containing building blocks m and n in a full helical conformation. In these expressions, k is the number of residues that need to become helical to form the helix(es) defined from the H1, H2, H3, and T lengths from the full protein MD ensemble. For instance, k for H1H2 is H1+H2 lengths determined from the two NCBD protein trajectories. We used this procedure to calculate pairwise interactions between helices 1 and 2 and helices 2 and 3. For the tail (T), we considered that its effect on a combined molecule is to extend helix 3 rather than nucleating a new one (w_{3T} only includes 1 nucleation and w_{23T} includes 2). After the pairwise tertiary interactions have been estimated, the same calculation can be carried out for the entire protein to estimate the overall folding cooperativity. In this case, the fully formed conformation includes three helices, and hence $w_{NCBD} = \sigma_{NCBD}^3 S_{NCBD}^{\mathrm{k}_{NCBD}}$, relative to the product of the statistical weights of the four elements. The overall folding cooperativity

is finally obtained as: $\Delta G_{\text{coop}} = \Delta G_{NCBD} - (\Delta G_{H12} + \Delta G_{H23T})$.
We performed these calculations for the experimental data using the helix-coil parameters, and for the MD simulations using nucleation and elongation parameters obtained from the analysis of the trajectories.

3.4 Results

3.4.1 Molecular LEGO design.

The design of the LEGO elements (locations and extension along the sequence) of highly disordered proteins is far from trivial, unless there are structures in complex with partners available. IPDPs, however, do have residual structure, which for NCBD was sufficient to enable the determination on an NMR ensemble based on chemical shifts and a few mid-range NOEs [62]. We used this NMR ensemble to divide the 59-residue sequence of NCBD into four building blocks that represent its local (secondary) structural segments: helices 1, 2, and 3 (H1, H2, H3) and the C-terminal tail (T). We further refined the limits of the α-helices based on the predictions of helical propensity from AGADIR [124], which delineates a clear helix profile (Figure 3.2).We then designed four combinations of consecutive building blocks (H1H2, H2H3, H3T, H2H3T) to recapitulate the various sets of native pairwise tertiary interactions. Finally, the comparison of LEGO elements with the entire protein informs on the overall contribution from global cooperativity to the NCBD ensemble. The complete molecular LEGO design is shown in Figure 3.1.

3.4.2 Strategy to Dissect Conformational Ensembles

.

We analyzed NCBD and its LEGO elements by experiment and simulation. Experimentally, we employed far-UV circular dichroism spectroscopy, which reports on the average peptide bond conformation of the protein/peptide and is particularly sensitive to α-helical structures (NCBD and most IPDPs are, or be-

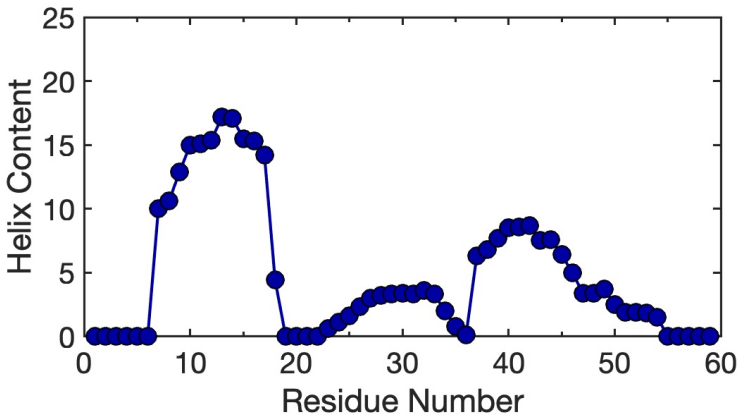

Figure 3.2: Predicted NCBD helical content from AGADIR shows 3 α-helices at the precise locations determined by NMR experiments.

come upon binding, α-helical). We use the cosolvent 2,2,2-trifluoroethanol (TFE) as thermodynamic variable to enhance the inherent conformational propensities of the LEGO elements and full/complete/entire protein. TFE is a polar/organic cosolvent that induces structure in peptides and proteins by strengthening the backbone intramolecular hydrogen bonds relative to the hydrogen bonds they make with water [125]. The TFE titration of the building block H1 monitored by far-UV CD is given in Figure 3.3 (left) as an example. In the absence of TFE, the CD spectrum of H1 indicates a population of $\approx 25\%$ α-helix with the remainder being random coil. The addition of TFE steadily increases the α-helical content of H1 until it plateaus (from 0.3 to 0.5 ϕTFE).

These results indicate that the interplay between TFE and the folding/structural propensities of the molecules in this study can be analyzed in terms of the statistical thermodynamics of the helix-coil transition [126],[127]. The helix-coil transition describes the formation of α-helices at the residue level as a nucleation (σ) and elongation (s) process [128]. The effect of TFE can be simply described as an increase in the elongation parameter (stronger hydrogen bonds), which promotes a cooperative (sigmoidal) transition to α-helix structure (Figure 3.3 right).

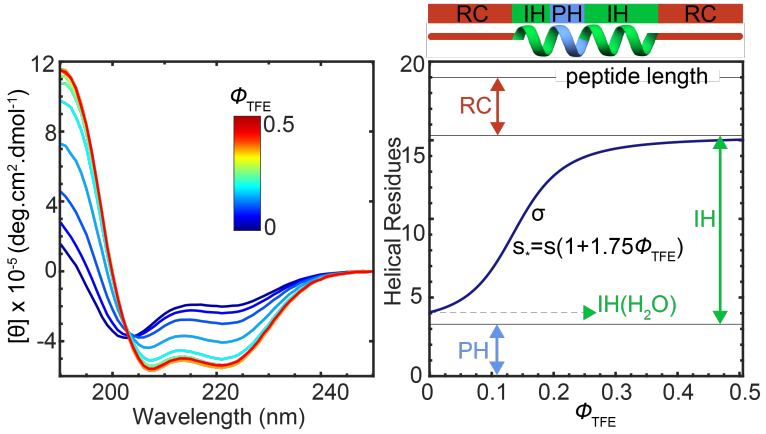

Figure 3.3: Experimental conformational analysis of NCBD and LEGO elements. To probe the energetic biases in the conformational ensemble of each molecule, we use TFE as a structure promoting agent and monitor the changes in conformation by far-UV CD. The left panel shows the CD spectra of H1 as a function of TFE concentration as an example. The right panel summarizes the tripartite helix-coil analysis of the TFE titration for each molecule: preformed helical residues (PH) in blue, TFE-inducible helical residues (IH) in green, and TFE-insensitive random coil residues (RC) in red. The average number of helical residues (dark blue) is obtained from the CD spectra.

For this purpose, we implemented a tripartite helix-coil model based on the original Zimm-Bragg homopolymer treatment [127]. The tripartite model divides any polypeptide chain into three different types of units (peptide bonds): PH, which are already α-helical without TFE; RC, which are random coil regardless of TFE; and IH, which have residual α-helix population that can be significantly enhanced by TFE. The model can be used to calculate the average number of helical peptide bonds on any peptide/protein with only four parameters: the number of PH units, and σ, s, and the number of IH units. We then analyzed the CD spectra as a function of TFE for each peptide/protein to determine the number of helical peptide bonds using singular value decomposition (SVD) analysis. Each dataset was expressed in molar ellipticity units; and a value of -39,500 deg.cm2.dmol-1 for the molar ellipticity at 222 nm of one helical peptide bond [129],[130] was used to convert the data into the number of helical peptide bonds as a function of TFE.

Computationally, we performed atomistic MD simulations in explicit solvent. For full NCBD, we performed two independent 12 μs simulations (total simulation time of 24 μs). For the LEGO elements, we performed 2-3 sets of 2 μs simulations, taking advantage of the expectation of much faster conformational dynamics. We used the CHARMM22* force field with the TIP3P water model, which have been shown before to produce a good agreement with experiments on partially disordered proteins [51],[131].

We first examined the MD conformational ensembles as a function of the fraction of native contacts (Q): all trajectories given in Figure 3.4. The simulations of the LEGO building blocks showed abrupt fluctuations in Q (their number of native contacts is small) taking place in ns timescales. The combined LEGO elements exhibited fluctuations in Q of smaller relative magnitude that are also somewhat slower, but several transitions are still observable in each 2 μs trajectory (Figure 3.4). The behavior of NCBD is similar, but the protein trajectories show an additional slowdown in the dynamics: six times longer trajectories produce similar numbers of transitions (Figure 3.4). The number of transitions

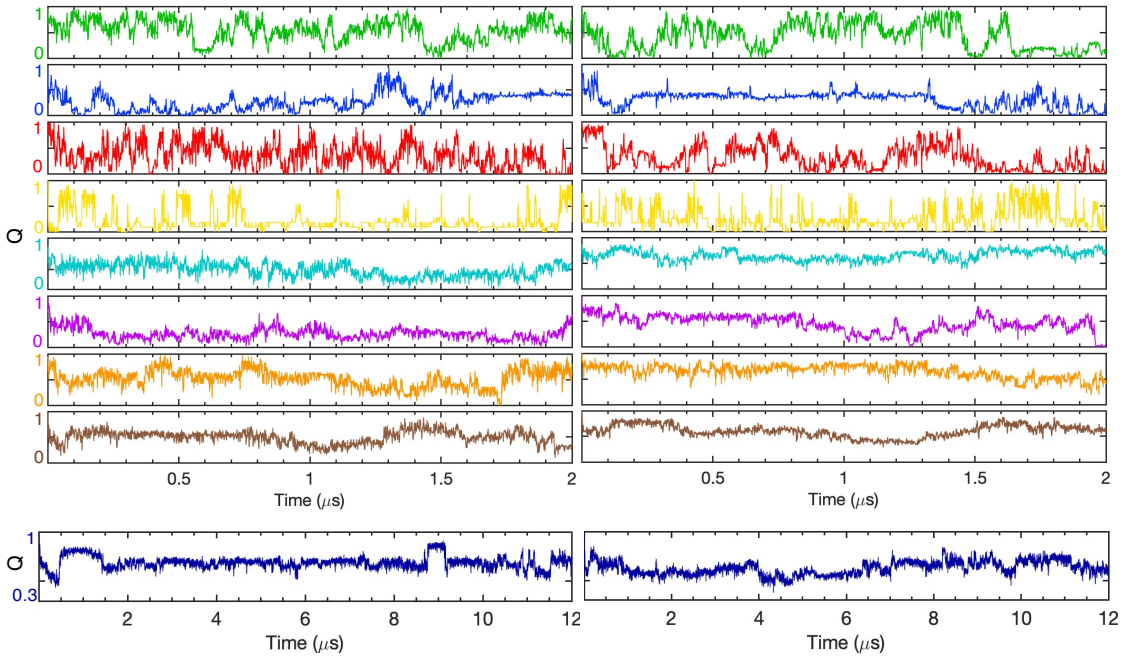

Figure 3.4: Time evolution of fraction of native contacts (Q) sampled in representative MD trajectories of all 8 fragments and the protein.

per trajectory and the consistency between the average behavior of separate trajectories suggests that the conformational sampling within these timescales is reasonable. The NCBD trajectories as a function of Q are also in good agreement with previous simulations in which the C-terminal tail was removed [30]. We then analyzed the trajectories to compute the fraction helix, as well as nucleation and elongation ZB parameters, for each peptide bond in each molecule (see methods). The agreement between the residue-specific helix populations obtained from individual simulations further supports that the simulated timescales afford reasonable sampling.

3.4.3 Conformational Propensities of LEGO Building Blocks.

The results of the conformational analysis for the four building blocks (H1, H2, H3 and T) are provided in Figure 3.5-Figure 3.6. In general, these results indicate that the three regions containing α-helices in the NCBD NMR structure have residual helical structure on their own and are very sensitive to TFE. Of all the elements, H1 has the highest residual helical structure, both in experiments

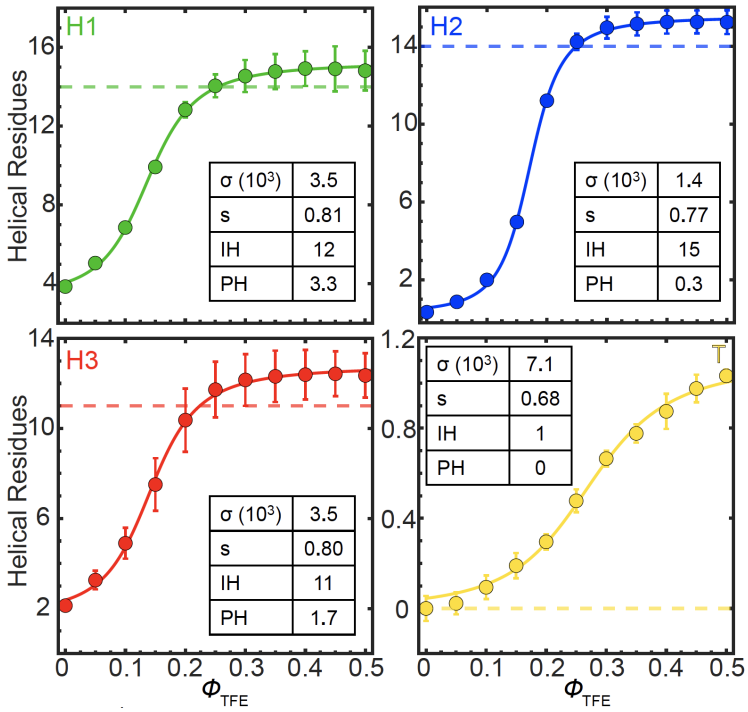

Figure 3.5: LEGO building blocks: secondary structure propensities and local interactions. The experimental conformational analysis of the 4 NCBD building blocks. Color coding as in Figure 3.1. The panels show the average number of helical residues (circles) and experimental error, obtained from two independent measurements, as a function of the TFE volume fraction for H1, H2, H3, and T. The colored curves represent the fit to tripartite helix-coil model, and the parameters from the fit are given in the inset. Dash lines indicate the number of helical residues determined from the NMR structure.

and simulations. The maximal helix length of H1, H2, and H3 (i.e., at the highest TFE) is only slightly longer (about one residue) than the helices in the NMR ensemble, which suggests that local signals tightly control the location and extent of the NCBD helices. It is also apparent that the tail (T) does not have a detectable helical structure, but a minimal helix of about 1 residue (i.e., 1 hydrogen-bonded peptide bond, or 1 helix turn) forms at the highest TFE.

The helix-coil parameters for each building block are given as an inset in each panel of Figure 3.5. This experimental analysis shows that the cost of nucleation (σ) for H1, H2, and H3 is within the range of the values found in idealized model

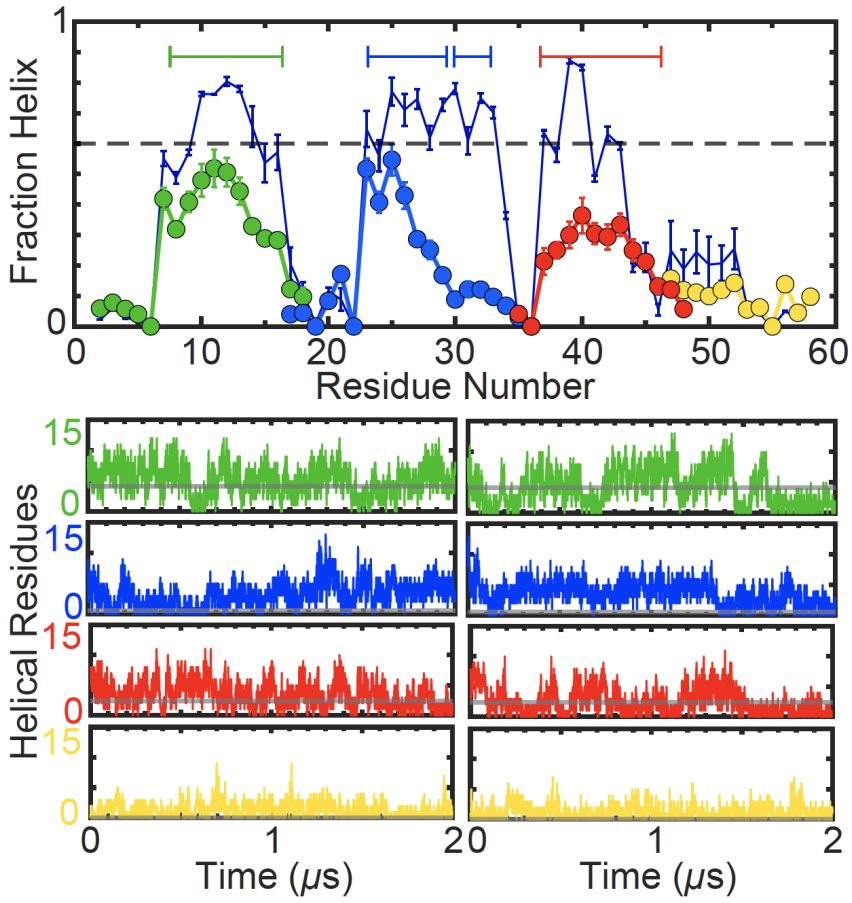

Figure 3.6: The helical propensity per residue for the 4 building blocks determined from two 2 μs MD simulations. The helical propensity profile for the full-length protein (discussed later) is shown with a thin navy line for reference. The horizontal lines signal the helix length (consecutive residues with at least 10% fraction helix) emerging from these simulations. The grey dashed line indicates a 60% helicity threshold. Error bars indicate the standard error of two trajectories. The bottom panels show the time evolution of the number of helical residues for each molecule in two separate 2 μs MD trajectories. The horizontal grey lines indicate the average number of helical residues determined from experiments at ϕTFE = 0 (ordinate intercept in Figure 3.5), shown for comparison.

	H1	H2	H3	T	H1H2	H2H3	H3T	H2H3T	NCBD
s	0.63	0.39	0.39	0.14	0.87	0.5	0.8	0.8	1.2
σ	0.018	0.019	0.021	0.004	0.025	0.023	0.018	0.018	0.024

Table 3.1: Helix–coil model parameters calculated from MD simulations of all components and NCBD.

peptides used to investigate helix stability [132]. H1 and H3 are slightly easier to nucleate and hence less cooperative than H2. Elongation is somewhat lower than 1 for all the sequences, which explains both their residual helix content (on an infinitely long helix s=1 results in 50% helix content) but also their high sensitivity to TFE (easy to raise s above 1). T is interesting because even though it has minuscule helical propensity overall, it contains a one turn region that seems primed to become helical by stabilizing factors.

The MD simulations are in good agreement with the experimental findings, including the presence of residual helical structure, the average helix population per molecule (particularly H1 and H3), and the detection of some marginal helical propensity in T. The simulations also indicate that the helical population is not uniform throughout each building block. The asymmetric helical distribution along each building block, most notably in H2, further supports the use of the tripartite helix-coil model to analyze experiments. In addition, the extension of the helical regions in the simulations is in excellent agreement with those of the NCBD NMR ensemble. This result further confirms that the helical regions in the NCBD ensemble are defined by strong local signals. On the other hand, the and s parameters obtained from the simulations (Table 3.1) differ from the experimental values in that they produce systematically lower nucleation costs (about 5-10-fold larger σ) and elongation (smaller s). The differences in both parameters compensate each other to produce similar helical contents (Figure 3.6). The implication is that the force field underestimates the cooperativity of the helix-coil transition, and generally of folding, in agreement with previous benchmark studies [133],[134].

The combined experiments and simulations on the LEGO building blocks demonstrate that the sequence of NCBD contains very specific local signals. Such signals prime certain regions on NCBD to form α-helices upon mild stabilization by other factors (i.e., TFE, tertiary interactions, partner binding) and also seem to define their limits. The consistency between the local conformational biases in the isolated building blocks and the structural ensemble of the full protein suggests that local interactions play a major role in determining the folding landscape of NCBD.

3.4.4 Conformational Biases Through Pairwise Tertiary Interactions.

The results for the combined LEGO elements are given in Figure 3.7-Figure 3.8. The behaviors of these molecules should highlight any contributions from pairwise tertiary interactions to the NCBD conformational ensemble. Qualitatively, the experimental and computational results are similar to those of the building blocks: i) residual helical structure in native conditions, ii) strong response to TFE, iii) sigmoidal TFE transitions, and iv) helix populations within the helix lengths of the NCBD NMR ensemble. However, the comparison between the combined LEGO elements and the compounded effects of their separate building blocks reveal significant differences that demonstrate the presence of transient interactions between elements.

Particularly, all of the combined elements exhibit enhanced sensitivity to TFE, as manifested by the curves with higher slopes and plateauing at lower ϕTFE, as well as their slightly higher σ and s helix-coil parameters. Notably, the experiments do not detect major increases in residual helical structure in the absence of TFE. Hence, the energetic biases introduced by pairwise tertiary interactions are insufficient on their own to increase the helical content by experimentally detectable levels. However, increases in helical content are observed in the simulations, possibly owing to their much higher sensitivity and resolution at the residue level. Another observation is that the thermodynamic coupling between consecutive LEGO building blocks seems to have a significant impact on redefin-

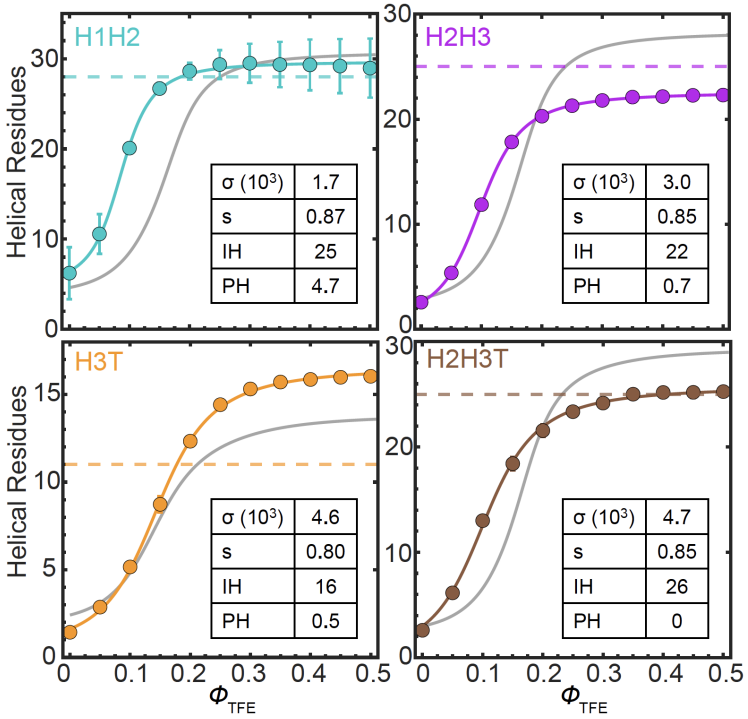

Figure 3.7: Combinations of building blocks: mapping pairwise tertiary interactions. The experimental conformational analysis of the 4 combinations of building blocks. Color coding as in Figure 3.1. The panels show the average number of helical residues (circles) and experimental error, obtained from two independent measurements, as a function of the TFE volume fraction for H12, H23, H3T, and H23T. The grey curves show the compounded curves of the relevant building blocks for each combination (e.g., H1 and H2 for H12) and represent the reference behavior expected for the combined fragment if the effect is additive (no tertiary interactions). Dashed lines as in Figure 3.5.

ing the maximal helix lengths, most notably for H3.

On an individual basis, we find that the interactions between H1 and H2 are stronger than between 2 and 3. H1H2 does, in fact, exhibit some increase in residual helical content in experiments and simulations. In fact, the fraction helix calculated from simulations is in very good agreement with the experiments (cyan in Figure 3.7). The effects on H2H3 are somewhat more subdued in simulations and only detectable from the response to TFE in experiments. The impact of the tail on H3 is interesting, as the added C-terminal sequence seems to stimulate the extension of the helix beyond what is observed in the NMR ensemble. The helix extension is clear in the experiments (3 residues longer maximal helix length) and the simulations (see the H3T profile in orange, Figure 3.8). In other words, whereas the tail does not nucleate much of a helix on its own, it effectively elongates a helix formed in its preceding sequence. The simulations indicate that this effect is entirely driven by local interactions (helix-coil cooperativity). The extension of H3 is also predicted by AGADIR (Figure 3.2), which further supports an entirely local origin for this effect. The effects of pairwise interactions on the three helix lengths are more individualized. For instance, simulations of H1H2 show that pairwise tertiary interactions between helices 1 and 2 increase the intrinsic helical population (mostly at the end of helix 2), but these interactions do not seem to change the maximal length of either helix in experiments or simulations. In contrast, experiments on H2H3 indicate a maximal helix of ≈ 23 residues, whereas, in the NCBD NMR ensemble, this region extends over 25, and the H2 and H3 building blocks sum up to 28. At least part of this difference seems to arise from local capping effects of the region connecting helices 1 and 2, which is absent in H2H3 and H2H3T (Figure 3.1). This is apparent in the simulations, which show residual helical population in that region, as well as the stabilization of the beginning of the helix 2 in H2 relative to H2H3 (Figure 3.6 vs. Figure 3.8). Furthermore, the presence of helix 2 seems to impede the elongation of helix 3 into the tail. This is readily apparent in experiments, which show that H2H3T has a maximum helix of 26 in perfect agreement with the NCBD NMR ensemble. In contrast, the maximal helix lengths of H2 and H3T add up to 30. The same

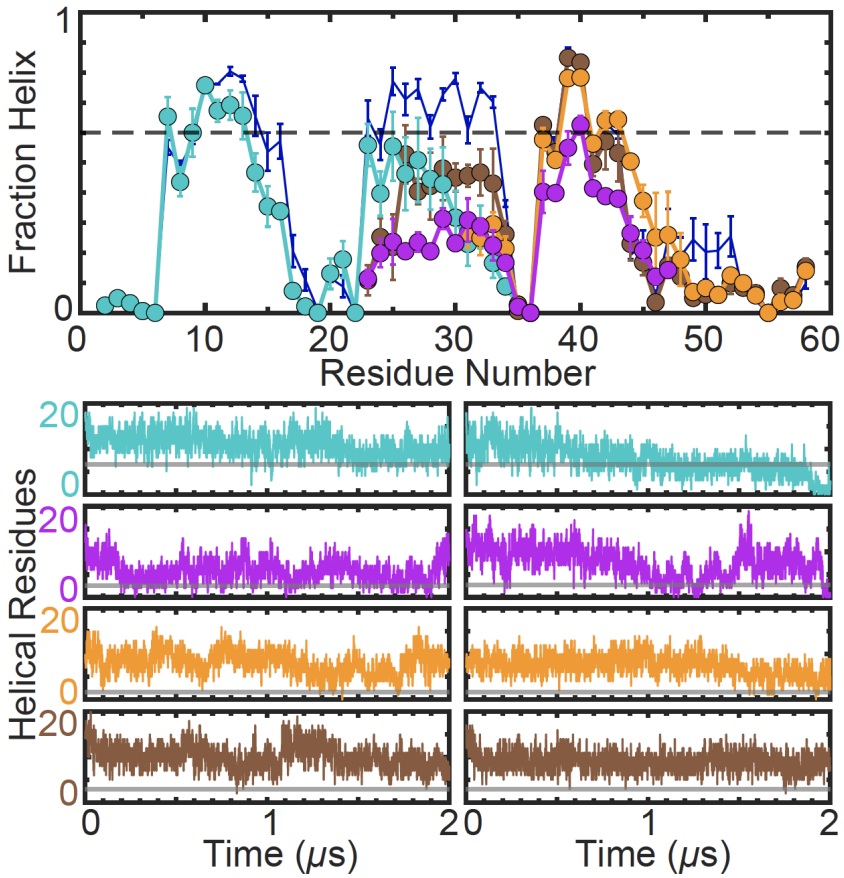

Figure 3.8: The helical propensity per residue for the 4 combined LEGO elements obtained from two 2 μs MD trajectories. The propensity of the full-length protein is shown as a thin navy-blue line for reference. The grey dashed line indicates a 60% helicity threshold. Error bars indicate the standard error of two trajectories for H23 and H3T and 3 trajectories for H12 and H23T. The bottom panels show the time evolution of the number of helical residues for each molecule in two separate 2 μs MD trajectories. The horizontal grey lines indicate the average number of helical residues for each fragment determined from experiments at ϕTFE = 0 (ordinate intercept in Figure 3.7), shown for comparison.

pattern is observed in simulations, with helix 3 constrained within the limits of the NCBD NMR ensemble in the H2H3T molecule. Strikingly, there also seem to be non-native interactions (not found in the NCBD NMR ensemble) between helix 2 and the tail. This is evident in simulations, which show that the tail stabilizes helices 2 and 3 without becoming itself helical (brown versus orange in Figure 3.8). The experiments are also consistent, showing an increase in elongation (s) for H2H3T relative to H3T, jointly with a reduced maximal helix length. The main discrepancy between experiments and simulations is quantitative: the interplay between helices 2 and 3 with the tail results in a strong stabilization of the two helices in the simulations. The effect is, however, more subtle in experiments. Hence, the simulations overestimate the helical population of the relevant molecules relative to experiments, most particularly H3T and H2H3T, and to a lesser extent, H2H3.

3.4.5 Global Stabilization Effects in the NCBD Ensemble.

The LEGO results provide useful references to interpret the uncooperative (non-sigmoidal) TFE transition of full NCBD (Figure 3.9), which is, in fact, much broader than that of its elements. By compounding different LEGO elements, we then establish the behavior that would be expected from only local interactions (grey profile), or after adding the interactions between helices 1-2 (green profile), or those between helices 2-3 and tail (pink profile).

The comparison reveals that NCBD has much higher helical content than expected from the sum of its parts: $\approx$ 24 helical residues in water relative to 6-7 residues for the three combinations of LEGO elements. The helix-coil analysis indicates that about 15 residues are fully helical (PH) in water, whereas the remaining helical content comes from the partial helical population (i.e., $\approx$ 30%) of many IH residues. Hence, in full NCBD the helix-inducible residues (IH) already have high helical content in water, which enormously facilitates nucleation: 10-fold higher σ relative to the LEGO elements. Elongation (s) is, on the other hand, comparable. In other words, the low TFE sensitivity of NCBD is

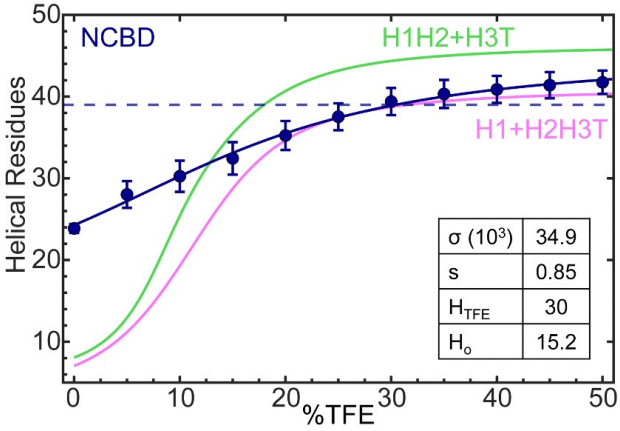

Figure 3.9: Cooperativity in the NCBD conformational landscape. Average number of helical residues (circles) and experimental error, obtained from two independent measurements, as a function of the TFE volume fraction for full-length NCBD. The grey curve shows the compounded curves of the 4 building blocks (H1, H2, H3, T). The pink and light green curves show the compounded curves of H12 with H3T and of H1 with H23T, respectively.

not because its conformational ensemble is disordered, but because it is already highly primed towards forming α-helical structure due to interactions that are only present in the entire protein. The effect of TFE on folded globular proteins is complex, switching from native-stabilizing at relatively low volume fractions to denaturing as TFE becomes the main solvent. What we see in NCBD is that the native-stabilizing effect extends to higher TFE volume fractions. Indeed, at 0.5 ϕ TFE NCBD reaches $\approx$ 41 helical residues, in agreement with the NMR ensemble (dashed line in (Figure 3.9). However, the helix-coil parameters indicate that, in contrast with its LEGO elements, NCBD continues to increase its helical content beyond 0.5 ϕ ($\approx$ 4 more residues), thereby starting to promote non-native conformations. The broader native-stabilizing range of TFE could due to the fact that NCBD is natively α-helical and lacks a defined hydrophobic core [125]. Hence this property could be common to other IPDPs. For NCBD, the simulations closely reproduce the main experimental results: overall helical content in water (Figure 3.10), nucleation and elongation (Table 3.1).

The simulations also show that helix 2, which has the lowest intrinsic propen-

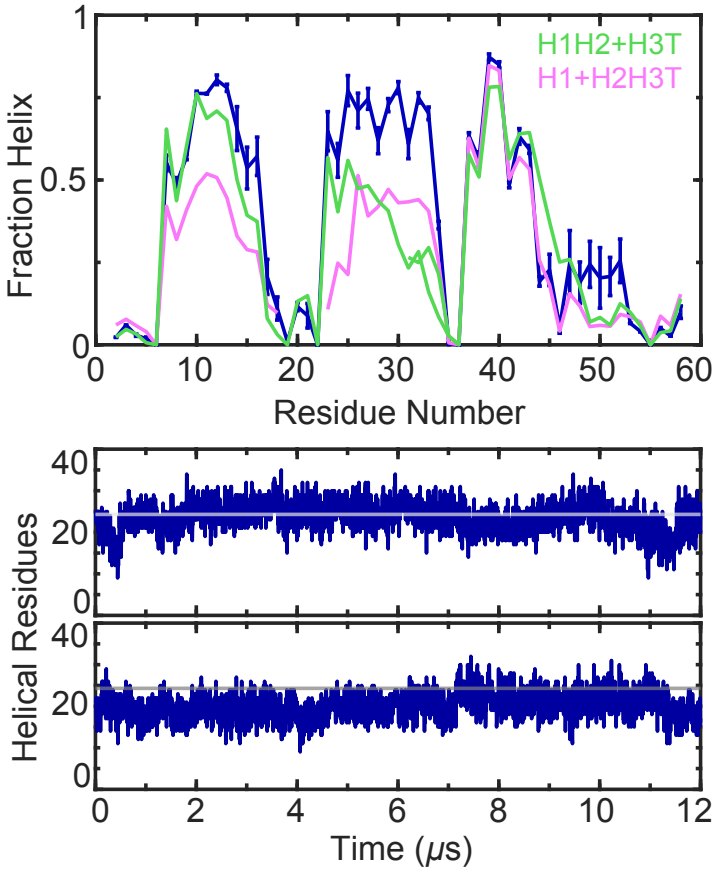

Figure 3.10: Helix fraction per residue for the full-length NCBD (navy-blue) obtained as the average of two 12 μs MD simulations, compared to the compounded helical propensity patterns of H12+H3T (light green) and H1+H23T (pink). Bottom) Time evolution of the number of helical residues in NCBD for two separate 12 μs MD trajectories. The horizontal grey lines indicate the average number of helical residues for each fragment determined from experiments at ϕTFE = 0 (Figure 3.9).

sity (Figure 3.6), is preferentially stabilized in the full protein (Figure 3.10) and engages in frequent interactions with the other two helices. The stabilization of helix 2 in presence of both flanking helices is evident in the comparison of the NCBD helix profile with the H1H2+H3T (green) and H1+H2H3T (pink) compounded profiles. This comparison also highlights that helix 1 is mostly stabilized by 1-2 interactions, and helix 3 is stabilized/delimited by its interplay with helix 2 and the tail. The NCBD simulations also show the transient formation of many long-range interactions that are not seen in the NMR ensemble (non-native); particularly between the tail and helix 1, and between helices 1 and 3. These interactions are not native but are still consistent with an antiparallel helix bundle fold. Moreover, they contribute significantly to the stabilization of helical structure in the NCBD ensemble. For instance, interactions with helix 1 make the tail regain some of the helix structure that is suppressed by helix 2 (Figure 3.10). Transient interactions between helices 1 and 3, which were not found by NMR [62] also contribute to stabilize the three-helix bundled ensemble in the simulations.

3.4.6 Interaction Network and Cooperativity.

The top panel of Figure 3.11 shows the time-averaged native contacts observed in NCBD (bottom right) versus those on the LEGO elements (top left). These maps reveal that the H1H2 and H2H3 mostly recapitulate the patterns of native interactions present in the full NCBD, although, in these molecules, the contacts are somewhat less probable. However, in the NCBD ensemble, there seems to also be a significant number of non-native interactions, and which are longer range than the super-secondary structural patterns recapitulated in the LEGO elements (Figure 3.10 top). In the simulations, these long-range non-native interactions as the differential factor in cooperatively biasing the conformational landscape of NCBD.

To estimate the energetic contributions from each set of interactions, we resorted to the helix-coil parameters from the LEGO analysis (Figure 3.5 - Fig-

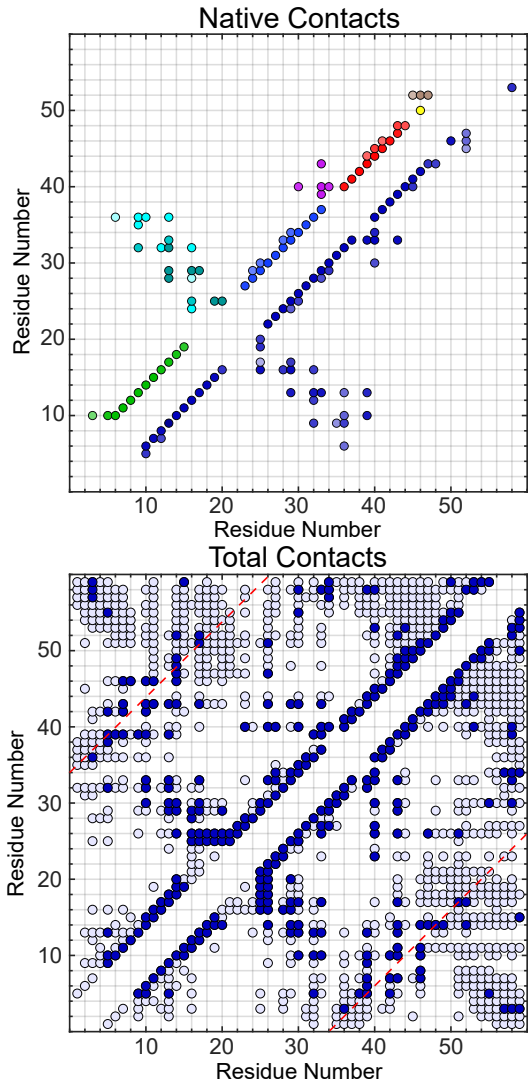

Figure 3.11: NCBD residue-residue interaction maps. Maps of the time averaged residue-residue contacts formed during the simulations. Top left triangle shows the native residue-residue contacts on all of the combined LEGO elements (local contacts shown in the color of the building blocks), and bottom right on the full-length NCBD. The color intensity reflects the time-averaged probability of observing the contact in the logarithmic scale, with the lightest color corresponding to a probability between 10^{-4} and 10^{-3} and the strongest intensity for probabilities between 10^{-1} and 1. bottom. total contacts (native and non-native) observed in the simulations of full NCBD. Contacts have been parsed in two groups: dark navy blue for contacts present at least 10% (>= 0.1 probability) and light navy blue for contacts present for at least 1% but less than 10%. The diagonal red dashed lines signal the maximum threshold for native interactions (<= i,i+34) defined as per the long-range NOEs reported in the NMR structure.

ure 3.9, Table 3.1). Using these parameters, we calculated the statistical weight for forming a full native α-helix conformation for each molecule. We then obtained the change in free energy from the ratio between the weight of a given combined LEGO element and the product of the weights of its building blocks (see Methods). This calculation can be performed for the experiments and simulations, thereby providing another comparative tool (Table 3.2). The experiments indicate that each set of pairwise tertiary inter-actions (helices 1-2 and 2-3) contributes $\approx$ 5-6 kJ/mol, which is comparable to the mean perturbation induced by single mutations on folded proteins [135]. The interplay be-tween helices 2, 3 and tail contributes $\approx$ 3 kJ/mol more. The overall NCBD stabilization amounts to $\approx$ 30 kJ/mol, which is comparable to the chemical denaturation free energies of many two-state folding proteins, even though NCBD is an IPDP. However, this comparison is misleading because the 30 kJ/mol for NCBD is in reference to a completely disordered ensemble (building blocks). In contrast, chemically unfolded states, especially of marginally stable/fast folding proteins, have large contents of local structure [27]. This procedure produces significantly stronger interactions in the simulations, particularly for pairwise tertiary interactions (helices 1-2 and 2-3).

To estimate the cooperative (non-additive) contributions we subtract the pairwise interactions from the NCBD total stabilization. This calculation leads to an experimental estimate of $\approx$ 17 kJ/mol, and of only $\approx$ 5 kJ/mol for the simulations (Table 3.2). The underestimation in the simulations could be due to imperfections in sampling and/or force-field. The total cooperativity presumably includes contributions from simultaneously forming interactions between helices 1-2 and 2-3, and from the non-native interactions (helices 1-3 and tail) that we see in simulations. The simulations also reveal that these sets of tertiary interactions compete with one another, resulting in alternating patterns. The conflict between tertiary interactions jointly with strong local propensities explains why NCBD does not form a unique structure but a broad, highly dynamic ensemble.

	$\Delta G_{exp}(kJ/mol)$	$\Delta G_{sim}(kJ/mol)$
H1-H2	-5.1	-25.1
H2-H3	-5.7	-10.2
H3-T	-1.0	-14.0
H2-H3-T	-8.6	-29.7
NCBD	-30.8	-59.5
Cooperativity	-16.6	-4.6

Table 3.2: Non-local energetic contributions. The change in free energy (ΔG) for given composite molecules (combinations or full protein) that is due to non-additive contributions (tertiary interactions) estimated from the σ and s parameters of the composite molecule relative to its building block elements from experiments and simulations. The cooperativity is obtained by subtracting the tertiary contributions for H1-H2 and H2-H3-T from the NCBD total change in free energy.

3.4.7 Discussion

Since IDPs were first identified, we have faced the challenge of explaining how these proteins integrate intrinsic disorder with the ability to select partners, fold upon binding, bind multiple partners, and switch among them in allosteric fashion. A key barrier has been the lack of suitable methods to dissect the conformational landscapes of IDPs in the absence of partners. Here we introduce a modular approach specifically designed to tackle this challenge (molecular LEGO), and apply it to investigate the folding landscape of NCBD, a partially disordered protein. The approach should in principle be easily generalizable to other IPDPs and hence it adds a powerful tool to the IDP re-search toolbox. In this regard, we outline some basic rules for applying the molecular LEGO to other disordered proteins: 1) A key element involves the design of the LEGO elements. The ideal scenario is to use a structural ensemble determined without partners using one of the existing approaches for applying the molecular LEGO to other disordered proteins:

- A key element involves the design of the LEGO elements. The ideal scenario

is to use a structural ensemble determined without partners using one of the existing approaches for generating IDP ensembles from experimental structural restraints [136],[137],[138]. An alternative could be a structure of the IDP folded when in complex with a partner. In the worst-case scenario, the design could be based on secondary structure prediction profiles.

- Since these proteins are disordered, it is convenient to use a structure-promoting cosolvent as thermodynamic variable. Inducing structure is also more significant to how these proteins fold upon binding. TFE is a good option, particularly for IDPs that form α-helical structure (whether free or upon binding). Other alternatives are osmolytes, such as betaine and TMAO, and salts, given that IDPs have very high net charges

- The conformational analysis should be carried out with techniques sensitive to the backbone conformation. Residue-averaged information is sufficient to address general mechanistic questions, as we show here using circular dichroism, or alternatively with infrared spectroscopy. NMR provides residue-specific structural information, but it is much more labor-intensive (especially for a complete analysis of all LEGO elements).

- To interpret the conformational biases of broad ensembles, it is essential to use a statistical thermodynamic treatment rather than assuming a two-state transition. The analysis could still be fairly simplified, but it should consider conformational entropy explicitly in terms of ensembles of microstates. In this regard, the molecular simulations allow the researcher to test the significance of the model used to analyze the experiments.

On a second front, the molecular LEGO study presented here sheds much needed light into key mechanistic questions related to the conformational behavior of IDPs in general, and NCBD in particular. Our results demonstrate that the amino acid sequence of NCBD contains strong local signals that prime the formation and define the limits of the native secondary structures formed in the ensemble. This observation supports the hypothesis that the conformational

behavior of IPDPs is connected to the energetics of downhill folding [29].

The LEGO multi-elements demonstrate that the few contacts observed by NMR in the full protein produce conformational biases that help maintain an overall helix bundle fold on the NCBD ensemble. However, these energetic contributions are relatively small (about 5-6 kJ/mol for each set of pairwise tertiary interactions: helices 1-2, and 2-3). From simulations we find that these native tertiary contacts form but are transient. These results explain the puzzling observation of specific long-range NOEs on an otherwise molten-globule-like ensemble [62].

The behavior of full NCBD relative to the LEGO elements provides other important clues about IPDP energetics. For instance, the tertiary interactions between helices 1-2 and 2-3 cooperate in the consolidation of NCBD's helical bundle fold (mostly via the stabilization of helix 2). However, we find that NCBD is much more ordered than expected from just its local and "native" tertiary interactions. Specifically, our experimental analysis reveals an extra of ≈ 17 kJ/mol stabilizing the NCBD ensemble. That is, the structural factors used to calculate the NMR structure (local conformation and NOEs) amount to less than 50% of the total ensemble energetics (Table 3.2). Our results reveal several such non-native factors. The C-terminal tail, which is fully disordered in the NMR ensemble, turns out to be a major player. The tail has intrinsic propensity to elongate helix 3 (see H3T, Figure 3.8), but the interactions of helices 2-3 impede such extension, and keep the tail disordered (H23T, Figure 3.8). The tail can also interact with helix 1, resulting on end-to-end contacts (Figure 3.11 right) that stabilize helix 1 and the formation of one turn of helix on the tail. This helix turn is disconnected from and bent relative to helix 3. In addition, the end of helix 1 interacts with the start of helix 3 in parallel fashion, which involves breaking many of the native helices 1-2 and 2-3 interactions. The pivotal role of the flexible tail is confirmed by comparing our results with previous simulations of NCBD with a truncated tail [30]. We note that all of these non-native factors can be inferred from, or are consistent with, the LEGO experiments. They are, however, most evident in the

simulations. This synergy highlights the importance of combining experiments and simulations in IDP research. Therefore, the picture that emerges from our dissection of the NCBD energy landscape is one of a protein with strong local conformational biases and a tug of war between sets of tertiary interactions, each stabilizing a distinct conformational sub-ensemble. Hence, the apparent disorder of NCBD arises from the conflict between competing tertiary interactions, which makes NCBD to dynamically alternate between sub-ensembles with slightly different folds. This behavior is in stark contrast with the usual interpretation of disorder as indicative of the lack of strong tertiary interactions. Remarkably, the conformational properties we find on NCBD uncover an internal mechanism that can drive its sophisticated, multi-partner, folding upon binding behavior. The 3D structure of NCBD in complex with p53-TAD [23] is fully consistent with the native sub-ensemble in which helices 1 and 3 interact with helix 2 but do not with each other, and the tail is disordered. These conformational biases are recapitulated by the LEGO elements H1H2, H2H3, and T. In contrast, ACTR binds NCBD by forming an intertwined complex in which the helices 2 and 3 of NCBD are set apart by ACTR, and helix 3 elongates onto the tail5, precisely as we see in H3T and H23T. Finally, the non-native interactions of helix 1 with helix 3 and tail are entirely consistent with the structure that NCBD forms in complex with the stably folded IRF3 [22]. Summarizing, the NCBD folding landscape has built-in energetic biases that compete for stabilizing the various conformational sub-ensembles that NCBD forms in complex with structurally diverse partners. This behavior un-covers an internal folding mechanism to select partners and modulate affinity that is likely essential for NCBD's recruiting role as transcription coactivator [6], indicating that molecular LEGO can be used to detect subtle energetic biases on IPDPs that are key to their biological function.

CHAPTER 4

Decoding Conformational Rheostats in Transcription: Morphing Coupled to NCBD Binding

In the previous Chapter, we focused on developing a novel tool, Molecular LEGO, for measuring the folding cooperativity and the energetic contributions of native interactions of IPDPs. We demonstrated that this approach solves the problem of mapping the energetics of partially disordered proteins, can readily be extended to other IDPs and offers the possibility to design IDPs for future engineering applications. However, the underlying structural heterogeneity of its complex native ensemble is not discussed in detail and further requires in-depth dynamical characterization. In general, interactions/transitions and competitive binding processes involving IDPs cannot always be described by classical mechanisms. To this end, this Chapter explores the conformational dynamics of an IDP native ensemble based on our hypothesis that many proteins classified as IDPs and implicated in a range of regulatory and signaling biological functions operate as conformational rheostats. Here we investigate the mechanistic underpinnings of what we have identified as a putative conformational rheostat playing a pivotal role in organizing the eukaryotic transcription complex: the nuclear coactivator binding domain (NCBD) of CBP (CREB Binding Protein) and its binding to other morphing partners such as the transactivation domain of p53 (p53-TAD) and ACTR, or well-folded proteins like IRF3.

4.1 Abstract

Conformational rheostats are defined as protein domains that naturally populate a broad, non-random conformational ensemble that gradually morphs onto different structures in response to cues, such as binding to multiple partners. Many intrinsically disordered proteins (IDPs) interact with numerous partners and frequently function as molecular hubs in protein interaction networks. Despite their growing repertoire of biological roles, the molecular mechanism that enables such promiscuous binding and morphing behavior remains largely unexplored. Previously we proposed many partially disordered proteins operate as conformational rheostats, which allow their diverse functioning. Here we performed extensive all-atom MD simulations of NCBD, an IDP characterized in substantial detail and known to display complex dynamical behavior. We obtain results on NCBD in its free and bound complex forms through a detailed structural characterization to investigate the dynamical features of a potential conformational rheostatic behavior. Our results reveal the hidden conformational biases in the dynamics of the native heterogeneous ensemble of NCBD in the absence of its partners. We find structural heterogeneity arising from the timescales of both local (short-) and non-local (long-range) interactions; the NCBD ensemble showcases less flexibility on tens of nanoseconds timescale and samples a broader ensemble on longer timescales. We demonstrate that NCBD populates sub-ensembles that distinctly resemble the various bound complexes in terms of topology and secondary structure, undergoing gradual conformational transitions, hinting at a working conformational rheostat in transcription.

4.2 Introduction

The emerging role of conformational disorder in protein folding and binding and its physical relevance has been highlighted across multitudes of biological processes and is implicated in many human diseases [10]. A large fraction of the proteome is now believed to contain naturally unstructured domains in their

functional states [2]. These proteins, in isolation, exist as broad, non-random, conformational ensembles of interconverting states rather than a unique structure [1],[7]. Certain disordered proteins adopt folded structures upon binding to their multiple partners, which has aided progress in determining how the Intrinsic flexibility of these proteins is advantageous from a functional or evolutionary point of view. These showcase the importance of conformational plasticity and heterogeneity in protein function, mostly attributed to high specificity for multiple targets and low-affinity binding, which allow them to operate as morphing proteins. Many such partially disordered states sample transient yet specific, secondary, and tertiary interactions. The question then becomes what fundamental molecular mechanism allows these proteins to morph onto different conformations. Indeed, this dynamical behavior cannot entirely be explained using conventional binary transitions rather a more gradual (un)folding behavior. Therefore, we previously hypothesized that many intrinsically partially disordered proteins (IPDPs) perhaps act as conformational rheostats (CRs) to support a continuum of conformational states and transitions tuned by diverse binding modes that regulate dynamics between subpopulations and subsequent ligand binding [29].

The preformed residual structures found on IPDPs are typically investigated using NMR experiments. However, the IPDP's inherent flexibility may bias the structural determination to one or a few structures because any observable will be averaged over a heterogeneous ensemble of structures. In this regard, MD simulations can probe these morphing transitions and provide the mechanistic insights needed to interpret the inevitably lower resolution experimental data that are obtained with techniques suitable for well-structured proteins. MD simulations were also not particularly well suited for the analysis of IDPs, but recent technical developments have led to simulations that capture the dynamical properties of IDP more accurately [45],[46].

One particularly interesting IPDP is the Nuclear coactivator binding domain (NCBD) from the CREB binding protein (CBP) that helps recruit the basal transcription machinery and binds to multiple structurally diverse partners. The

structural and binding properties of NCBD have been of much interest in the past decade. The intrinsic disorder underlying NCBD has been probed via extensive biophysical experimental techniques and MD simulations, highlighting its molten globule-like features [62],[17], heterogeneous ensemble [30], and folding coupled to binding to its multiple partners with distinct affinities [23],[22] . Here we propose that NCBD operates as a CR. The wealth of biophysical data provides us a unique opportunity to examine NCBD in a new light, to gain possible insights into its morphing behavior. To this end, we performed ≈ 60 μs of all-atom MD simulations of NCBD in its NMR characterized free form and various bound forms to delineate its detailed structural characterization and dynamics. Our results reveal the hidden conformational biases in the dynamics of the native heterogeneous ensemble of NCBD in the absence of its partners. We find structural heterogeneity resulting from the timescales of secondary (local) and tertiary (non-local) interactions in defining its intra-molecular interaction network. We observe less variability on tens of nanoseconds timescales and a broader ensemble on slower timescales. We then derive its kinetic map using Markov state modeling and compare the structural features of the most populated sub-states with the known NCBD structures. Intriguingly, NCBD populates sub-ensembles that resemble the various bound complexes in terms of topology and secondary structure, which undergo morphing transitions, hinting at a working conformational rheostat in transcription.

Background

NCBD, a well-characterized IPDP, has been probed via extensive experimental and computational techniques. This section covers the most relevant literature on known NCBD structural forms (free and in the complex) and its native conformational ensemble to provide a fundamental background for our proposed work.

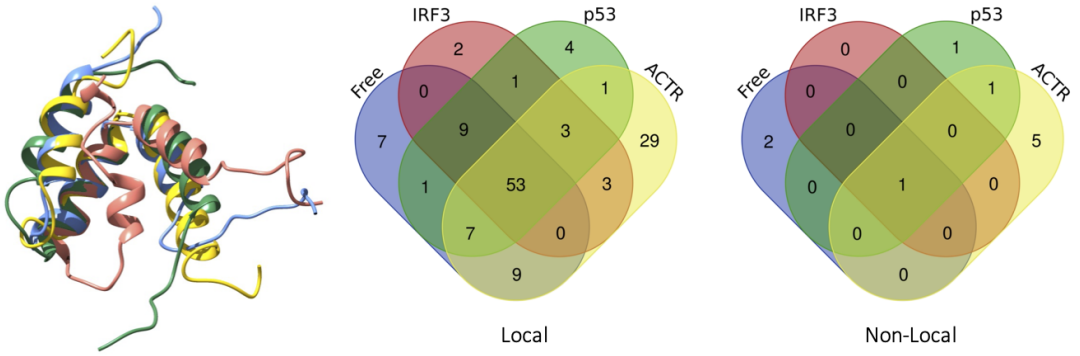

Figure 4.1: Structural alignment and topological variations in NCBD structures. Local (left) and non-local (right) interactions among known NCBD structures. Venn diagram with each enclosed curve representing NCBD structure (free form and bound). There are 53 core local contacts and one non-local contact common among these four known structures. Local contacts are defined as contacts between Cα atoms within 0.65 nm and are less than five residues apart, and non-local contacts are five residues above in sequence.

4.2.0.1 Known NCBD structures and Complexes

NCBD, in its free state, is compact and has a high degree of helicity reminiscent of a molten globule but does not exhibit cooperative thermal unfolding [62]. The NMR-determined structure of NCBD comprises of a three-helix bundle (Figure 1.2), and structures of NCBD in complex with its diverse interaction partners reveal large topological variations as illustrated in Figure 4.1, with significant different arrangements of the helices as in the IRF3 and ACTR bound states.

In characterizing the structure and affinity of the p53TAD and NCBD complex, Lee et al. performed heteronuclear multidimensional NMR and isothermal titration calorimetry (ITC) of various lengths of p53-TAD constructs with NCBD and found the affinity of full-length p53TAD (1-61)/NCBD $\approx$ 1.7 μM. In the complex structure, the p53-TAD comprises two helical regions that dock into the exposed broad hydrophobic groove formed by the three NCBD α-helices (Figure 1.2). The binding interface of TAD and NCBD consists of hydrophobic interactions, although both molecules are highly charged [23].

The structure of the complex between the NCBD and ACTR was determined by NMR (Figure 1.2). The ACTR helices almost completely encircle helix 3 of NCBD, and this assembly of two proteins forms a rich hydrophobic core. Both IDPs combine with high affinity to form a folded helical heterodimer with a dissociation constant of $\approx$ 34 nm measured by ITC [17]. Further studies using NMR spectroscopy of ACTR:NCBD show that the unbound ACTR is devoid of any long-range intramolecular contacts. The free NCBD retains most of its helical content as seen in the complex but is largely flexible [139].

The crystallographic NCBD-IRF3 bound complex (Figure 1.2) comprises a hydrophobic interface of $\approx$ 180 nm^2 [22]. Helices H3 and H4 of IRF3 form the major interaction surface for the NCBD. Using ITC, the dissociation constant for the NCBD:IRF3 interaction was determined to be around 100 μM indicating, a lower affinity than ACTR binding but in the same range as for p53-TAD [140].

4.2.0.2 Free from folding ensemble

A study on replica exchange MD simulations of free NCBD, initiated from an unfolded state, revealed a large amount of residual secondary structure in its native ensemble. The simulations show that the most common pair of coexisting helices are I and III, with the II helix being rarely structured (Figure 1.2.A). The middle helix (II) is less critical for binding ACTR. However, it is implicated in most of the intra-protein contacts in the ACTR bound conformation of NCBD (Figure 1.2.B), overall indicating a scenario of a preformed binding interface in the unbound ensemble [63]. On the other hand, a study using implicit solvent and replica-exchange sampling shows that the free NCBD appears to sample distinct conformations that have been observed experimentally in complexes. Overall simulations observe that free NCBD is highly compact. Although the poly-Q segment of NCBD (residues 2082-2086, part of the II helix) is disordered in the NCBD:ACTR complex, it is highly helical in the unbound state [115]. Naganathan et al. studied the free form NCBD (Figure 1.2.A) with a hierarchy of models, ranging from Ising-like models, Go-model, and explicit solvent atom-

istic simulations (without the NCBD C-terminal tail), and showed that NCBD displays many conformational properties of being a 'global downhill folder' and samples a heterogeneous native ensemble consisting of conformations reported in other bound-forms [30].

All the above computational studies find that none of the structures (clusters) sampled in the unbound ensemble closely resemble the NMR unbound/free form.

4.3 Methods

Atomistic Molecular Dynamics Simulations

Based on the setup described in Chapter 3, we performed GPU-accelerated 3*12 μs long trajectories at 310 K. In addition, 3 sets of simulations, 9 μs long each, were produced starting from the bound NCBD conformation of each of the three NCBD complexes, ACTR (PDB ID: 1KBH), p53-TAD (PDB ID:2L14), and IRF3 (PDB ID: 1ZOQ), in the absence of these partners using the same protocol. The NCBD conformer bound to IRF3 has 12 residues missing (N- and C-terminal ends), which were modeled using the i-Tasser protein structural modeling tool [141]. For all analyses, the first 200 ns of all trajectories were discarded.

Native probability contact map. The time-averaged native probability contact map has been obtained from the simulation using the following definition of native contact: when the distance between any heavy atom of the two interacting residues (>3 apart) is less than 5.5 Å in the native NMR structure (PDB ID: 2KKJ). The contact map was calculated using a total of 100,000 frames (every 100 ps) from the trajectory to compare with the Glutton-derived contact map.

Local and non-local native fraction The number of native contacts per residue was calculated from each MD trajectory using the NMR structure as the reference of native contacts. Contacts were defined using a 0.5 nm cutoff between any two pairs of heavy atoms that are at least 3 residues apart in the sequence for local and more than 5 residues apart for non-local. The number of these contacts

trajectories were then converted into the fraction of local and non-local native contacts repectively.

Time-averaged probability contact maps. We determined the time-averaged probability of finding each contact (whether native or non-native) in three NCBD free trajectories between two residues that are at least 3 apart in the sequence. A contact is considered formed at any given 10 ns interval when at least one heavy atom of residue i is within a cutoff distance of 0.5 nm of at least one heavy atom of residue j (where j >= i+3) with a probability > 0.7 during such time interval.

Bound specific contacts in NCBD free form. We first calculated all possible native contacts in PDB structures of NCBD bound to its three partners and free form. Then evaluated the unique list of contacts among all native contacts in the three bound conformations (bound specific) that are not present in the free form and computed the time trajectories of NCBD free form with reference to bound specific contact list (every 1 ns time frame). A contact is considered formed at any given 10 ns interval when at least one heavy atom of residue i is within a cutoff distance of 0.5 nm of at least one heavy atom of residue j (where j >= i+3) with a probability > 0.7 during such time interval.

Timescales of C_β contacts. We first extracted the list of C_β contacts in the NMR free form structure. A contact is considered formed when the minimum pairwise distance between C_β atoms of the interacting residues across the two proteins is <=0.6 nm and the residue pair is >3 residues apart in the protein sequence. There are 40 C_β contacts in the reference structure. Next, computed the time trajectories of the distances and the autocorrelation function of the distances with a lag time of 6 μs (Figures in Appendix). The characteristic timescales are calculated based on ACF $\approx$ 0.5. We then segregated these 40 contact indices based on sequence separation between each contact pair into short (below 5 residues apart in the sequence), mid (5-10), and long (greater than 10) range contact. The error bars indicate the standard error of the three NCBD trajectories. Similarly, we use the 40 C_β contacts reference list to evaluate the

timescale vs. the sequence separation for the NCBD unbound trajectories of ACTR, p53-TAD and IRF3.

Principal Component Analysis (PCA). PCA was performed using the PCA function in MATLAB. First, we evaluated the center of mass (COM) distances between each helix pair, H1-H2, H2-H3, H1-H3, using the gmx distance tool in GROMACS. Then, the obtained distances were transformed into PC space to capture the variability in terms of the helical arrangement. We used the principal component 1, which accounts for 65% of the total variance, for further analysis.

Conformational Landscapes of NCBD. The map was obtained from the normalized probability distribution as a function of the relevant set of order parameters. The probability distribution was converted into an energy scale using the following expression:

$$\Delta A_{ref \to i} = -RTln(\frac{P_i}{P_{ref}}) \tag{4.1}$$

where the probability of going from a reference state (ref) of the system to any state i (e.g., from folded to unfolded) at constant temperature and constant volume is evaluated. R is the ideal gas constant, T is the temperature, and p_i and p_{ref} are the probabilities of finding the state i and state ref system, respectively. We project the conformational space onto two order parameters: PC1 and the fraction of native contacts (Q). In this calculation, a contact is considered formed when the minimum pairwise distance between atoms of the interacting residues is <=0.5 nm and the residue pair is >3 residues apart in the protein sequence. Conformations collected at 100 ps intervals were projected onto the Q–PC1 plane using a 32 × 32 grid (1024 cells) and sampling statistics were compiled to evaluate Equation 4.1. The grid cell with the largest population was used as reference state.

Markov state modeling (MSM). MD simulations are difficult to analyze because the obtained result—a set of trajectories recording each particle's cartesian coordinates—can contain numerous data points in a large number of

dimensions. To this end, a quantitative statistical model of the structure and dynamics of the system is of interest. The model should be able to describe the long-timescale processes in the data and should be interpretable. In this regard, Markov modeling offers all of these desired properties to dissect the underlying dynamics of biomolecular systems [142],[143],[144]. For disordered proteins, transitions between metastable states are typically fast and not always accompanied by significant changes in the protein's overall structure. In comparison to structured proteins, dividing the space into discrete states can be challenging. In the case of NCBD, based on our thorough investigation in Chapter 3, we found marked structural alterations in its free ensemble, so we decided to leverage MSM [145],[146] to characterize its conformational kinetics using the PyEMMA suite [145]. We performed the MSM analysis on the positions of backbone atoms. Then, used the time-independent component analysis (TICA) to identify slow degrees of freedom, i.e., find a projection that maximizes the autocorrelation functions in the TICA space to filter out the fast dynamics in the space of the positions of the backbone atoms [147]. We used a lag time of 15 ns and 95% variance to account for the number of dimensions for the TICA subspace. TICA is based on the following equation to solve for the eigenvalues (λ) and eigenvectors (v):

$$C^{(\Delta t)}v = \lambda \Sigma v \tag{4.2}$$

where $C^{(\Delta t)}$ is the time lag correlation matrix defined by:

$$C_{ij}^{(\Delta t)} = \mathbb{E}\left[X_i(t)X_j(t+\Delta t)\right] \tag{4.3}$$

We discretized the TICA space into 100 clusters using the k-mean clustering algorithm and were found to represent the slowest modes reasonably well (Figure 4.13).

Further, we built a Markovian transition matrix of 100X100 microstates by counting for all possible transitions between any two clusters using a lag time $\tau =$

7 ns based on the implied timescales of the 10 slowest processes (Appendix). To make the obtained MSM interpretable, we used the hidden Markov model (HMM) algorithm [148] for grouping the 100 clusters into few macrostates. Further, to visualize the kinetic relationship between these functionally relevant states, we estimated the average transition times based on the emission probabilities of the derived HMM transition matrix of size 7X7 using equation:

$$\omega = \frac{-\tau}{\ln(1-p)} \tag{4.4}$$

where ω is the average timescale for the transition.

We extracted 10,000 snapshots from each macrostate and superposed all frames onto each of the reference structures (four experimentally determined NCBD structures) to assess structural diversity in the macrostates. We then calculated the mean backbone RMSD of residues ranging from 6 to 47 and the standard deviation shown in Figure 4.14.

4.4 Results

NMR structure determination methods aim to determine the high-resolution 3D structure (or closely related structures) that simultaneously satisfies all the experimental restraints, most typically Chemical Shifts (CS) and Nuclear Overhauser Effect (NOE). This approach is inadequate to analyze IDPs or partly disordered proteins (e.g., folding intermediates) because in these cases, the NMR parameters, most notably CS, are averages over an ensemble of conformations that is highly heterogeneous but has strong native-like structural biases. The need for specific analytical methods that interpret conformational preferences rather than structures is pressing, especially given the key roles that disorder plays in protein folding, binding, and function. The most serious difficulty resides in constructing structural ensembles that are entirely consistent with the experimental data and capture the underlying conformational heterogeneity. To this end, we have implemented a novel relational database, termed Glutton, that links all existing

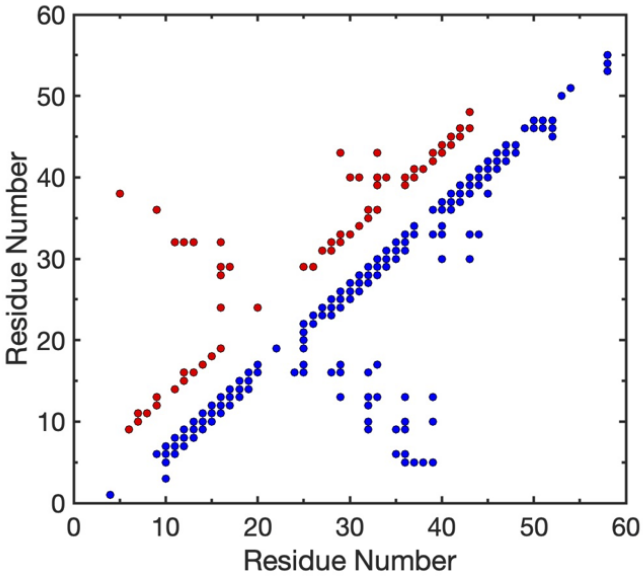

Figure 4.2: Contact map of the lowest energy NMR structure and associated NMR distance restraints (NOEs) in blue and red respectively.

CS data with corresponding protein 3D structures to enable the conformational analysis of IDPs directly from their experimental CS [138]. Glutton's uniqueness focuses on dihedral angle distributions consistent with a given CS set rather than with unique structures. Such dihedral distributions define how native-like the ensemble is and lead to the practical calculation of large ensembles of structures that efficiently sample the available conformational space. The structural ensembles obtained from Glutton are based on geometric considerations and CS.

The NMR experimental conditions of NCBD permitted a small number of distance restraints (NOEs) that sufficed to determine a native structure using standard procedures to assess high-resolution NMR structures [62]. This NMR structure is most likely over-fitted relative to the structural ensemble populated by NCBD at physiological conditions, and the standard structure determination protocol enforced that all restraints be satisfied simultaneously as unique distances rather than as distance distributions. To showcase this, we compare the NOEs with the residue-residue contacts formed in the NMR structure (PDB ID: 2KKJ) within a 0.5 nm cutoff, as shown in Figure 4.2. In this regard, Glutton provides a reliable way to compare the NCBD ensemble generated by MD sim-

ulations starting from the NMR structure with the ensemble derived from the statistical distribution of dihedral angles based on chemical shifts.

4.4.1 Free NCBD Conformational Ensemble

To investigate the conformational properties of the NCBD native ensemble, initially, we performed a 10 μs long all-atom MD simulation of NCBD in explicit solvent starting from the NMR structure and using the Charmm22* (c22*) force field [52]. The c22* forcefield has been shown to describe ensembles of partially disordered proteins in relatively accurate agreement with the NMR experiments in terms of secondary structure propensity and topology. As a first step, we evaluate the conformational bias in the NCBD NMR structure and the MD sampling statistics. We examine the MD conformational ensemble compared to the structural ensemble generated from the Glutton database that is specifically designed to generate conformational ensembles of partially disordered proteins from NMR chemical shifts. The Figure 4.3 (left) compares the time-averaged native contact map of the MD ensemble, and the contact map averaged over the entire Glutton ensemble obtained from the experimental CS [1].

The MD ensemble reproduces the highly dynamical properties observed in Glutton, including a high probability for the local, α-helical contacts and a much lower probability for long-range contacts. The patterns of contacts are also very similar, including transient contacts between helices 1 and 2 and no tertiary contacts elsewhere, even though the NMR structure has multiple contacts between the three helices. Figure 4.3 (right) compares the mean phi and psi angles from the MD simulations and the Glutton ensemble, showing considerable agreement overall. The entire phi-psi distributions demonstrate the level of agreement in more detail. Figure 4.4 shows two examples L10, which remains helical in both ensembles; and A49, which visits the α and β regions.

[1] Yi He, Suhani Nagpal, Mourad Sadqi, Eva de Alba, Victor Muñoz, Glutton: a tool for generating structural ensembles of partly disordered proteins from chemical shifts, Bioinformatics, Volume 35, Issue 7, 01 April 2019, Pages 1234–1236, https://doi.org/10.1093/bioinformatics/bty755

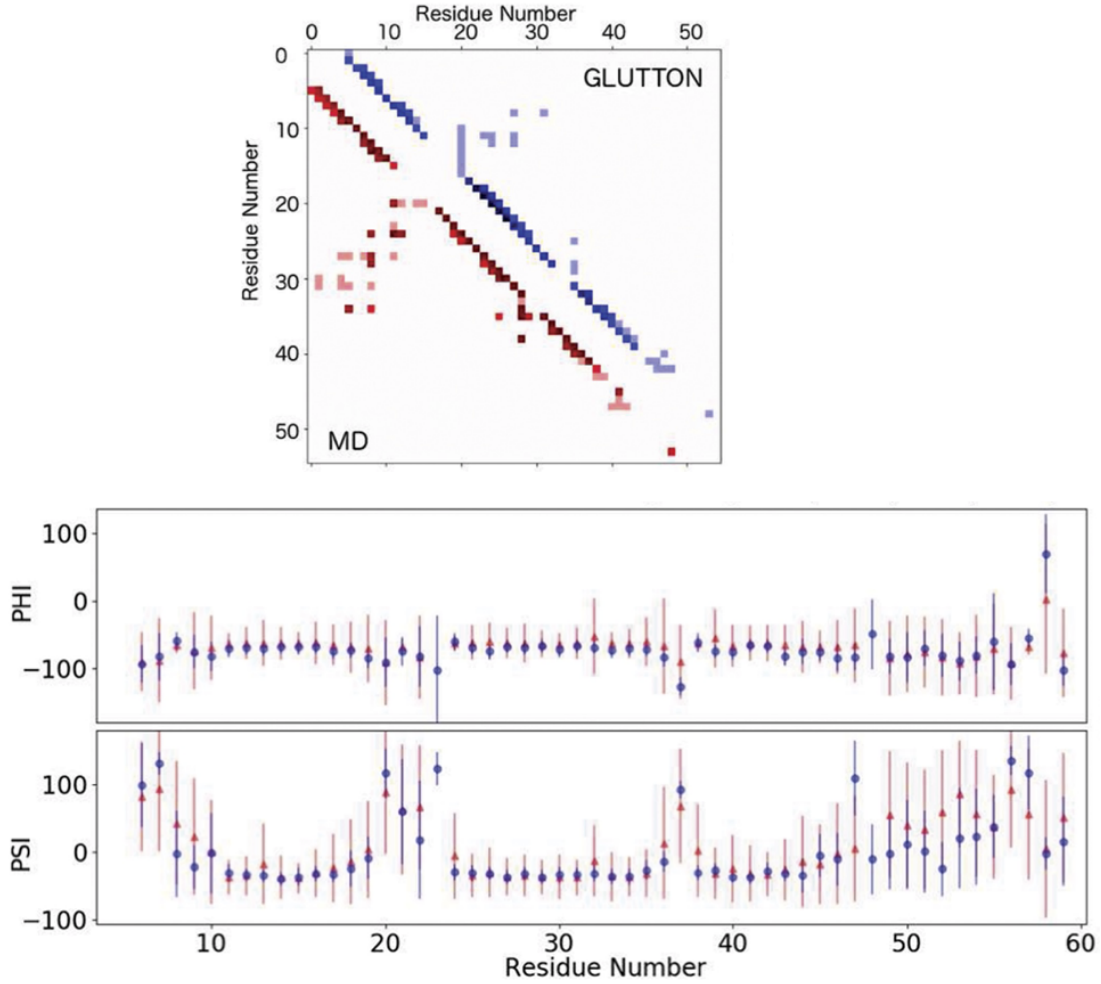

Figure 4.3: Time-averaged (bottom left triangle) and ensemble-averaged (upper right triangle) native contact map of NCBD. Darker indicates closer to 1 and lighter closer to 0. (C) Mean and standard deviation of $\phi\psi$ angles for NCBD from Glutton (circles) and MD (triangles). Figure reproduced with permission, copyright © 2018, Oxford University Press.

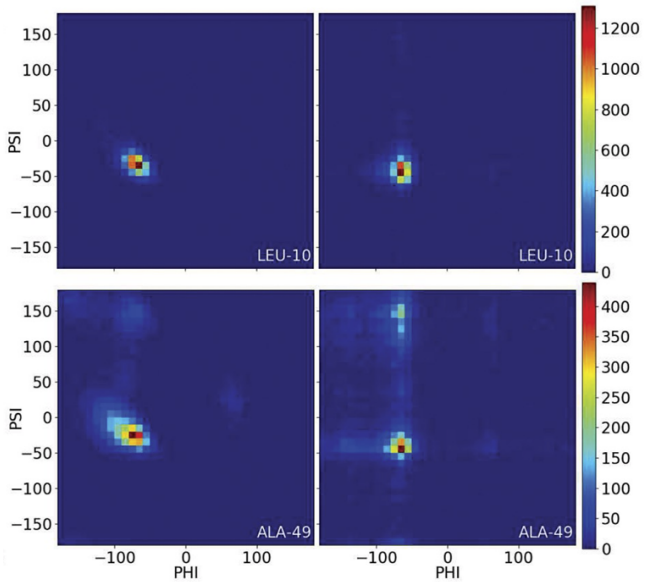

Figure 4.4: Examples of $\phi\psi$ angle distributions of NCBD residues from MD simulations (left) and Glutton (right). Figure reprinted with permission, copyright © 2018, Oxford University Press.

These examples highlight how each ensemble captures some residues' restricted conformation and more heterogeneous conformations of others. Overall, this analysis reveals that NCBD populates a highly dynamic ensemble with native-like features and is consistent across both ensembles. This observation also corroborates our choice of the c22* force field. Also, it is noteworthy to point out that the regions populated on the Glutton structural ensemble are co-centered with the corresponding regions populated on the MD trajectory. Still, for some residues, the regions populated on the Glutton ensemble are also wider, indicating that more extended and multiple MD simulations are needed to achieve better sampling statistics.

4.4.2 NCBD as a Model to Investigate Conformational Rheostat Mechanism

To explore the concept of NCBD operating as a potential conformational rheostat, we performed three sets of MD simulations totaling 36 μs starting from the NMR

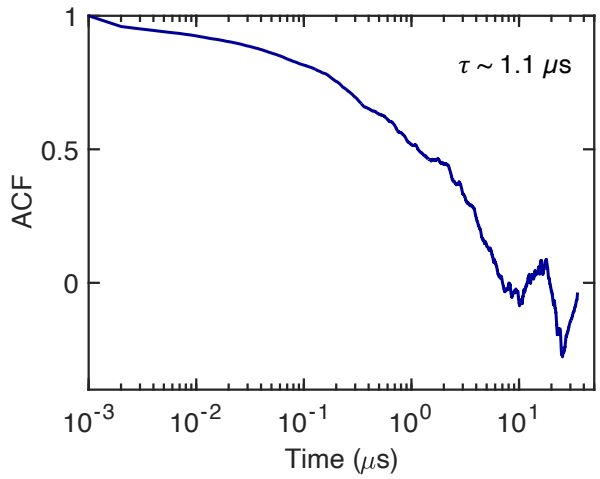

Figure 4.5: Autocorrelation function of the fraction of native contacts (Q_{free})

structure in explicit solvent at 310 K. We determined the autocorrelation function of the fraction of native contacts (Q_{free}) sampled across the three MD trajectories to assess the sampling quality.

Figure 4.5 shows an apparent characteristic time of $\approx$ 1.1 μs which translates to 33 statistically independent Q_{free} values [149], [150] and demonstrates the reliability of the trajectories for further probing NCBD morphing dynamics. This observed relaxation decay for NCBD is consistent with the reported relevant timescales for a catalog of fast-folding proteins that undergo microsecond folding kinetics displaying smooth structural disorder [26],[86],[151],[152]. In addition, we produced a total of 27 μs long all-atom MD trajectories of each of the three bound NCBD ($NCBD_{bound}$) structures in the absence of their partners (ACTR, p53-TAD, and IRF3, see Figure 4.1) to assess how fast they re-equilibrate to the free form strucural ensemble. Figure 4.6 shows similar relaxation times for the $NCBD_{bound}$ trajectories with reference to the native contacts in the NMR structure (free form), indicating that the overall sampling is not biased towards the starting structure and that the obtained ensembles have characteristic features of a downhill (un)folding behavior.

We then monitor the obtained free from MD ensemble with various structural probes to characterize its dynamical properties as shown in Figure 4.7. The probability distributions as a function of RMSD and radius of gyration (R_g) display

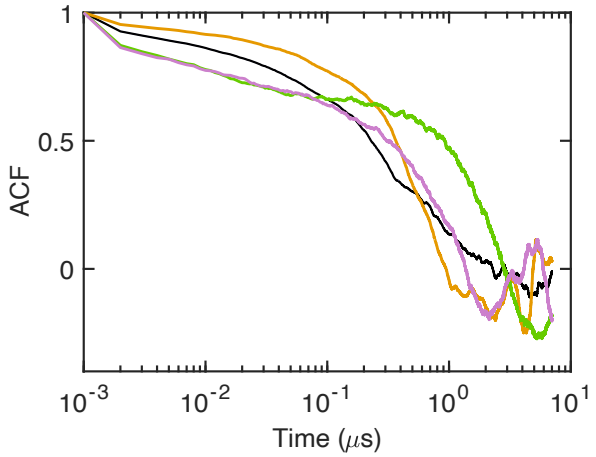

Figure 4.6: Autocorrelation function of the Q_{free} for the NCBD bound structure of ACTR (yellow), p53-TAD (green) and IRF3 (pink). Black profile represents the average behavior of NCBD free ensemble for reference.

a unimodal distribution in which conformations are sampled over a wider range along the RMSD than R_g, indicating a compact ensemble with high structural variability at physiological temperature. Given that NCBD has molten globule-like properties, the observed compaction is not surprising. Although the compact MD ensemble is slightly comparable to that obtained previously from REMD simulations ($R_g \approx 13.7$ $\AA$) at 304 K [63] but is lower than the estimated experimental value ($R_g \approx 15.5$ $\AA$) from Small Angle X-ray scattering (SAXS) measurements in solution [62]. This discrepancy between experiment and simulation can be ascribed to both force field deficiencies (overly large compaction of proteins is a common problem) and a lack of thorough understanding of the role of solvation on the SAXS characteristics of partly disordered proteins [53]. Still, the derived ensemble highlights the molten globule-like characteristics.

The distribution as a function of the fraction of local native contacts shows a similar pattern undergoing fast dynamic fluctuations, and the major conformational changes are primarily due to the fraction of non-local (tertiary) contacts, which varies from 0.1 to 0.4. This fractional contribution from local and non-local native interactions provides an interesting parameter to estimate the intrinsic stability of partially disordered proteins relative to natively folded domains from the

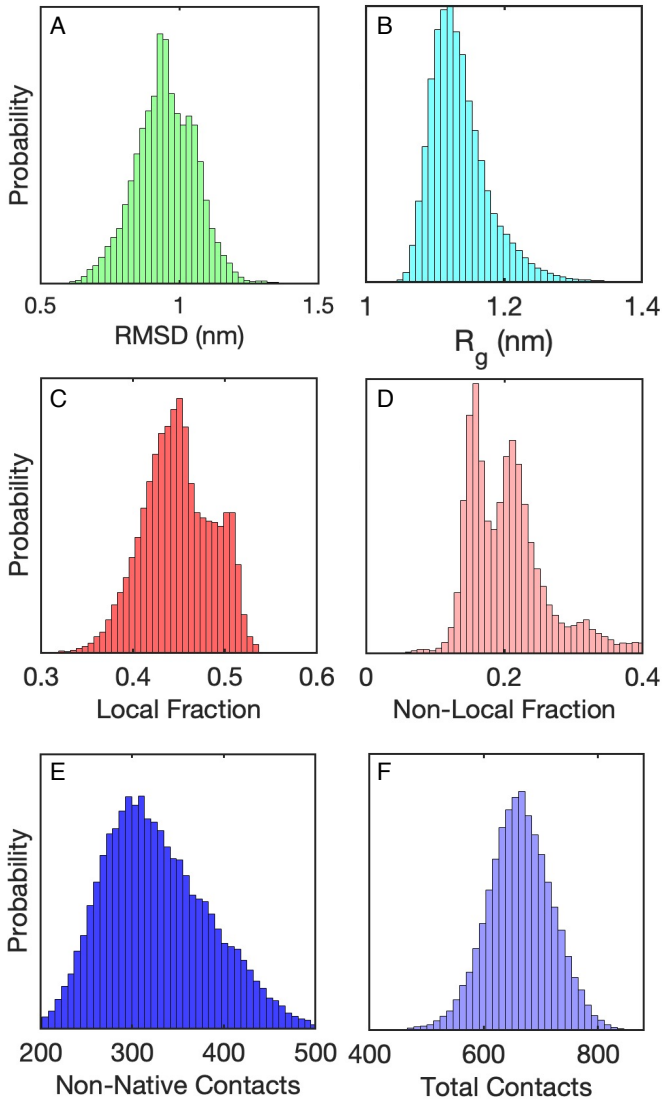

Figure 4.7: Structural properties of NCBD free ensemble. Probability distributions from one-dimensional projections as a function of (A) root-mean-square deviation, (B) radius of gyration, (C) local native fraction and (D) non-local native fraction, (E) non-native contacts and (F) total contacts.

perspective of folding cooperativity.

In comparison to folded proteins, which have a non-local fraction higher than 0.55 and a local fraction lower than 0.33, we observe a higher local fraction of 0.45 and a relatively low non-local fraction of 0.2, which conforms to the gradually (un)folding scenario expected for IDPs [29]. The unimodal distributions observed as a function of numbers of non-native and total contacts sampled further illustrate the gradual conformational behavior of NCBD, where total contacts are the amount of native and non-native contacts. Overall, the one-dimensional projections along these multiple parameters exhibit a heterogeneous yet compact conformational ensemble undergoing morphing transitions at physiological temperature.

In Figure 4.8, we plot the time-averaged probability contact map of the NCBD MD ensemble (averaged over three trajectories). A contact is defined based on the cutoff of 0.5 nm with a threshold that for at least 70% of 10 ns, residues i and j remain in contact (<=3 residues apart in the sequence). The total interaction matrix captures the magnitude of its structural disorder at physiological temperature. Examining the entire map compared to the native contacts in the NMR structure provides critical insights into the nature of its underlying structural heterogeneity. The applied threshold removes the highly transient contacts from the analysis. The three helices show a high probability of secondary structure formation (in logarithm scale) and significant diversity arising from non-local interactions. There are multiple weak non-specific long-range interactions sampled between helix 1 and the C-terminal disordered tail, helix 1-2 (H1-H2), and helix 1-3 (H1-H3). We observe numerous strongly interacting residues across the structure such as L13, F43, and Y51 and find the tail communicating with all the structural elements highlighting the dynamical conformational biases in the NCBD ensemble.

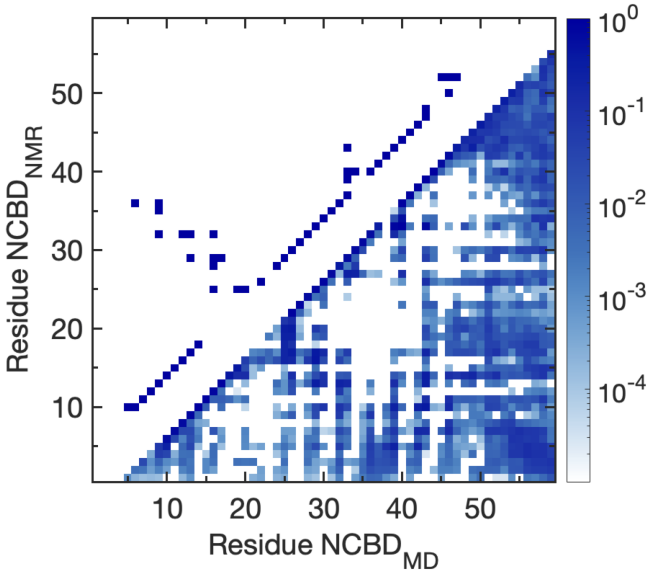

Figure 4.8: Contact map of NCBD. NMR free form structure (above triangle) and the time-averaged total interaction matrix of the MD ensemble comprising of all native and non-native interactions (below triangle).

4.4.3 Intramolecular Interactions Specific to Bound Conformations

To better understand the extent of the observed morphing behavior, we evaluated the occurrence of any contact specific to the known bound conformations of NCBD (Figure 4.1) sampled across the entire NCBD free ensemble. To do this, we first calculated the common contacts between the free form and all the bound conformations. Then monitored each of the trajectories with reference to only the unique list of 1027 contacts that are only formed in the native bound conformations, and the results are shown in Figure 4.9. We find all three MD trajectories sample on an average $\approx$ 150 intramolecular interactions specific to the bound conformations as a function of time. These multiple structural fluctuations are dynamic and reversible over the simulation time and suggest that free NCBD can morph onto conformations with structural features relevant to the bound structures of ACTR, p53-TAD, and IRF3 complexes. Moreover, the probability distribution hints at a gradual morphing pattern.

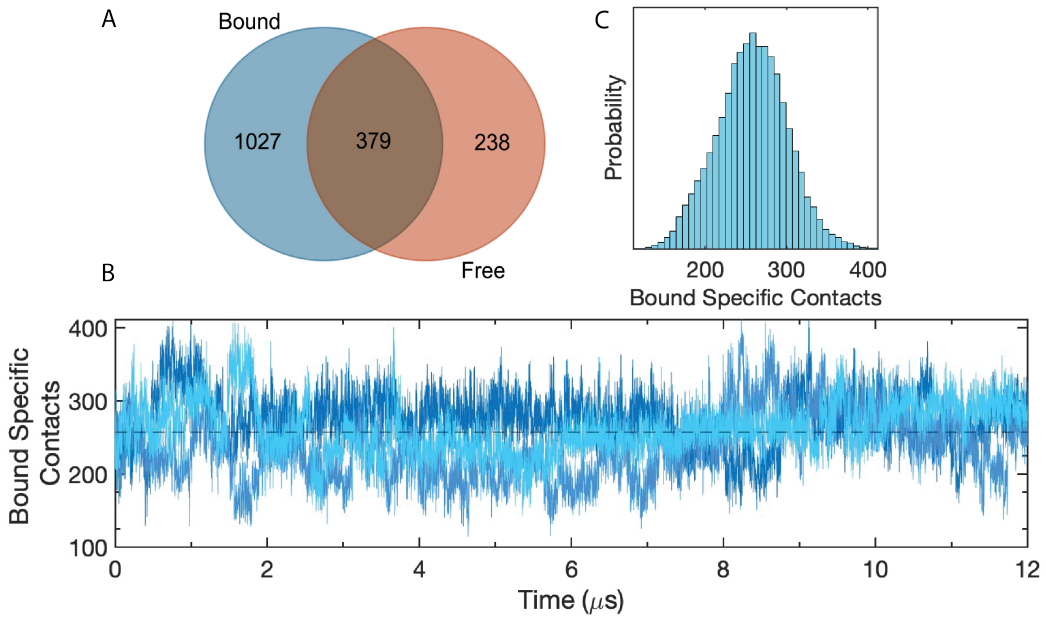

Figure 4.9: A. Venn diagram of native contacts found in NMR free form structure and all three bound structures. B. Time evolution of bound specific contacts across three $NCBD_{free}$ trajectories with respect to the unique list of 1027 contacts that are not found in the NMR free form structure within 0.5 nm. The black dashed line indicates the average number of the contacts sampled. C. Probability distribution of the three trajectories as a function of the bound specific contacts.

4.4.4 Timescales of Secondary and Tertiary Contacts

To further quantify the level of structural heterogeneity in the underlying NCBD conformational ensemble, we examine the dynamics of intramolecular interactions in detail. First, we analyzed the time evolution of the distances between center-of-mass (COM) of C_β contacts across the trajectories with reference to the C_β contacts present in the free from structure (0.6 nm cutoff). There are 40 C_β contacts in the reference conformation (table in Appendix). The Figure 4.10 depicts the time evolution of distances between each contact pair COM for one of the representative trajectories. We find many contact pairs moving further apart to almost 2 nm in the atomic distance and multiple reversible transitions along the simulation length. Next, evaluated the autocorrelation function of the distances of each contact pair across all three MD trajectories to estimate their characteristic timescales (Appendix). We then segregated these 40 contact indices based on sequence separation between each contact pair into short (below 5 residues apart in the sequence), mid (5-10), and long (greater than 10) range contact. Figure 4.10 shows the apparent relaxation timescale as a function of sequence separation for each contact pair. We observed heterogeneity in the dynamics of short- and mid-range contacts ranging from a few nanoseconds to almost more than 1000 ns and long-range contacts ranging from 100 ns to 1000 ns. Overall, the analysis of native C_β contacts captures relatively low conformational variability on early nanoseconds timescales and a broader ensemble on longer timescales, providing insights into the complex dynamical properties of NCBD. This observation that NCBD is mostly rigid in fast timescales and shows greater motions at slower timescales was also recorded in a recent MD analysis on NCBD using the c22* force field combined with NMR parameters [53] and is consistent with our findings as in Figure 4.10 and Figure 4.5.

4.4.5 Structural Rearrangement in NCBD Free Ensemble

To characterize the topological variation in the context of the three α-helices, we calculated the distances between the COM of each helix pair as a function of

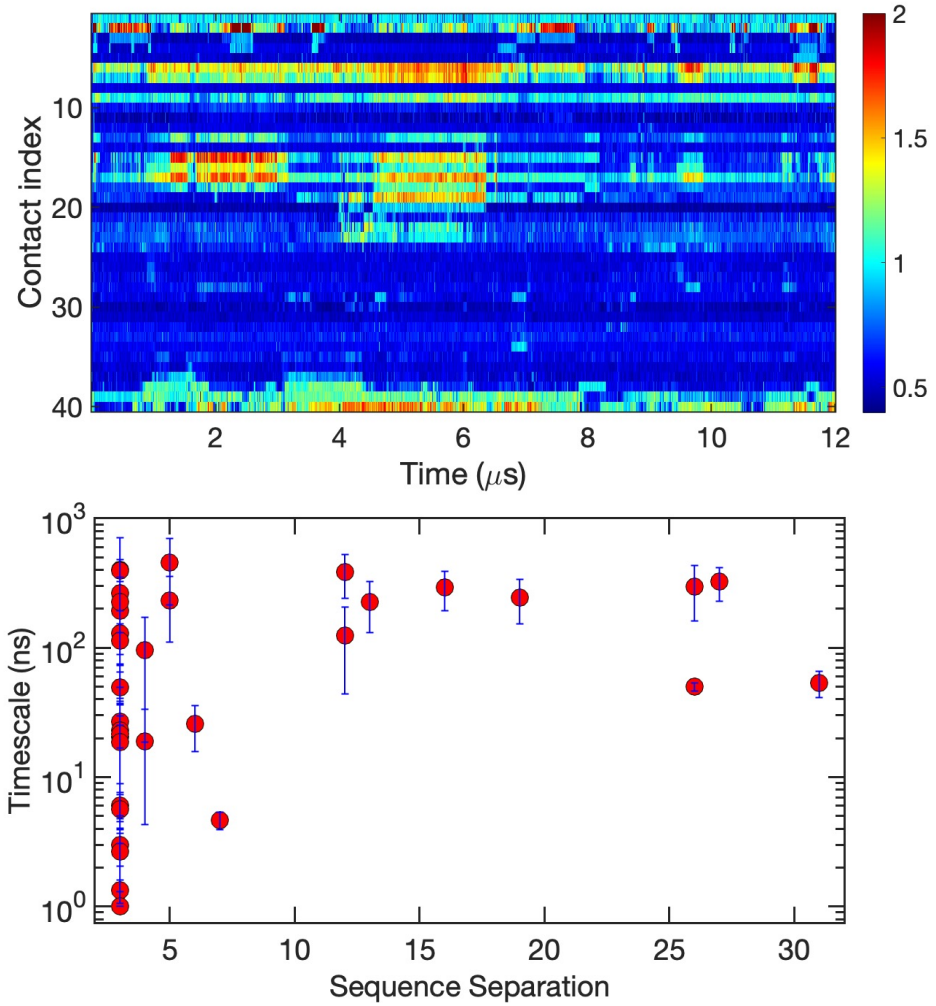

Figure 4.10: Dynamics of intramolecular contacts. Time evolution of the distances between center-of-mass (COM) of C_β contacts extracted from the NMR free form structure across a representative $NCBD_{free}$ trajectory (top). The 40 C_β contact indices are then evaluated on the basis of sequence separation between the interacting residue pairs into short (below 5 residues apart in the sequence), mid (5-10), and long (greater than 10) range. (Right) Timescale (ns) vs. sequence separation of the C_β contacts. Note that we do not include contacts with a fast relaxation decay of less than 1 ns.

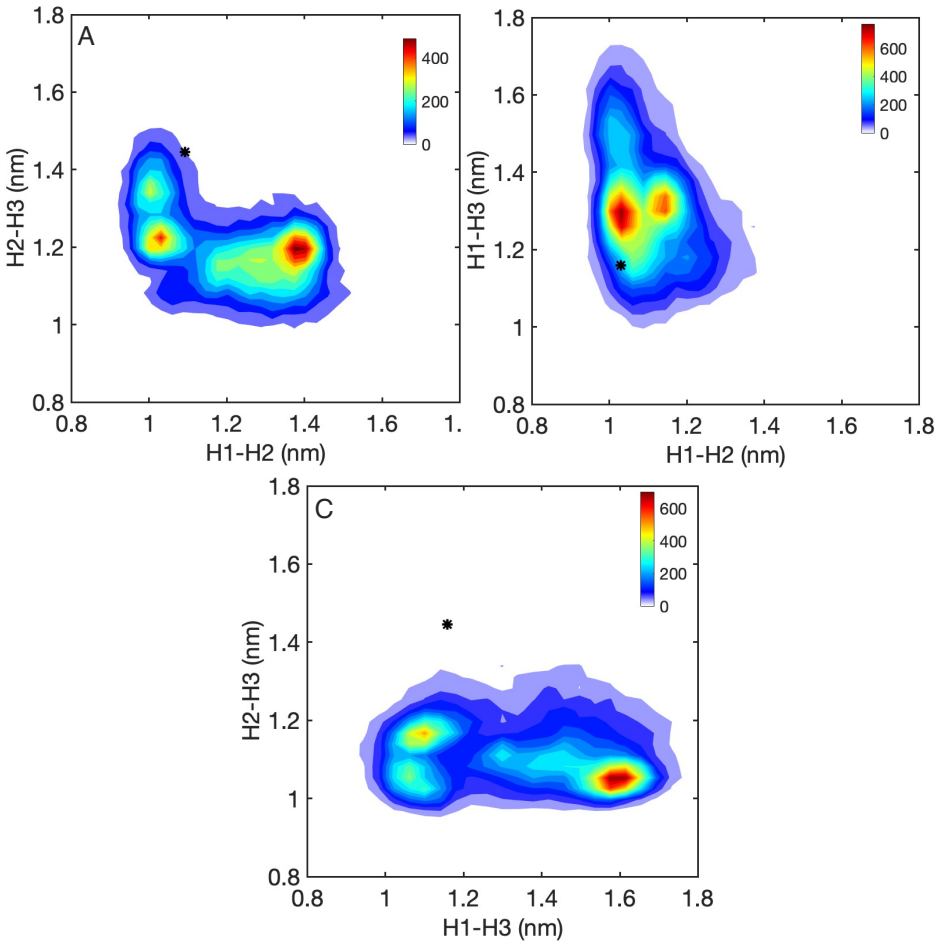

Figure 4.11: Structural rearrangement in $NCBD_{free}$ ensemble. The two-dimensional distributions of distances between the center-of-mass (COM) of all possible helical pairs; (A) H2-H3 vs. H1-H2, (B) H1-H3 vs. H1-H1 and (C) H2-H3 vs. H1-H3. Black marker indicates the respective distances between the COM of all helical pairs in the NCBD conformer bound IRF3 to highlight the relevant structural alterations in $NCBD_{free}$ ensemble.

the simulation length. The time trajectories of the distances between the helix pairs, H1-H2, H2-H3, and H1-H3, of all the trajectories are given in Appendix. We observe global motions within NCBD conformations revealing dynamical features of H1-H3 undergoing multiple reversible transitions concerted with H2-H3 fluctuations in the opposing direction. Figure 4.11 shows the two-dimensional distributions of each of the helix pair vs. the other pair. Figure 4.11.A highlights the interplay between H1-H2 and H2-H3, and we find that the system behaves like an oscillator between H1 and H3 to interact with H2. Figure 4.11.B indicates H1 remains mostly engaged with H2 and H3, sampling conformations that resemble the IRF3 bound helical arrangement. Figure 4.11.C shows a bimodal distribution with H1 and H3 interactions ranging over a broad range compared to the dynamics between H2 and H3, indicating decorrelated behavior among H1-H3 and H2-H3. We observe H2 couplings with either of the end helices, which is consistent with our findings in Chapter 3. NCBD's conformational propensities are enhanced by this dynamical coupling, contributing to the overall marginal folding cooperativity. In addition, this observation relates to the binding interfaces for its structurally diverse partners where the structural elements in NCBD alternate to adopt a conformation suitable for the specific partner. The structural variability in the ensemble is further extracted by the principal component analysis (PCA) of the distances between the COM of all helical pairs and discussed in the following section.

4.4.6 Two-dimensional Projection of NCBD Free Ensemble

We visualize the conformational landscape of free NCBD projected along Q_{free} and the principal component 1 (PC1) of the distances between the COM of helix pairs in the Figure 4.12. Q informs on the overall degree of the native structure with reference to the NMR PDB, and PC1 reports on the structural variability in the dynamics of helical arrangement. PC1 accounts for 65% of the total variance. The resulting conformational landscape is highly broad, with the two most populated minima with $0.71 < \mathrm{Q} < 0.76$ and $0.54 < \mathrm{Q} < 0.58$ respectively, and less

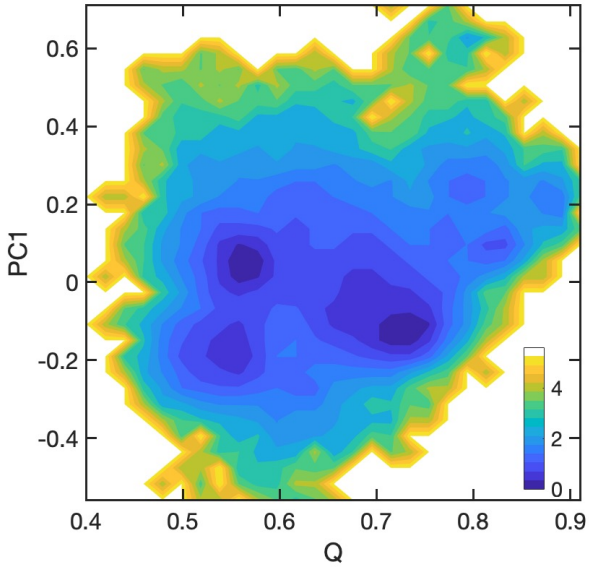

Figure 4.12: Projection of the $NCBD_{free}$ trajectories on the order parameters Q_{free} (fraction of native contacts wrt. NMR structure) and PC1 (principal component 1) of the distances between the COM (center of mass) of all helical pairs, H1-H2, H2-H3 and H1-H3. Color bar is in kcal/mol.

than 1 kcal/mol energy barrier between the two sub-ensembles. Overall, the observed 7 minima demonstrate NCBD's ability to sample multiple sub-states with distinct structural characteristics without crossing a significant energy barrier.

4.4.7 Kinetic model of NCBD

We sought to obtain more direct insight into the morphing dynamics and constructed the kinetic map of the $NCBD_{free}$ conformational ensemble to extract relevant sub-states. As feature vectors to describe the system, we transformed the cartesian coordinate trajectories into the positions of backbone atoms. For dimensionality reduction, we conducted a linear transformation on these feature vectors using time-lagged independent component analysis (TICA) [147] to find a projection containing the slowest kinetic modes by maximizing the autocorrelation function in the reduced space. Note when building Markovian matrices and estimating kinetics, the TICA space is always more reliable than an original space because it aids in separating correlated signals from noise in the original

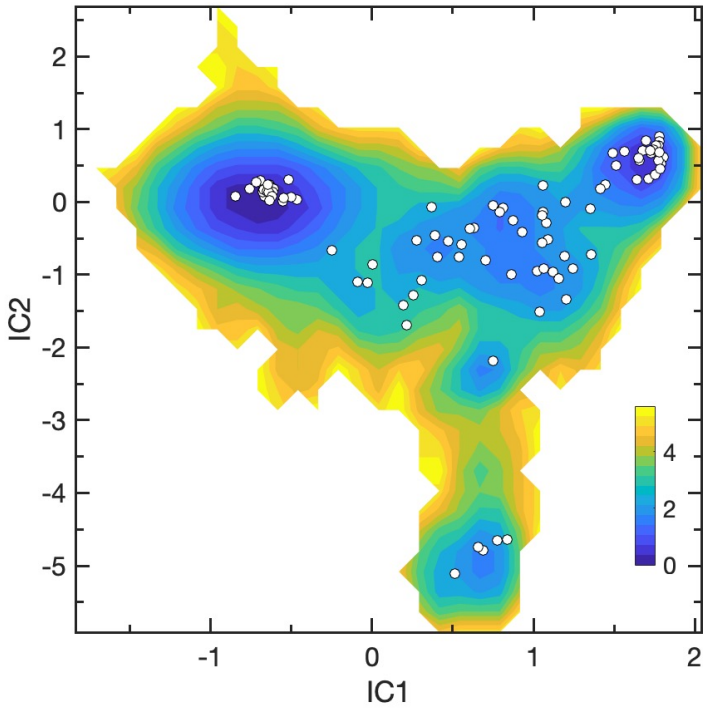

Figure 4.13: Elucidation of relevant $NCBD_{free}$ states. 100 microstates (white) plotted onto the free energy profile of the data transformed by time-lagged independent component analysis (TICA) with a lag time = 15 ns and further computed over the first two independent components (IC). Color bar is in kcal/mol.

data by optimizing correlations among all variables.

Next, we performed k-means clustering to decompose the low-dimensional MD data into hundreds of relevant discrete microstates such that each frame of the trajectories can be assigned to one of these microstates. The clusters are plotted onto the free energy profile projected along the two slowest TICA reaction coordinates, as shown in Figure 4.13. Our analysis reveals that the kinetic states are well distributed in the low-dimensional TICA subspace, and most clusters are located around the local free-energy minima. The overall essence of the obtained profile is consistent with our observation in Figure 4.12 that captures multiple sub-states. Note that the kinetic information is lost in this projection (Figure 4.12), unlike the TICA subspace.

We estimated the Markov state model (MSM) from the discretized trajectories to extract the kinetically relevant states and the interconversion timescales

between them. Based on the relaxation timescales,, we reduce the high-resolution MSM into a more tractable representation by the hidden Markov model (HMM) algorithm into seven macrostates (details in Methods). The Figure 4.14 illustrates the kinetic model of the NCBD conformational ensemble with the average transition timescales between the states and their representative structures. To examine how each macrostate of NCBD relates to the bound conformers, we computed the mean RMSD and associated uncertainty of various conformations belonging to each state with reference to the free form NMR, ACTR, p53-TAD and IRF3 bound structures and are enlisted and color-coded according to Figure 4.1.

We find several relevant states with significant transitions and lifetimes revealing crucial insights into NCBD morphing behavior. State 1 makes up 2.6 % of the total ensemble used to construct HMM and is topologically closest to the free form, ACTR, and p53-TAD known bound structures, and transitions to either state 2 or 4 with average timescales of 1.9 μs and state 6 in sub-microsecond. The structural features of state 2 indicate a more compact topology that closely resembles state 1 with marked variation in helix 1. State 2 rapidly transitions to state 6 within $\approx$ 400 ns. The lowly populated state 3 is structurally closest to all the known conformers and captures the orientation of helix 1 in the process of transforming to the IRF3 bound structure.

Interestingly, state 6, with 18.8 % of the total sampling, represents the IRF3 associated topology as it has the lowest mean RMSD value of 0.3 nm and varies significantly compared to the other structures. This observation indicates the rheostatic capability of NCBD to gradually morph onto conformations with significant topological variations that resemble the bound conformations of NCBD. Multiple computational studies on NCBD have also previously reported observing a cluster corresponding to the IRF3 conformation [30],[63]. In fact, state 6 serves as the connection hub that most states transition to and therefore has an extended lifetime. The most disordered conformations are found in states 4, 5, and 7. State 5 quickly relaxes to state 7 in 9.1 ns, while state 4 accounts for more

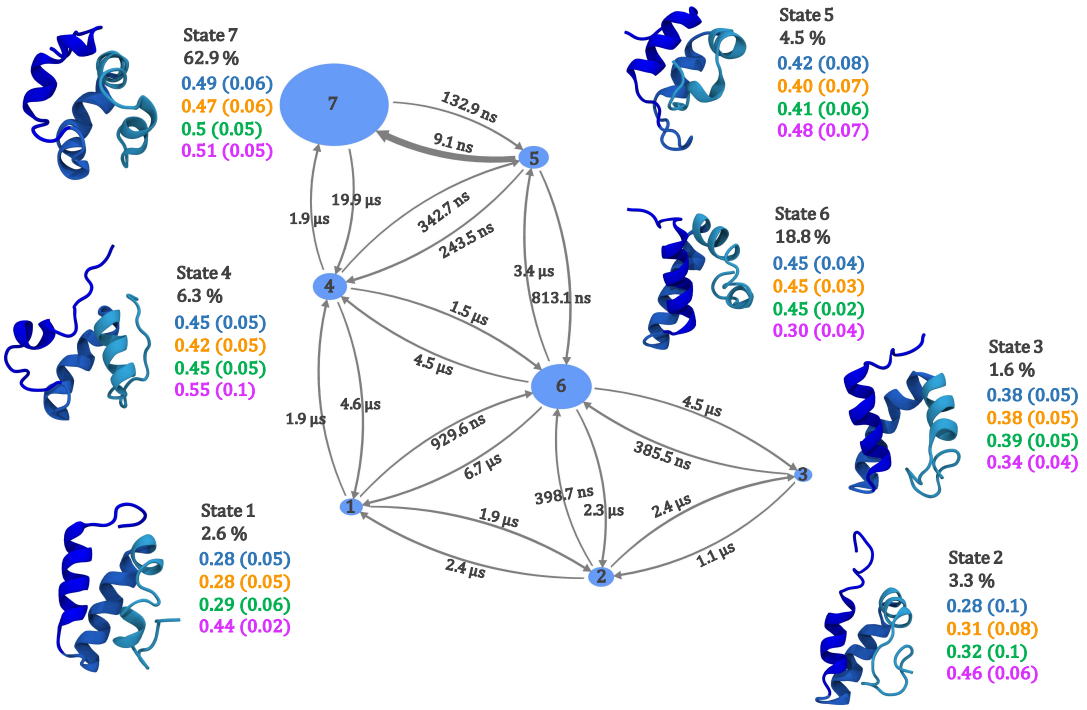

Figure 4.14: Kinetic network of $NCBD_{free}$. Kinetically metastable conformations (macrostates) obtained from kinetically coupled microstates via Hidden Markov Model (HMM) analysis. The relative population of each macrostate is proportional to the volume of each representative sphere and interconversion kinetics are shown with thickness of the connections proportional to average transition time between two macrostates. The minimum average RMSD of each experimentally determined NCBD structure versus 10,000 randomly selected macrostate conformations is stated (after superposition of all backbone atoms of residues 6 to 47). RMSD values and related uncertainties are color-coded in the following order: NCBD free form (blue) and bound to ACTR (yellow), p53-TAD (green), and IRF3 (pink).

extended conformations with largely unstructured helix 1. Finally, macrostate 7, which is linked to states 4 and 5, is the largest populated state and the model estimates the average transition timescale of ≈20 μs from 7 to state 4 that is not observed in the original trajectories which are 12 microseconds long. The structural features of state 7 show an average topology of NCBD compared to all four reference structures with RMSD ≈ 0.5 nm.

This observation highlights that the gradual morphing phenomenon that we observe here in NCBD largely samples conformations that represent an average structural arrangement of all the three α-helices that are partially structured with closely interacting regions. While our comprehensive analysis shows that NCBD samples conformations that are compatible to binding with its structurally diverse ligand partners, it mostly remains in a partially disordered state, suggesting that NCBD is readily available to morph into a pre-binding competent structure when in the presence of a partner, without being structurally biased towards a more ordered conformation. Overall, these mechanistic insights provide a fundamental basis of the functional importance of NCBD and demonstrate a working conformational rheostat in transcription.

4.4.8 Conformational Propensities of NCBD bound ensembles

Investigation of the (un)folding dynamics of NCBD in the absence of their respective partner (see Methods) allows us to determine how fast the ensemble re-equilibrates between the different sub-ensembles that are typically selected by partner binding. The fast conformational relaxation (Figure 4.6) to a broad ensemble relative to the free form shown in Figure 4.5 confirms the stochastic variability of the simulations. Monitoring the time trajectories as a function of Q with reference to NMR free form PDB show strikingly marked conformational differences (Figure 4.15). ACTR trajectory initiates with a high Q ≈ 0.8 since the bound conformer is structurally similar to the NMR free form with an extended H3 and then undergoes large structural changes with a grey swath representing the wide range of the free form ensemble. P53-TAD bound conformer is a bit

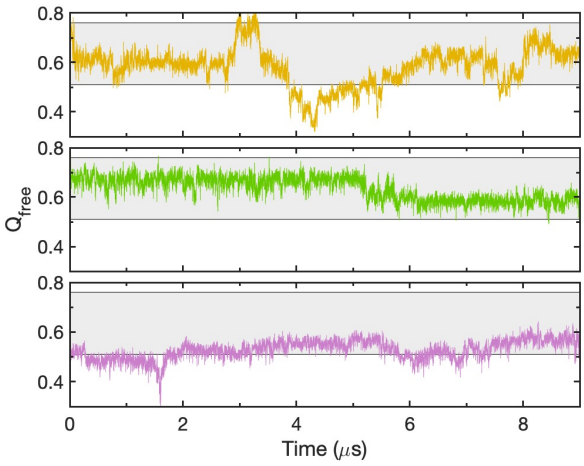

Figure 4.15: Time evolution of fraction of native contacts wrt. free form NMR structure across three distinct NCBD MD ensembles in the absence of ligand partners; ACTR (yellow), p53-TAD (green) and IRF3 (pink).

open but almost identical to the NMR PDB (Figure 4.1), and the trajectory converges to the free form ensemble. The deviation of IRF3 bound conformer from the free form structure results in a relatively low Q in the initial phase of the trajectory and experiences structural alterations towards the free form ensemble. For further analyses, we use the last 4 μs of the trajectories from Figure 4.15.

Inspection of the characteristic timescale as a function of sequence separation for each ensemble (Figure 4.16) shows similar dynamical behavior relative to free form (Figure 4.10). To further dissect their structural characteristics, Figure 4.17 shows the proportion of time spent by each residue of NCBD in a helical state; the helices in each of the respective starting structures are also indicated. This analysis follows the same definition as in Chapter 3 for Molecular LEGO. The helical regions are consistent with free form ensemble and display significant helicity in H1 and H3, and greater variability in H2 and C-terminal tail. The conformations sampled in each trajectory differ significantly from their starting bound form and closely resembles the free form native structural ensemble in the microsecond timescale. In general, the native NCBD ensembles are in excellent agreement with the CD experiments discussed in Chapter 3 in the context of average helical content $\approx$ 50 %. Taken together, these results highlight the large inherent

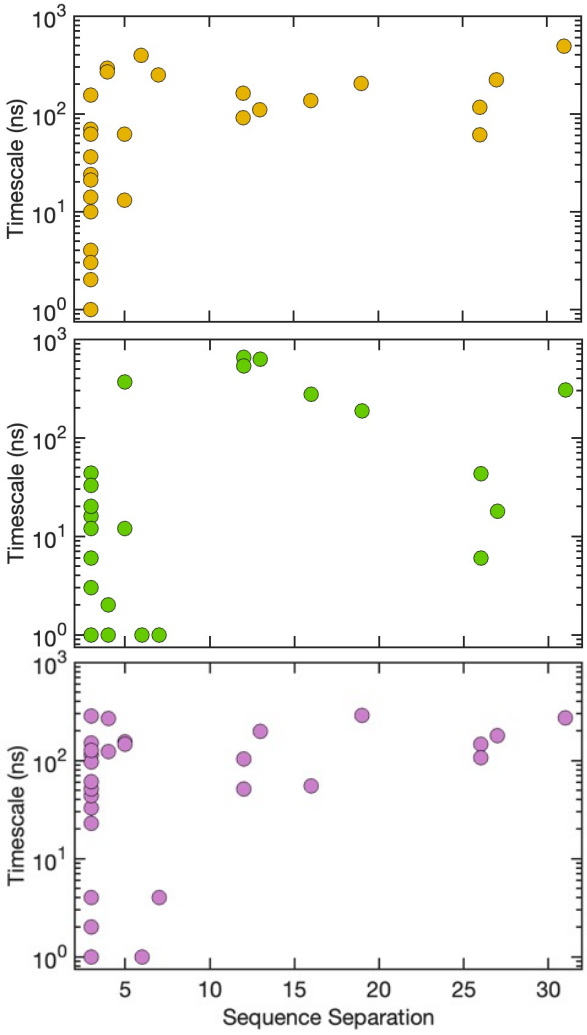

Figure 4.16: Dynamics of intramolecular contacts. Timescale (ns) vs. sequence separation of C_β contacts following the same definition in Figure 4.10 across three distinct NCBD MD ensembles in the absence of ligand partners; ACTR (yellow), p53-TAD (green) and IRF3 (pink).

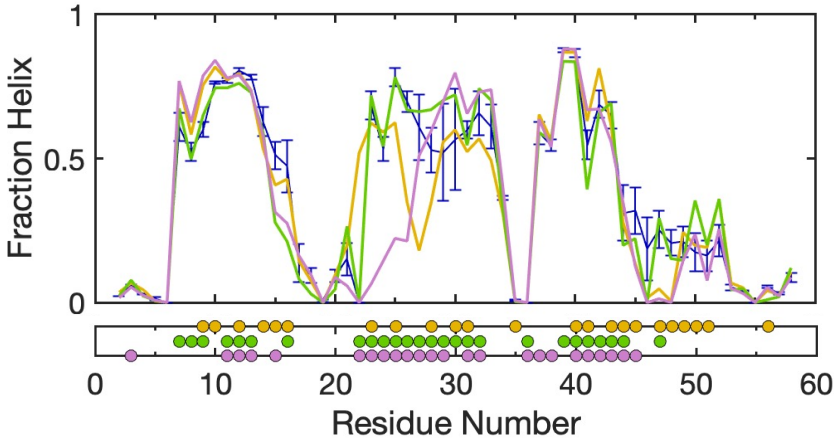

Figure 4.17: NCBD helical propensities. Fraction helix per residue across three distinct NCBD MD ensembles in the absence of ligand partners; ACTR (yellow), p53-TAD (green) and IRF3 (pink). $NCBD_{free}$ profile is in blue for reference. The helical component in each of the respective starting structures shown below in filled circles. The error bars indicate the standard error of three $NCBD_{free}$ trajectories.

conformational fluctuations of NCBD. The presence of diverse functional conformations in its native ensemble strongly supports the conformational rheostat behavior that potentially enables its promiscuous binding to multiple partners.

4.5 Discussion

NCBD is a well-studied IDP that binds to structurally diverse ligand partners. We performed extensive μs-long MD simulations of multiple experimentally determined NCBD structures (free and bound forms) and found that NCBD exhibits gradual (un)folding behavior in microseconds timescale sampling bound-specific conformations, essentially acting as a conformational rheostat. The simulations provide a detailed picture of the NCBD morphing phenomenon, illustrating how a network of high local and transient non-local interactions results in a more rigid conformational ensemble at fast timescales and a highly dynamic ensemble at longer timescales and can lead to distinct functionally relevant structural characteristics. The simulated NCBD ensembles show a high probability to pop-

ulate a topology that is an average form of all the known bound conformations, whether associated with other IDPs or structured protein. This important finding provides a mechanistic explanation into NCBD morphing coupled to binding behavior in which NCBD samples structural features in a partially disordered state to remain readily available and potentially morph onto distinct conformations compatible with its structurally diverse biological partners. The somewhat average structure of NCBD supports the diffusion-limited mechanism of binding to partners [153].

Although the force field description and sampling limit these MD ensembles, we find our data is in excellent agreement with the biophysical characterization using CD and subsequent ensemble-based analysis of NCBD as discussed in Chapter 3. In addition, simulations started with different conformations in the absence of partners re-equilibrate to the free form ensemble within microseconds confirming the reliable sampling of the simulations. We note that NCBD populates a highly heterogeneous ensemble but retains native-like features compared to the NMR free form. Despite the dynamic nature of the domain, it is being shown here that the NCBD ensemble contains a significant fraction of a conformer similar to NCBD in the NCBD:IRF3 complex, suggesting that the ligand is able to select a pre-folded conformer from the ensemble. Although in lower proportions, it also samples conformers consistent to ACTR and p53-TAD bound complex. These analyses support a conformational mechanism to select the partner and agree retrospectively with the previous reports on NCBD. Furthermore, NCBD's distinct conformational dynamics is remarkably similar to that of a downhill folding module. Overall, this study provides crucial insights into how a conformational rheostat exerts a key functional role in transcription and potentially extends to other IPDPs at the hubs of cellular networks linked to different regulatory and signaling functions associated with various human diseases.

CHAPTER 5

Dissecting the Interplay between NCBD and its Binding Partners

Chapter 4 illustrated the underlying structural and dynamical features of NCBD conformational ensembles in the absence of its ligand partners. We find that the short-lived NCBD sub-states can morph to functionally relevant structures with distinct topological variations, which resemble the bound conformations of NCBD to other morphing proteins and a well-structured ligand partner. These observations adequately provide high-resolution insights into a relatively unexplored conformational rheostatic mechanism governing unbound NCBD and are also consistent with the established results on NCBD. However, to determine the descriptive states of its complex binding modes, we should consider the influence of its various structurally diverse ligand partners in the bound states. To embrace this task, in this Chapter, we first study NCBD's IDP partners, p53-TAD and ACTR, and decipher their structural properties in the unbound forms. Next, we elucidate the interplay of NCBD bound to the IDP partners and the structured protein, IRF3, using all-atom MD simulations.

5.1 Background

This section provides detailed background on NCBD coupled folding and binding to ACTR, p53-TAD and IRF3 studies using extensive experimental and computational biophysical techniques (experimentally determined structures are shown in Figure 5.1).

Despite their small size, the folding and binding kinetics of ACTR and NCBD

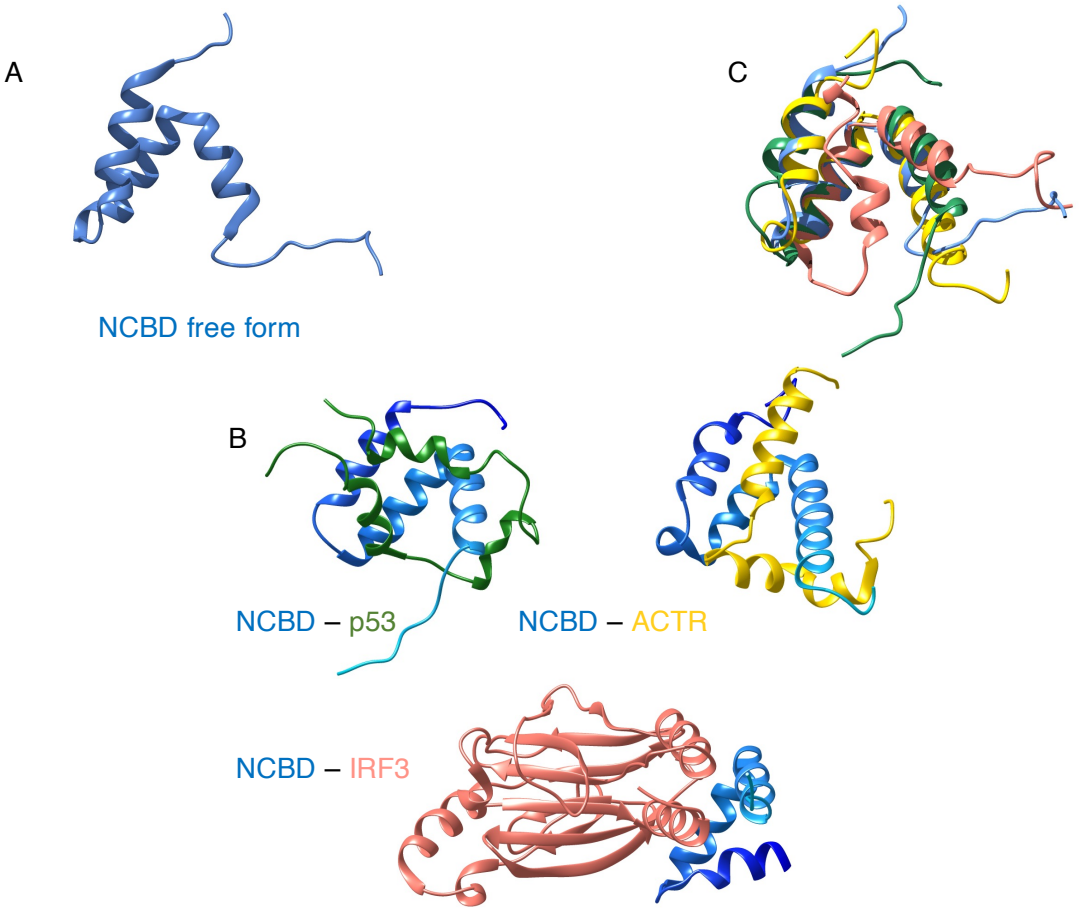

Figure 5.1: Experimentally determined A. NCBD free form structure (PDB ID: 2KKJ). B. NCBD bound to ligand partners: p53-TAD, ACTR, and IRF3 shown in cartoon representation. Dark to light blue represents N- to C-terminal. C. Structural superimposition of NCBD free form with bound conformers color–coded according to its respective ligand partner to illustrate its structural plasticity.

have been observed to be complex. Free NCBD has been previously described as a molten globule ensemble with similarities to the structure found in complex with ACTR, which led the authors to propose a conformational selection mechanism for its binding [62]. On the other hand, studies using topology-based models, implicit solvent simulations, and replica-exchange sampling of NCBD and ACTR reported that there is synergistic coupled folding and binding of two IDPs leading to mixed induced fit and conformational selection mechanisms and a prevalent role for electrostatic interactions in their specific recognition [115],[64]. Further, using topology-based model simulations (calibrated to balance intermolecular interactions) of NCBD:ACTR complex, authors suggested that even though binding and folding of NCBD and ACTR is cooperative on a baseline level, the tertiary folding of NCBD is best described by the 'extended conformational selection' model that involves multiple staged of selection and induced folding. Further, these simulations predicted that mainly second and third helices of NCBD and ACTR are involved in the recognition to form a mini folded core [154].

Recent sm-FRET studies to capture the transition path times for the association of NCBD-ACTR complex showed that both IDPs form a folded dimer upon binding. The results revealed an electrostatically driven metastable encounter complex with a lifetime of 80 μs that transits to the final bound state [155]. This observation is similar to a fast-kinetic phase observed during NCBD:ACTR binding in a different study using temperature-jump experiments [156], which was attributed to the conformational exchange within NCBD as observed by NMR [62]. Thus, it has been suggested that the lifetime of the transient encounter complex might be associated with the internal dynamics of NCBD (molten-globule-like). Further, a recent study using sm-FRET showed NCBD undergoes slow conformational switching (tens of seconds) between two subpopulations that differ by the conformation of a proline residue where NCBD pro20 in trans binds ACTR with higher affinity than in cis conformation (by order of magnitude). The difference in affinity is primarily caused by a change in the dissociation rate coefficient. Further MD simulations indicate reduced packing of the complex for the cis isomer, overall suggesting peptidyl-prolyl cis/trans isomerization may be an important

mechanism for regulating IDP interactions [116].

A study of different mutants of ACTR (without interfering with the intermolecular interactions between ACTR and NCBD) using NMR and fluorescence-monitored stopped-flow kinetic measurements show that the secondary structure content in helix 1 of ACTR influences the binding kinetics. These results demonstrate that helical propensity in ACTR modulates its binding to NCBD, both in terms of association and dissociation [157]. Similarly, in a computational study, topology-based modeling, and simulations have shown increasing ACTR helicity enhances the binding rate and that residual helices mainly promote more efficient folding upon binding [158].

Moreover, using a multi-state coarse-grained simulation model, it has been shown that the binding of NCBD to either binding partner ACTR or IRF3 appears to occur via an induced-fit mechanism [159].

A total of 800 ns of explicit solvent MD simulations of p53TAD:NCBD (at room and high temperatures) suggested a local induced fit and a global conformational selection in recognition mechanism between the two IDPs [160]. Unbound TAD is mainly disordered and undergoes a large conformational change upon binding to NCBD. Kim et al. using sm-FRET, investigated the molecular mechanism of binding of p53TAD and NCBD. The analysis of photon trajectories shows that the lifetime of transient complex (TC) of TAD and NCBD binding is much longer than that of the association of two folded proteins. The long lifetime ($\approx 183\ \mu$s) results from the stabilization of TC by non-native electrostatic interactions, which makes diffusion-limited association possible [65].

Overall, the studies on binding promiscuity of NCBD to its structurally diverse ligand partners have shed light on its coupled conformational change and ligand binding in context to both conformational selection and induced fit mechanisms. Our findings in the previous Chapter corroborate such a mixed mechanism as we demonstrated a conformational rheostat mechanism regulating free NCBD conformational ensemble populating pre-competent binding structures that suggests a conformational-selection type aspect, which should then relax to complete

the binding, suggesting induced-fit-like features.

However, the structural properties of NCBD bound complexes in terms of their conformational dynamics after undergoing coupled folding and binding remain elusive, especially how NCBD-IDP bound complexes (associated with ACTR or p53TAD) compare to NCBD bound to a structured protein (IRF3). Furthermore, since our current findings on NCBD revealed its rheostatic capabilities, deeper investigation into NCBD bound complexes and its IDP partners are needed to fill in the gaps in our understanding of how a conformational rheostat functions during transcription.

The binding between two IDPs is more challenging to investigate computationally due to large entropic contributions, especially if each protein spends significant time in conformations that are incompatible with binding to the partner. To this end, here, we performed explicit solvent long-scale all-atom MD simulations to obtain conformational sampling of NCBD in association with each of its three ligand partners in the vicinity of their respective essential subspace (Figure 5.1). Our results shed light on the interplay between NCBD and its structurally diverse ligand partners.

5.2 Methods

All-atom MD simulations. We carried MD simulations in explicit solvent using the GROMACS package [118],[119], and the Charmm22* force field [52]. Water molecules were described using the TIP3P model. Periodic boundary conditions were used, and long-range electrostatic interactions were treated with the Particle Mesh Ewald (PME) [120] summation using a grid spacing of 0.16 nm combined with a fourth-order cubic interpolation to derive the potential and forces in-between grid points. The real space cutoff distance was set to 1.2 nm, and the van der Waals cutoff to 1.2 nm. The bond lengths were fixed [121], and a time step of 2 fs was used for the numerical integration of the equations of motion. Coordinates were recorded every 10 ps.

System (PDB ID)	Size (no. of atoms)	No. of simulations	Time (μs)
ACTR (1KBH - chain A)	18,920	2	14
P53-TAD (2L14 - chain B	23,719	2	14
NCBD:ACTR (1KBH)	19,327	2	14
NCBD:p53-TAD (2L14)	28,954	2	14
NCBD:IRF3 (1ZOQ)	51,673	1	10

Table 5.1: Simulation details of NCBD partners and the complexes.

Simulation details of all systems are enlisted in Table 5.2. For unbound ACTR, NCBD:ACTR and NCBD:p53-TAD, simulations were started from the lowest energy structure of the NCBD NMR ensemble (see Figure 5.1). For P53-TAD system, the protein was modeled using i-Tasser protein structural modeling tool [141] to account for the missing 1-13 residues in PDB ID: 2L14 (chain B). The homology modeled structure was first subjected to energy minimization using UCSF Chimera and then prepared for the simulation. The X-ray crystallography characterized NCBD:IRF3 complex structure was used as the starting structure for the 10 μs long simulation.

All systems were placed in a dodecahedral water box large enough to contain the protein and at least a 1.5 nm layer of solvent on all sides. The structure was solvated with water molecules, and ions were added to neutralize the system. The unbound ACTR and P53-TAD were acetylated and amidated. The CHARMM22* force field was then adjusted to include the parameters for N-acetylation and C-amidation. Box dimensions were kept sufficiently large to account for the flexibility of partially disordered proteins. In all the simulations, the protein became compact and never interacted with its mirror image in the periodic boundary conditions. In all cases, the starting structure was subjected to energy minimization using the steepest descent method. All systems were equilibrated at a constant temperature of 310 K utilizing the two-step ensemble procedure (NVT and NPT). First, the system was subjected to NVT (constant number of particles, volume, and temperature) equilibration for 500 ps with the position of the protein restrained, followed by NPT (constant number of par-

ticles, pressure, and temperature) equilibration for 4 ns each. The simulations were subjected to the modified Berendsen thermostat with a 0.1 ps relaxation time [122] to maintain the temperature. The structures were then subjected to Parrinello-Rahman with 0.2 ps relaxation time for pressure coupling [123] at 1 bar before the production run was started at 310 K. All the simulations were run on the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputing center (SDSC)

Conformational landscapes of NCBD Complexes. The map was obtained from the normalized probability distribution as a function of the relevant set of order parameters. The probability distribution was converted into an energy scale using the following expression:

$$\Delta A_{ref \to i} = -RTln(\frac{P_i}{P_{ref}}) \tag{5.1}$$

where the probability of going from a reference state (ref) of the system to any state i (e.g., from folded to unfolded) at constant temperature and constant volume is evaluated. R is the ideal gas constant, T is the temperature, and p_i and p_{ref} are the probabilities of finding the state i and state ref system, respectively. We project the conformational space onto two order parameters: Fraction of native intermolecular contacts (Q_{inter}) and the fraction of intramolecular contacts (Q_{intra}). In Q_{inter} calculation, a contact is considered formed when the minimum pairwise distance between heavy atoms of the interacting residues across the two proteins is <=0.5 nm. In Q_{intra}, a contact is considered formed when the minimum pairwise distance between heavy atoms of the interacting residues is <=0.5 nm and the residue pair is >3 residues apart in the protein sequence. Conformations collected at 100 ps intervals were projected onto the Q_{intra}-Q_{inter} plane using a 32 × 32 grid (1024 cells) and sampling statistics were compiled to evaluate Equation 5.1. The grid cell with the largest population was used as reference state.

Time-averaged probability contact maps. We determined the time-averaged probability of finding each contact (whether native or non-native) in

NCBD, ACTR and p53-TAD between two residues that are at least 3 apart in the sequence. For each time frame, a contact was considered formed when at least one heavy atom of the first residue was within 0.5 nm distance of at least one heavy atom of the second residue with a threshold of at least 70% of 10 ns, the two interacting residues remained in contact.

Inter-protein interactions in NCBD and ligand partner. Based on the above mentioned threshold, we computed the time-averaged probability of finding each contact (whether native or non-native) between each NCBD: ligand partner bound trajectories. A contact was considered formed when at least one heavy atom of one protein's residue was within 0.5 nm distance of at least one heavy atom of the other protein's residue.

Hydrogen bond analysis. The number of intermolecular hydrogen bonds were computed using gmx hbond tool with donor (OH and NH)-acceptor (O) cut-off of 0.35 nm between the two interacting residues across three NCBD complex trajectories (bound to ligand partners; ACTR, p53-TAD, and IRF3).

5.3 Results

We performed $\approx$ 70 μs aggregate explicit solvent all-atom MD simulation of NCBD's IDP partner proteins, ACTR, and p53-TAD in their unbound forms and in complex with NCBD. In addition, we also performed NCBD bound to the structured protein, IRF3, to characterize the conformational and dynamic differences between these bound ensembles. Our comprehensive investigation of NCBD in Chapters 3 and 4 provides an excellent platform to compare the dynamical properties of its morphing ligand partners when in complex with NCBD.

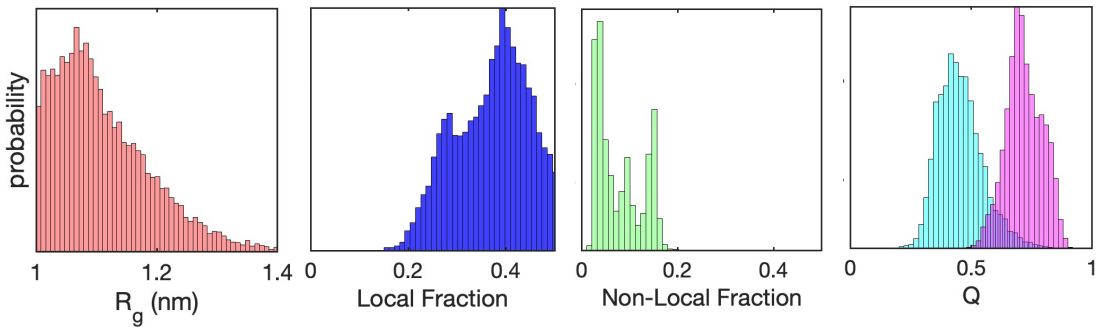

Figure 5.2: Structural properties of ACTR. Probability distributions from one-dimensional projections as a function of the (left to right) radius of gyration, local and non-local native fraction in the absence of NCBD, and the fraction of native contacts in the absence (cyan) and presence (magenta) of NCBD.

5.3.1 Coupled Folding and Binding of Morphing Proteins

5.3.1.1 Structural properties of ACTR in the absence and presence of NCBD

Figure 5.2 shows the probability distributions of ACTR in the absence and presence of NCBD as a function of multiple structural properties. The unbound ACTR undergoes significant structural transitions that populate relatively extended conformations. This observation differs from what we observed for the isolated NCBD MD ensemble. The dynamical content of the local fraction results in transient partly disordered states, even though the fraction of tertiary contacts is almost negligible. The MD ensemble lacks adequate non-local interactions and agrees well with unbound ACTR NMR experiments [139].

On the other hand, the ACTR conformational ensemble in the presence of NCBD shows a higher fraction of native contacts (mean Q $\approx$0.75) with reference to the NMR determined ACTR bound structure, compared to the unbound ACTR ensemble (mean Q $\approx$ 0.48). Interestingly, both ensembles display a unimodal distribution as a function of Q, implying that ACTR dynamics are characterized by gradual morphing behavior.

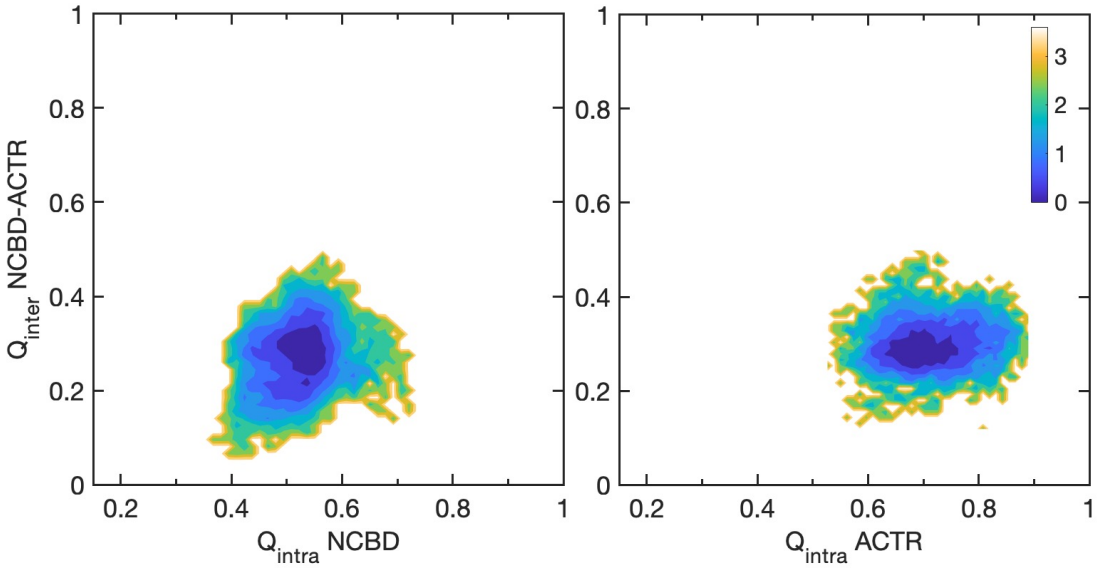

Figure 5.3: Projection of the NCBD:ACTR complex trajectories on the order parameters Q_{intra} (fraction of native contacts wrt. NMR structure) of NCBD (left) and ACTR (right), and Q_{inter} (fraction of native intermolecular contacts between NCBD and ACTR). Color bar is in kcal/mol.

5.3.1.2 Conformational propensities of NCBD:ACTR ensemble

We examined the bound-state ensemble of the NCBD:ACTR complex to understand the extent of the coupling among intermolecular interactions and the structural order of the two associated IDPs. Here the rationale is to determine the conformational dynamics that ensue after the association event. Our MD trajectories initiated from the bound NMR structure would provide detailed information in the vicinity of its essential sub-space. To this end, the Figure 5.3 shows the 2D projection along the fraction of native intermolecular contacts (Q_{inter}) and the fraction of native intra-molecular contacts (Q_{intra}) for both the morphing proteins.

We find that NCBD remains party disordered, sampling heterogeneous conformations, and ACTR populates a more structured ensemble, as discussed above. Despite the entangled binding pose of the two IDPs (Figure 5.1.B), nearly 30 % of the native bound fraction is observed, indicating that they have lost most of their specific intermolecular contacts. Additionally, even in the bound-state

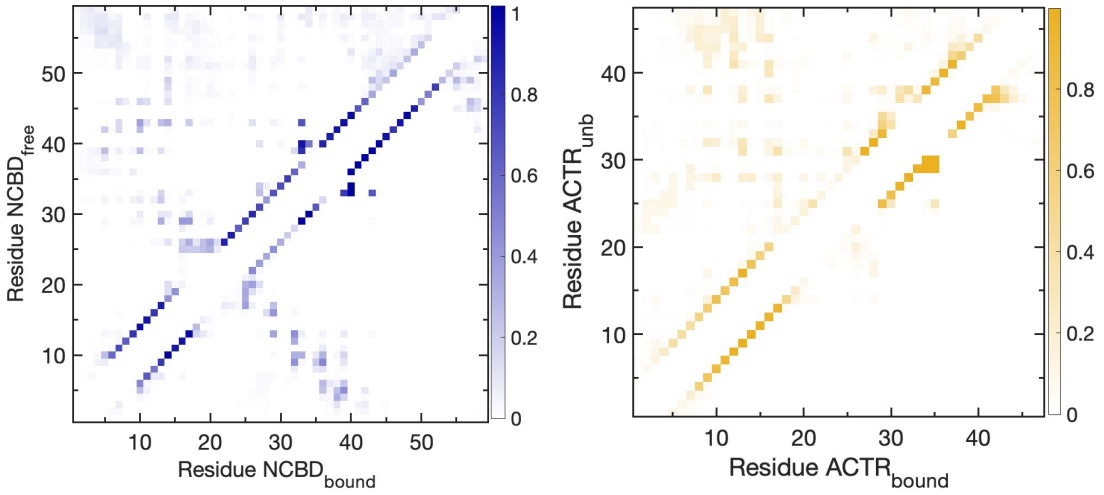

Figure 5.4: Total interaction map of NCBD and ACTR. Residue-residue total contacts (native and non-native) observed in the simulations of NCBD (left) and ACTR (right) bound (lower triangle) and free/unbound forms (upper triangle) for comparison. The color intensity reflects the time-averaged probability of observing the contact, with the light to dark color intensity corresponding to weakly to strongly interacting residues respectively.

ensemble, we find a single broad free energy minimum highlighting their gradual conformational transitions.

To visualize the structural feature, we analyzed the probability contact maps of the two IDPs in the absence and presence of each other depicted in Figure 5.4. H2 (residue 23-37) in NCBD remains largely disordered compared to the free form ensemble. Its interaction with ACTR results in highly non-specific inter-protein contacts not found in the NMR determined structure depicted in Figure 5.5. Furthermore, NCBD's conformational variability in combination with ACTR is highlighted by the weak non-local interactions between H1-H2 (Figure 5.4 left). The bound trajectories capture the high probability of Helix 3 extension on to C-terminal tail (residue 45-55). This characteristic feature is consistent with our findings in Chapters 3 and 4, which show that in the absence of helix 2 interactions with helix 3, Helix 3 propagation is favorable. The intricate binding pattern between the two IDPs weakens the helix 2-3 interactions, allowing considerable structural rearrangement in NCBD.

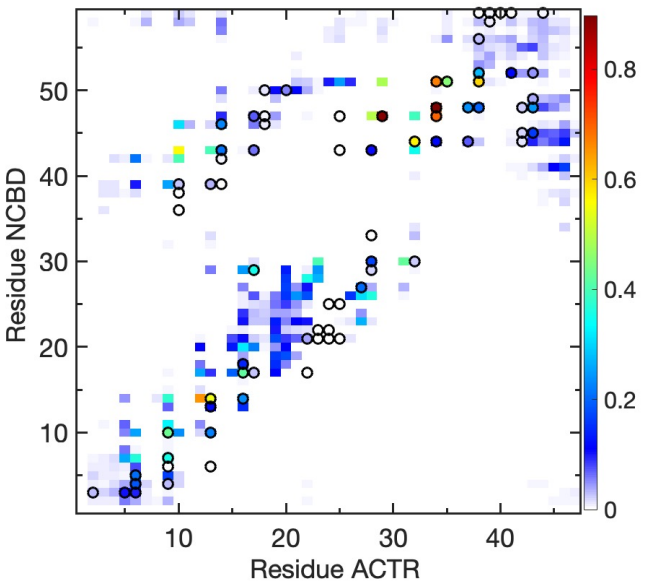

Figure 5.5: Inter-protein interactions in NCBD:ACTR complex. Residue-residue total contacts (native and non-native) observed in the simulations of NCBD bound to ACTR. Black circles indicate native contacts derived from the NMR structure. The color intensity reflects the time-averaged probability of observing the contact, with the blue to red color corresponding to weakly to strongly interacting residues respectively. A contact is considered formed based on the definition in Figure 5.4

.

Although the helix 3 of NCBD causes an apparent kink in the helix 2 of ACTR, ACTR shows a high probability of local native contacts across both helices (Figure 5.4 right). Unbound ACTR, on the other hand, stays relatively disordered in that region, sampling non-helical contacts. There are no long-range interactions found in the bound ACTR, implying that the overall enhanced stability of local interactions is due to NCBD associativity, which is further evident in Figure 5.3. The inter-protein probability contact map shows the occurrence of the native-specific and non-specific interfacial contacts in the complex trajectories. This analysis reveals the dynamical binding of the two IDPs as they retain significant native interfacial contacts for most of the simulation time that restricts their binding pose to an extent but at the same time sample numerous non-specific contacts (non-circle) highlighting their inherent flexibility. The native salt bridge between R47 of NCBD and D29 of ACTR is reported to contribute to the binding specificity of the NCBD:ACTR complex [17],[66] and is found to be strongly connected in the obtained MD ensemble ($> 95\%$ of the simulation time).

5.3.1.3 Structural characteristics of p53-TAD in the absence and presence of NCBD

Next, we focused on p53-TAD (1-61 residue) to characterize its structural properties in isolation and the presence of NCBD. Despite having similar protein lengths, unbound p53-TAD samples significantly more extended conformations than ACTR and NCBD, as depicted in Figure 5.6, where R_g ranges from 1.15 to 1.4 nm, and greatly disordered conformations, as evidenced by high variability in local contacts and a negligible non-local fraction. Surprisingly, there were no large-scale differences between the conformational ensembles of the unbound and bound states, as shown along Q. Unlike ACTR in combination with NCBD, bound p53-TAD (magenta) exhibits largely unstructured conformational fluctuations. Note that the unbound protein spans 1-61 residue, whereas the bound protein spans 13-61 based on the NMR complex.

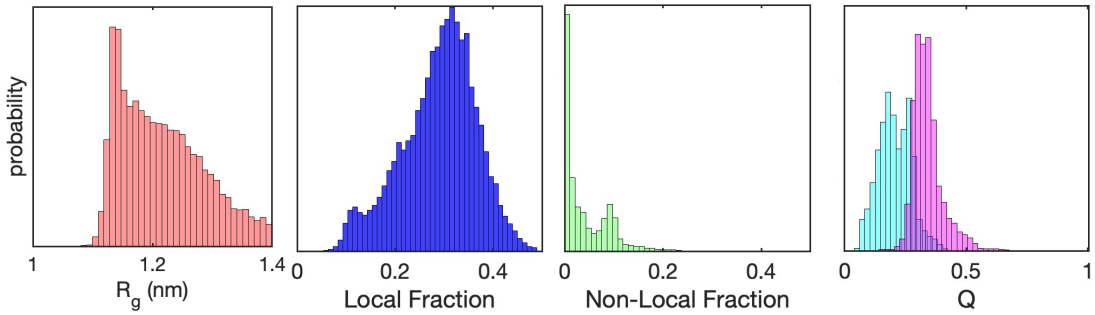

Figure 5.6: Structural properties of p53-TAD. Probability distributions from one-dimensional projections as a function of the (left to right) radius of gyration, local and non-local native fraction in the absence of NCBD, and the fraction of native contacts (Q) in the absence (cyan) and presence (magenta) of NCBD.

5.3.1.4 Conformational propensities of NCBD: p53-TAD ensemble

To further gain insights into the origin of this structural heterogeneity of the bound p53-TAD, in Figure 5.7, we project the NCBD:p53-TAD complex trajectories as a function of Q_{intra} and Q_{inter} for both IDP proteins. This protein-ligand complex manifests a strikingly opposite conformational behavior from what we observed for ACTR. With reference to the NMR determined structure of the complex, NCBD samples a substantially balanced ensemble, as shown by the higher fraction of intramolecular contacts with 77 % of native-ness intact.

Despite being poised in an entangled binding pattern (Figure 5.1.B), only a small fraction of native binding contacts between the two proteins are observed, with Q_{inter} varying from 0.4 to almost 0, and having minimal impact on NCBD's conformational properties. p53-TAD, on the other hand, samples largely disordered conformations and no discernible pattern with the varying degree of interfacial interactions. This examination, in addition to the ACTR profile, reveals an interplay between the two morphing proteins, in which weak native-specific binding between them drives the folding of one protein while the other protein remains substantially unstructured.

Compared to the free form ensemble, which depicts much heterogeneity, the probability contact maps displaying the degree of intramolecular contacts, whether

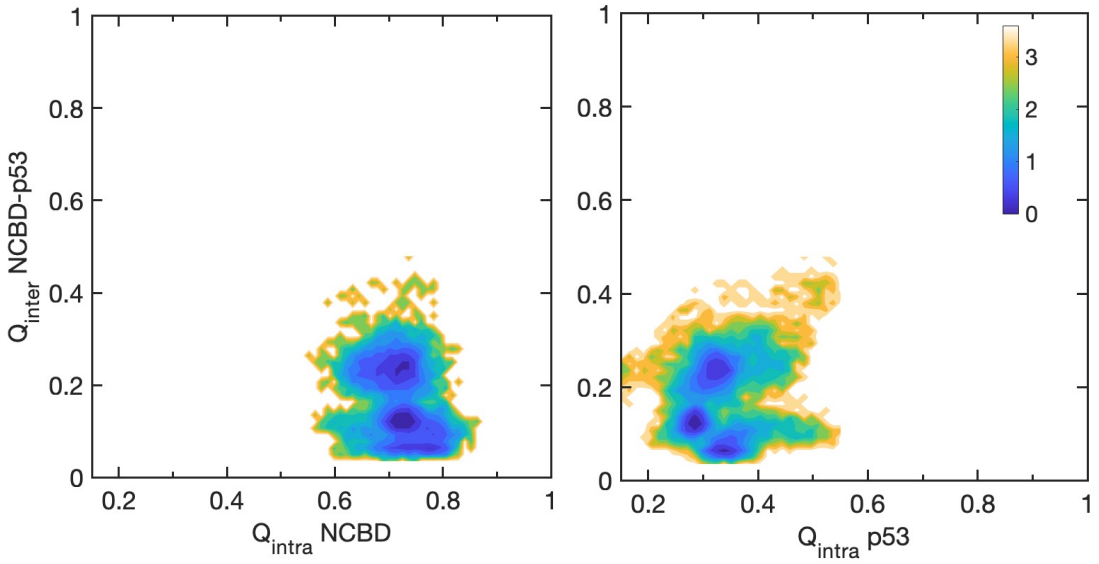

Figure 5.7: Projection of the NCBD:p53-TAD complex trajectories on the order parameters Q_{intra} (fraction of native contacts wrt. NMR structure) of NCBD (left) and p53-TAD (right), and Q_{inter} (fraction of native intermolecular contacts between NCBD and p53-TAD). Color bar is in kcal/mol.

native or non-native, illustrate the enhanced NCBD structural characteristics associated with p53 in Figure 5.8. The increased stability of NCBD is due to the strengthening of native contacts between helices 1 and 2, leading to the increased helical propensity in H2 (Figure 5.8 left). Furthermore, this enhanced coupling pattern causes H2 and H3 to interact less and H3 to expand onto the disordered C-terminal tail. Unlike the ACTR binding pattern, the binding poses of p53-TAD cause little interference with NCBD helices 1 and 2 (Figure 5.1.B), which allows sampling of a more structured NCBD ensemble.

With respect to transiently populated tertiary contacts, the unbound p53-TAD explores a broad conformational ensemble, while secondary structural elements remain flexible (Figure 5.8 right). The bound ensemble has a strikingly similar profile for the local elements with enhanced propensities in the C-terminal end of helix 1. The probability of non-local contacts between the two ends of p53-TAD is around 0.2-0.4 (Figure 5.8 bottom-right), indicating that the protein's binding poses undergo considerable variability, which can be further visualized in the highly dynamical binding interface of the two proteins shown in Figure 5.9.

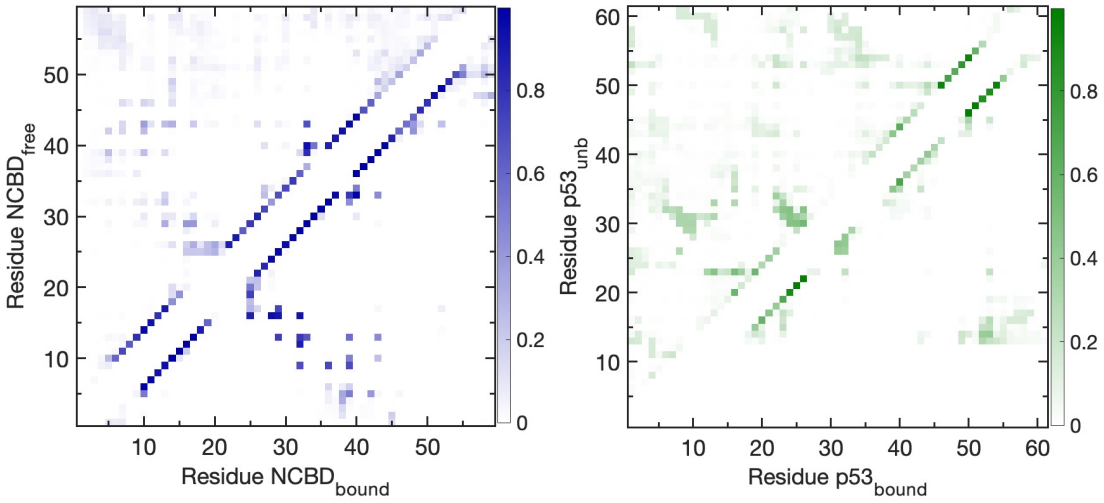

Figure 5.8: Total interaction map of NCBD and p53-TAD. Residue-residue total contacts (native and non-native) observed in the simulations of NCBD (left) and p53-TAD (right) bound (lower triangle) and free/unbound forms (upper triangle) for comparison. The color intensity reflects the time-averaged probability of observing the contact, with the light to dark color intensity corresponding to weakly to strongly interacting residues respectively. A contact is considered formed based on the definition in Figure 5.4

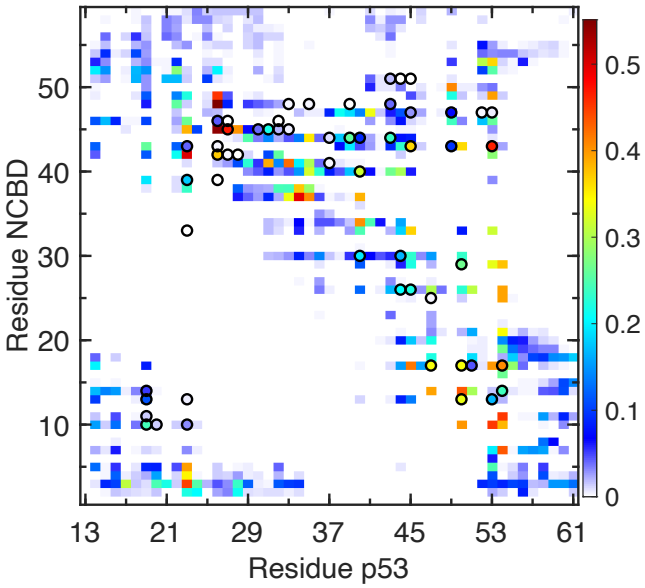

Figure 5.9: Inter-protein interactions in NCBD:p53-TAD complex. Residue-residue total contacts (native and non-native) observed in the simulations of NCBD bound to p53-TAD. Black circles indicate native contacts in the NMR structure. The color intensity reflects the time-averaged probability of observing the contact, with the blue to red color corresponding to weakly to strongly interacting residues respectively. A contact is considered formed based on the definition in Figure 5.4

The evaluation shows a high degree of non-specific contacts leading to diverse binding poses and a lack of multiple native protein-ligand interactions (empty black circles). For instance, the native salt bridge between R47 of NCBD and D49 of p53-TAD is known to contribute to the binding specificity of the NCBD:p53-TAD complex [23] and is found to be weakly connected in the obtained MD ensemble ($\approx$ 10% of the simulation time).

Overall, NCBD coupled to ACTR, and p53-TAD ensembles show a unique thermodynamic coupling between the two associated IDPs. One has a stronger native intramolecular effect, and the other populates mostly partly disordered conformations while sampling a large number of non-specific intermolecular interactions.

5.3.2 Coupled Folding and Binding of Morphing Protein to a Structured Protein

Next, we investigate the conformational dynamics of NCBD bound to the structured protein, IRF3, to obtain a better understanding of NCBD conformational properties when bound to different, structurally diverse ligand partners. For that, we carried out a 10 μs-long MD simulation starting from the crystallographically determined conformer of NCBD:IRF3 bound complex in explicit solvent. We find that IRF3 stabilizes onto a low RMSD of 0.25 $\mathring{A}$ sampling rigid, native-like conformations. The B-factors from the X-ray crystallography data show distinct profiles for NCBD and IRF3. High B-factors indicate greater uncertainty about the atom position suggesting that NCBD structure has higher flexibility across its topology compared to IRF3. We evaluated the conformational landscape of the NCBD:IRF3 ensemble in the vicinity of its bound conformation, and the results are shown in Figure 5.10.

When compared to IDP-IDP binding interactions, its dynamical binding mode has a very distinct profile. We observe reversible intermolecular transitions between native-like bound (Q >0.8), medium bound with Q ranging from 0.65-0.55, and loosely bound between 0.45 and 0.28.

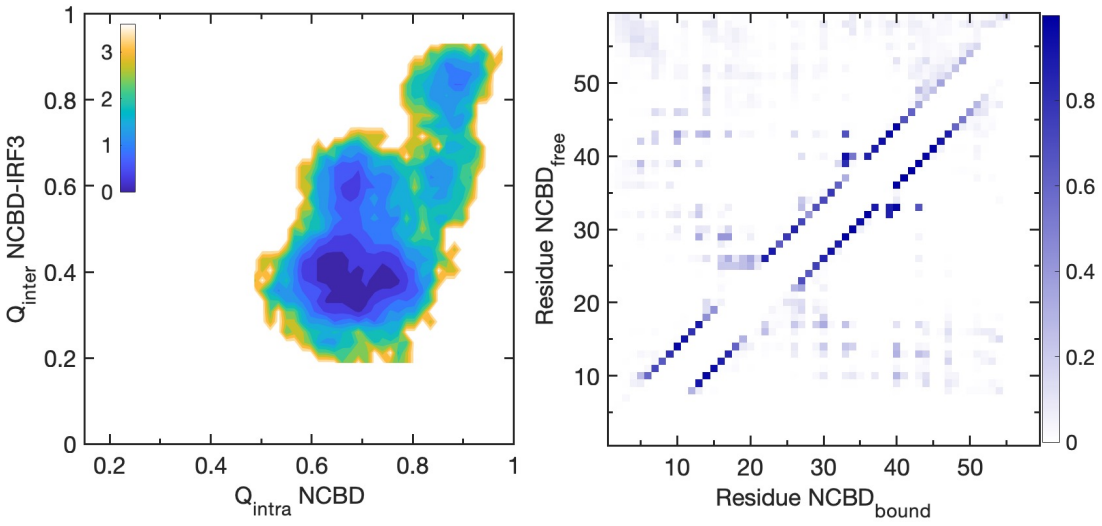

Figure 5.10: Structural features of NCBD associated to structured protein. Left. Projection of the NCBD:IRF3 complex trajectories on the order parameters Q_{intra} (fraction of native contacts wrt. NMR structure) of NCBD and Q_{inter} (fraction of native intermolecular contacts between NCBD and p53-TAD). Color bar is in kcal/mol. Right. Total interaction map of NCBD. Residue-residue total contacts (native and non-native) observed in the simulations of NCBD bound (lower triangle) and free forms (upper triangle) for comparison. The color intensity reflects the time-averaged probability of observing the contact, with the light to dark color intensity corresponding to weakly to strongly interacting residues respectively. A contact is considered formed based on the definition in Figure 5.4.

Although the binding pattern in IRF3:NCBD complex is drastically more superficial than the above-studied interaction partners, it showcases reasonably stronger specific binding. The highly interacting bound conformations in the early phase of the trajectory enabled sampling of more native-like conformers. Further, NCBD undergoes significant structural changes because of the marginal interaction between the two proteins. When loosely paired with IRF3, it populates a dramatically broad conformational ensemble (as observed for free form ensemble in Chapter 4) with Q ranging from 0.6 to 0.78. The overall conformational preferences observed here suggest that tighter binding induces more structuring in NCBD.

Compared to the free form MD ensemble, the structural diversity of NCBD (associated with IRF3) is further illustrated in the contact probability matrix in Figure 5.10 (right). Note that 12 residues are missing from the bound PDB structure, 1 to 6 residues from the N-terminal and 54 to 59 residues missing from the C-terminal end, and more importantly, it has a distinct topology in which the C-terminal end of helix 1 interacts with the N-terminal end of helix 3 as shown in Figure 5.1.C. Inspection of the contact map reveals that the secondary structural elements have dynamical properties similar to the free form ensemble with weakly interacting helices, manifested by the transient non-local contacts. The derived NCBD ensemble maintains its structural variability, is indeed quite diverse with a broad distribution of Q and exhibits surprisingly low conformational biases against its specific structural arrangement of the helices (helices 1 and 3 interactions) even when bound to the structured protein, IRF3.

The analysis of the binding interface dynamics indicates that the loss of specific contacts of NCBD helix 1 with IRF3 induces the conformational flexibility in NCBD, as shown in the Figure 5.11, and that the H2 and H3 of NCBD mostly retain their native binding contacts (black circles). In addition, compared to ACTR and p53-TAD, the overall binding pattern is observed to be much less variable as fewer non-specific contacts are sampled. This finding is interesting since NCBD:IRF3 binding pattern is superficial compared to the entangled IDP

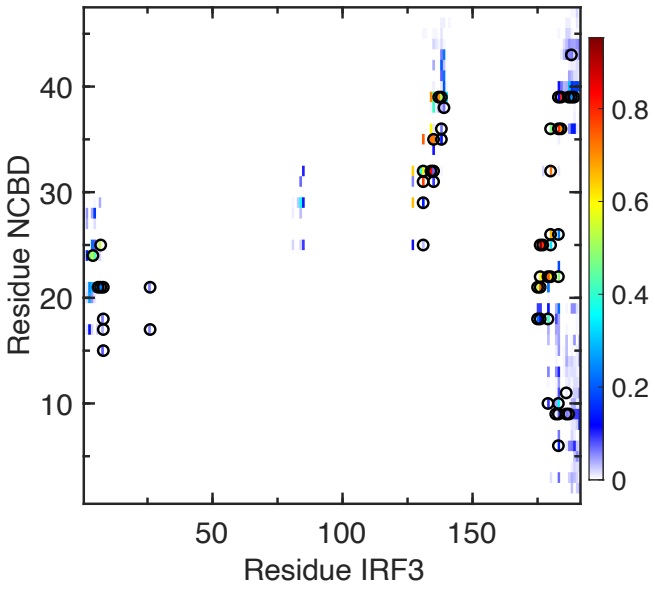

Figure 5.11: Inter-protein interactions in NCBD:IRF3 complex. Residue-residue total contacts (native and non-native) observed in the simulations of NCBD bound to IRF3. Black circles indicate native contacts in the NMR structure. The color intensity reflects the time-averaged probability of observing the contact, with the blue to red color corresponding to weakly to strongly interacting residues respectively. A contact is considered formed based on the definition in Figure 5.4.

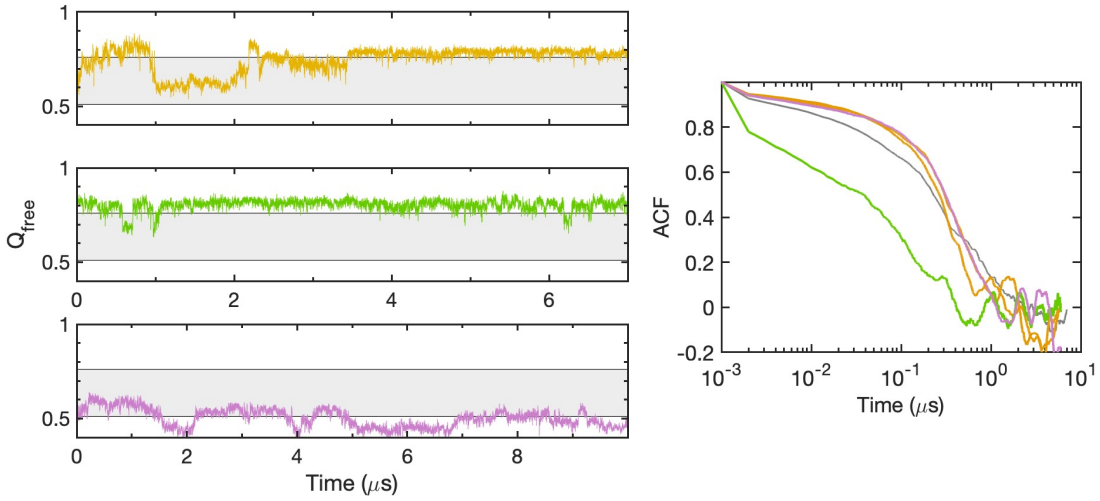

Figure 5.12: Left. Time trajectories as a function of fraction of native contacts wrt. to free form NMR structure (Q_{free}) across three distinct NCBD MD ensembles in the presence of ligand partners; ACTR (yellow), p53-TAD (green) and IRF3 (pink).The grey swath marks the limits of the Q_{free} in the free form ensemble. Right. Autocorrelation function of the Q_{free} for the NCBD bound with ACTR (yellow), p53-TAD (green) and IRF3 (pink). Grey profile indicates the average behavior of NCBD free ensemble for reference (from Figure 4.6).

conformers ACTR and p53-TAD binding poses, where we observed extensive non-specific interactions.

Finally, we examined the NCBD relaxation times when bound to its different interaction partners to gain insights into its dynamical behavior with and without the partner context. To do so, we looked at the time evolution of the fraction of contacts in all the bound NCBD trajectories with reference to the free form NMR structure. The time trajectories as a function of Q_{free} are depicted in the Figure 5.12. The ACTR-bound trajectory undergoes significant structural changes, while the p53-TAD bound trajectory, consistently populates enhanced native-like conformations, sampling a reasonably stable MD ensemble. NCBD associated with IRF3 undergoes reversible transitions exhibiting an interplay between the decrease in native intermolecular contacts, there is an apparent increase in free form-like conformations.

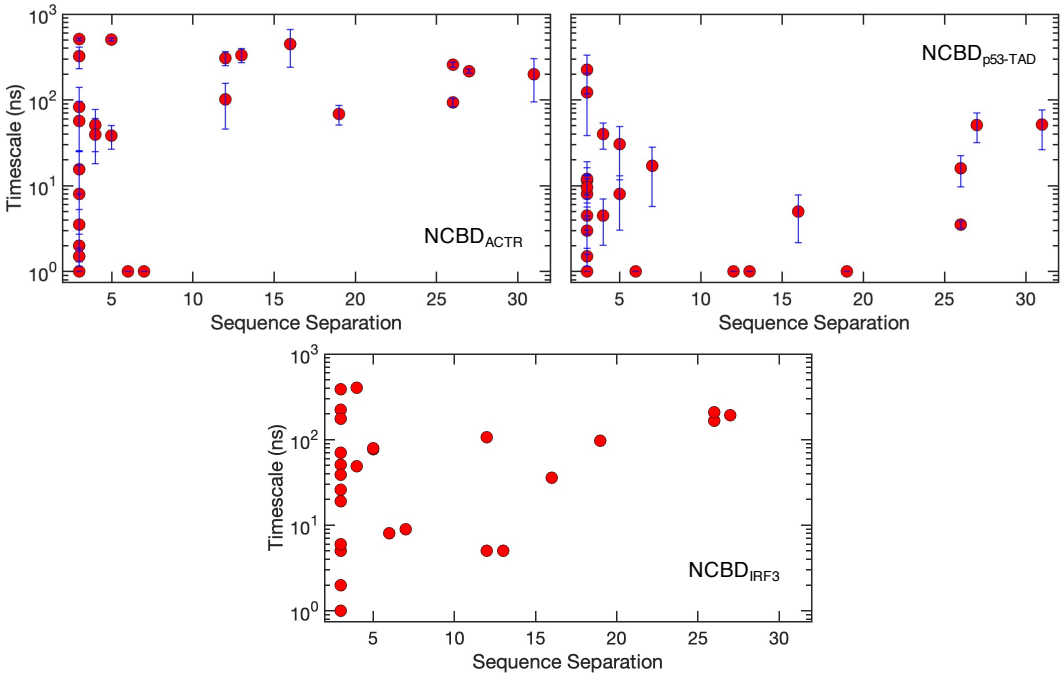

Figure 5.13: Dynamics of intramolecular contacts. Timescale (ns) vs. sequence separation of C_β contacts extracted from the NMR free form structure and their center-of-mass (COM) distances computed against the three distinct NCBD MD ensembles bound to ligand partners; ACTR (yellow), p53-TAD (green) and IRF3 (pink). The analysis follows the same definition as in Figure 4.10.

The autocorrelation function of Q_{free}, as shown in the Figure 5.12, demonstrates interesting morphing features of NCBD. The grey curve represents the average characteristic profile of free form NCBD ensemble from Chapter 4. All bound profiles show uncertainty at longer lag times. The NCBD bound to p53-TAD undergoes the fastest relaxation in less than 100 ns since it re-equilibrates onto a rather stable ensemble while ACTR and IRF3 bound exhibit sub-μs characteristic times. In addition, we calculated the timescales of the short-, mid-, and long-range contacts from the C_β COM distance evolution over time with reference to the free form NMR structure as defined in Chapter 4 (Figure 4.10), which further highlights the differential timescales of intramolecular interactions of the various bound NCBD conformers (Figure 5.13).

Overall, these findings show that NCBD's intrinsic versatility differs depending on the partner context; when bound to ACTR and IRF3, it remains party

disordered, while when bound to p53-TAD, it stabilizes into a structured ensemble, offering mechanistic insights onto its coupled morphing and binding behavior.

5.4 Discussion

The emerging synergy between simulations of various resolution and experiments has established simulations as a promising tool for delineating the structural properties of disordered proteins. This Chapter provides an in-depth assessment of one such IDP, NCBD, and its association with three structurally diverse biological partners. NCBD aids in recruiting the transcription machinery via mediating interactions with other morphing proteins such as ACTR, p53-TAD, and a structured protein IRF3 [6]. Together, these partners exemplify the most dissimilar NCBD complexes.

First, we provide a detailed picture of ACTR and p53-TAD conformational dynamics in the absence of NCBD. We observed greater extent of structural disorder in both IDPs as reported in various studies [9],[17],[34] compared to NCBD free ensemble in Chapter 4. Next, we studied each of the two IDPs bound to NCBD using multiple atomistic simulations. We found that the bound IDP-IDP conformations exhibited heterogeneity of binding modes instead of a single, stable complex. The MD ensembles of NCBD bound to ACTR and p53-TAD demonstrate a distinct thermodynamic coupling between the two, with one having a greater intramolecular effect and the other sampling partly disordered conformations while loosely bound to each other with respect to the native binding pattern. NCBD morphing in association with IRF3 shows an intriguing observation: NCBD retains much of its fundamental disorder even when bound to a structured protein, revealing a subtle interplay between the conformational landscape of NCBD and partner binding.

Most contacts between NCBD and the three ligand partners are mediated by hydrophobic interactions and several hydrogen bonds. Thus, we examined the intermolecular hydrogen bond (I-Hbond) network in all three complex trajectories

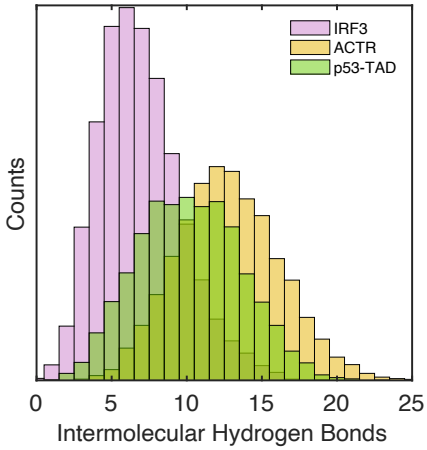

Figure 5.14: Distribution of intermolecular hydrogen bonds in three NCBD MD ensembles bound to ligand partners; ACTR, p53-TAD and IRF3.

and found a pattern that corroborates the known binding affinities of the interaction partners. The distribution for each of the bound ensembles is plotted as a function of the number of I-Hbonds between NCBD and each of its three ligand partners in Figure 5.14. NCBD:IRF3 ensemble has the fewest I-Hbond count and a low binding affinity with $K_d \approx 100$ μM [22], followed by P53-TAD with $K_d \approx 1.7$ μM [23] and relatively higher affinity for ACTR ($K_d \approx 34$ nm) [17], which displays the maximum I-Hbonds. The obtained network shows marked behavior, as intermolecular hydrogen bonding discriminated NCBD:ligand partner dynamics rather distinctly, with number and relation to the known binding affinity of each NCBD complex offering high-resolution insights into its morphing coupled to binding behavior.

The structural flexibility of IDPs has key functional consequences in biology. The broader conformational sampling of the three NCBD complexes capture the various non-native binding poses, which reflects the inherent structural diversity of NCBD and its functioning as a rheostat in coordinating dynamic assemblies, advancing our understanding of the complex binding modes of IPDPs.

CHAPTER 6

Conclusion and Future Direction

In biology, IDPs play an important role. IDPs populate various ensembles, from expanded states with little residual structure to more compact ensembles with residual secondary and tertiary interactions, and they often fold as they bind to their partners. Such IPDPs are found at the hub of various protein interaction networks, especially cellular signaling and regulation. How they use their intrinsic disorder to interact with multiple partners by folding upon binding mechanism and mediate protein functions remains elusive.

Our working hypothesis is that the IPDPs central to the regulation of essential protein interaction networks are conformational rheostats (CR), i.e., that their dynamic ensembles have innate capabilities to morph onto structures that are specifically compatible with their diverse partners gradually. The built-in conformational biases in CRs can enable new functionalities ranging from broadband sensitivity (in the partner(s) concentration), sub-microsecond (fast timescale) response, and promiscuous binding with multiple partners. These remarkable properties link IPDPs to downhill folding regime. This dissertation investigates different morphing proteins using protein engineering, long-scale all-atom MD simulations, and integrated computational and experimental analyses to gain important mechanistic insights into the functional role of morphing proteins.

In our quest to seek insights into the working of a conformational rheostat for practical applications, Chapter 2 illustrated an integrated computational and experimental approach to engineer molecular scaffolds using a fast-folding protein that loses/gains structure gradually (downhill folding) upon change in analyte concentration. Such a protein-based design can exhibit a broadband sensitivity in a rheostatic fashion and has exciting advantages over extant protein-based

transducers, which conventionally operate as a conformational switch. The histidine grafting strategy we employed introduces sensitivity to pH in the mildly acidic to neutral range systematically. Selecting the degree of histidine burial and the number of grafts highlight that the strategy is customizable. MD simulations at the few microseconds time scale seem to reproduce the conformational effects of local perturbations (mutation and ionization). The observed progressive structural (dis)ordering can reach >6 orders of magnitude in [H+], demonstrating that the rheostatic conformational transducer adds a new tool to the biosensor engineering toolbox and can be extremely useful in tracking widely fluctuating biological variables.

Despite advances in describing IDP structural features and discovering correlations between sequence and conformational features [47], the underlying interaction network and folding energetics of IDPs remain poorly understood. In Chapter 3, we devised a novel modular approach to measure the folding cooperativity and the energetic contributions of local and tertiary interactions in defining the conformational ensemble and binding properties of IPDPs. Here, we dissect the NCBD protein domain into all its elementary components (secondary and super-secondary structural elements) defined in a staggered fashion. Then, using an integrated strategy of computation and experiments, we perform an ensemble-based conformational analysis of all the components and establish the interactions between them by directly comparing relevant components. We found dynamical coupling of various secondary and non-local interactions and significant conformational biases in NCBD, supporting its conformational rheostatic capabilities. The agreement achieved in our experiments and simulations ensures that the observations made in this study are statistically reliable, which is of considerable relevance given an inherently flexible system such as NCBD. This approach provides an exciting tool for analyzing morphing transitions that should generally apply to any IPDP.

Further, in Chapters 4 and 5, we analyzed NCBD unbound ensembles and dissected the interplay of NCBD with its structurally diverse partners. Investigating

dynamical interactions in the unbound IDP can provide quantitative information about the properties of the free state and functional consequences of any free state interactions in regulating biological activity. Using extensive atomistic simulations with high-resolution residue-based analysis and Markov State modeling, we decipher the underlying heterogeneity in NCBD. NCBD populates distinct structural ensembles in its free state that it forms in complex with partners. We gain a fundamental understanding of how the structure and interaction of NCBD are controlled and regulated via their conformational-folding landscape. We find there is indeed a coupling of NCBD conformational dynamics and its binding specificity. Overall, our analysis captures the gradual conformational transitions of NCBD, providing substantial evidence of a partially disordered protein functioning as a conformational rheostat in recruiting the transcription machinery.

Taken together, our research shows that advancements in force fields and computing hardware have now enabled the development of innovative methods that combine computational and experimental techniques to uncover insights that cannot be obtained using either method alone. This dissertation, we believe, provides a unique perspective on the structure, thermodynamics, and kinetics of morphing proteins, allowing us to better comprehend their dynamic behavior.

6.0.1 Future Directions

In understanding the physical relevance of conformational disorder, our present research work provides key insights into the working of molecular rheostats in biology.

First, our work on engineering protein-based scaffolds to thermodynamically couple their gradual (un)folding to histidine ionization paves the way for fast-folding proteins to be targeted for real-time analog outputs with broad sensitivity (to the analyte of interest), enabling the development of conformational transducers for biosensing applications. The insights learned from the histidine grafting methodology should apply to the development of conformational transducers for analytes with a structurally defined binding site. In such circumstances, engi-

neering the folding characteristics of a protein domain that already contains the target ligand's binding site may be the most straightforward way. By enhancing secondary structure propensity and decreasing the hydrophobic core, one may modify the protein sequence to make it both intrinsically unstable and downhill-like and balancing the total perturbation such that it folds gradually upon binding to the ligand (folding upon binding transducer). Protein engineering combined with microseconds long all-atom MD simulations and experiments provides an integrated platform to engineer scaffolds for developing biosensors with remarkable properties. Based on these design principles, our group is currently working on engineering calcium, ATP, and COVID-19 conformational transducers with fluorescence readouts for real-time biosensing applications and has achieved promising results.

Second, IDPs offer novel advantages as therapeutic targets as they are implicated in human diseases and protein design as they make attractive scaffolds for potential engineering applications. However, these endeavors rely heavily on their structural and folding information, which have been an ongoing challenge to probe and analyze. To this end, combining energetic data with structural information can yield even more insights into IDP unique properties, highlighting the importance of the LEGO approach. MD simulations are a powerful tool to examine such morphing proteins as they can probe the rapidly interconverting conformations to provide a high-resolution molecular characterization of their dynamics. Further enhanced sampling and improved force fields can further help optimize and accurately characterize more disordered proteins using the LEGO approach. Particularly, fragmentation of the IDP is a critical step to make sure all relevant structural components are preserved that help resolves significant interactions in a rather heterogeneous system. Because of the extensibility of this tool, other IPDPs can be investigated in detail to gain mechanistic insights into their morphing behavior.

Computationally, atomistic simulations of NCBD and its partners highlight the importance of studying the conformational dynamics of morphing proteins

after the binding event. Numerous studies have focused on characterizing the binding event and have increased our knowledge on the formation of encounter complexes stabilized by 'non-specific' electrostatic interactions. Our obtained MD ensembles in the vicinity of the essential subspace of the bound complexes (PDBs) point to the inherent structural disorder in morphing proteins even when bound to the interaction partners offering key details into their binding specificity. Further NMR experiments in combination with simulations can dissect their morphing coupled to binding behavior in unprecedented detail.

Finally, the hidden conformational biases and dynamics observed in NCBD provide the first demonstration of a working conformational rheostat regulating a critical biological process, the eukaryotic transcription. The concept of a conformational rheostat offers a novel molecular mechanism for controlling protein function gradually rather than in a binary fashion. In addition, our group's recent experimental data on another morphing protein shows evidence of CRs mediating gene regulation in eukaryotes. The relationship between morphing behavior in IDPs and their rheostatic capabilities deserves further exploration. Since IDPs are still considered undruggable, these research endeavors can tremendously advance our understanding of their fascinating dynamic nature.

APPENDIX A

Appendices

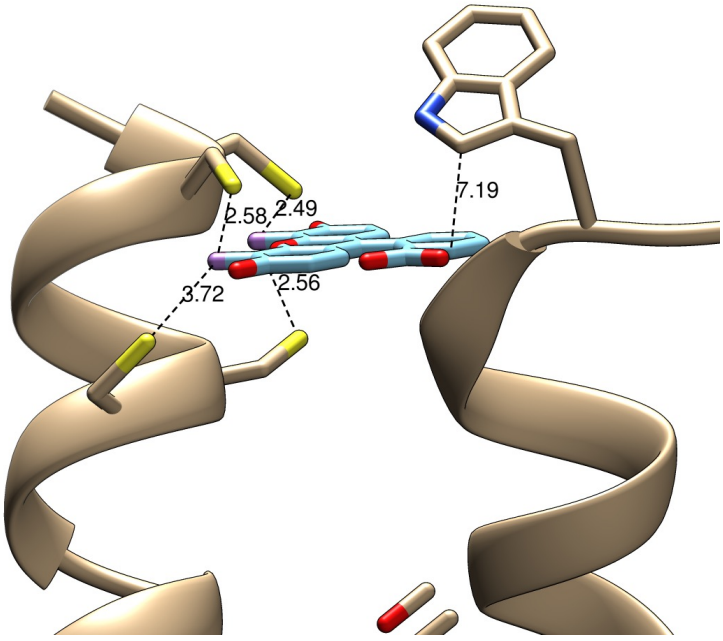

Figure A.1: Designed FlAsH-EDT_2 dye binding motif in gpW F35H mutant for signal readout.

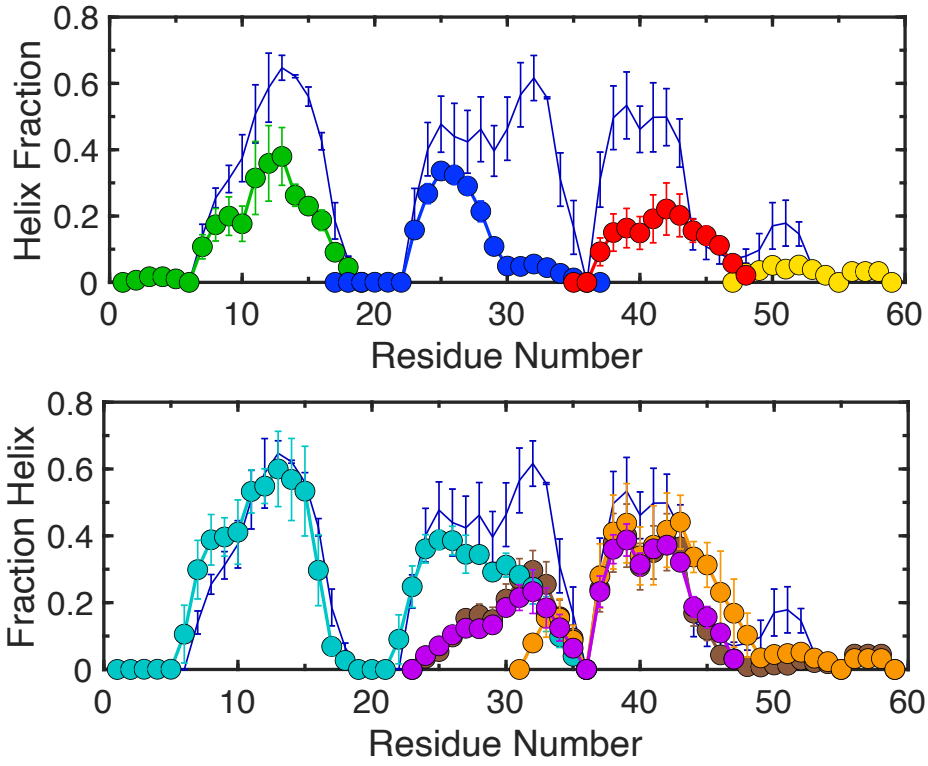

Figure A.2: Helix fraction per residue based on hydrogen bond definition for all MD ensembles. Top: building blocks. Bottom: combinations of building blocks. The full-length protein is shown with thin, navy blue lines as reference. Color coding as in Figure 3.1

Residue	Residue
P1	S4
I5	N36
S6	A9
P7	L10
A9	D12
A9	S35
A9	N36
L10	L13
L10	N36
Q11	L14
D12	R15
L13	T16
L13	I32
L14	L17
T16	Q28
T16	V29
T16	I32
L17	V29
P20	Q25
S22	Q25
Q24	Q27
Q24	Q28
Q25	V29
Q28	N31
V29	I32
L30	L33
N31	K34
I32	S35
L33	L39
L33	M40
N36	L39
P37	M40
Q38	A41
L39	A42
M40	F43
A41	I44
A42	K45
F43	Q46
R47	K50
R47	V52

Table A.1: List of native C_β contacts in the NMR free form NCBD structure within 0.6 nm distance.

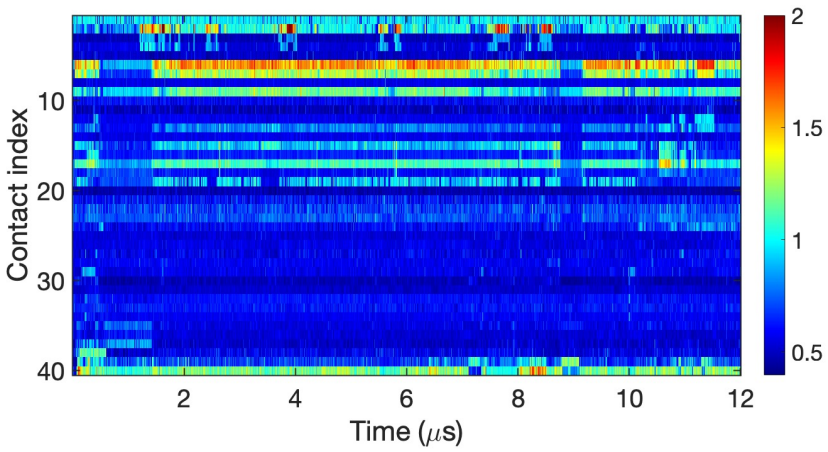

Figure A.3: Distance evolution of C_β contacts of NCBD trajectory 2.

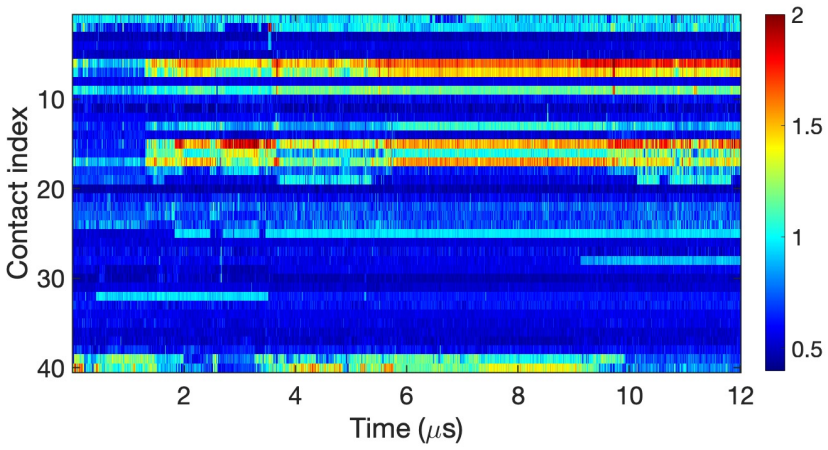

Figure A.4: Distance evolution of C_β contacts of NCBD trajectory 3.

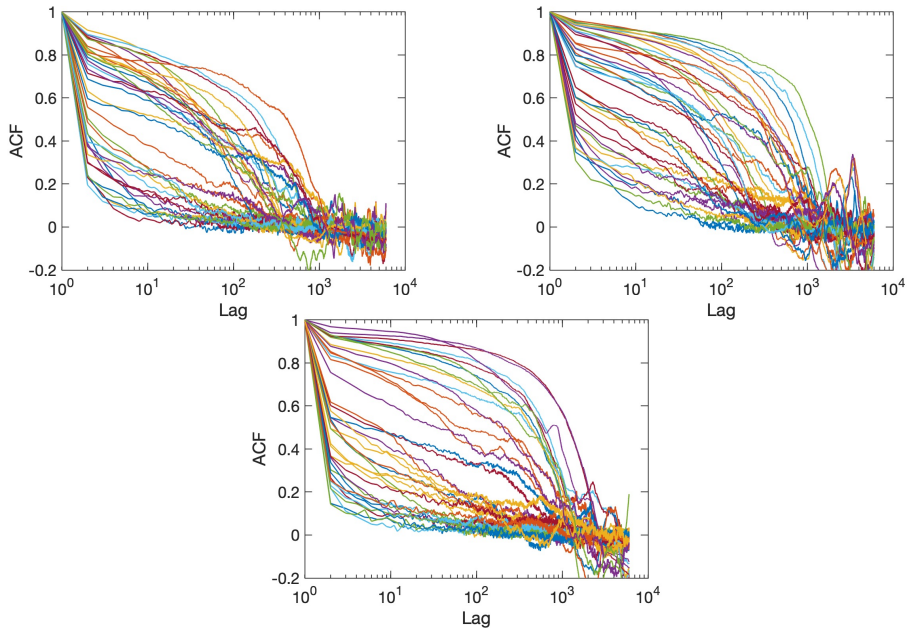

Figure A.5: Autocorrelation function of C_β contacts distances across three free form NCBD trajectories.

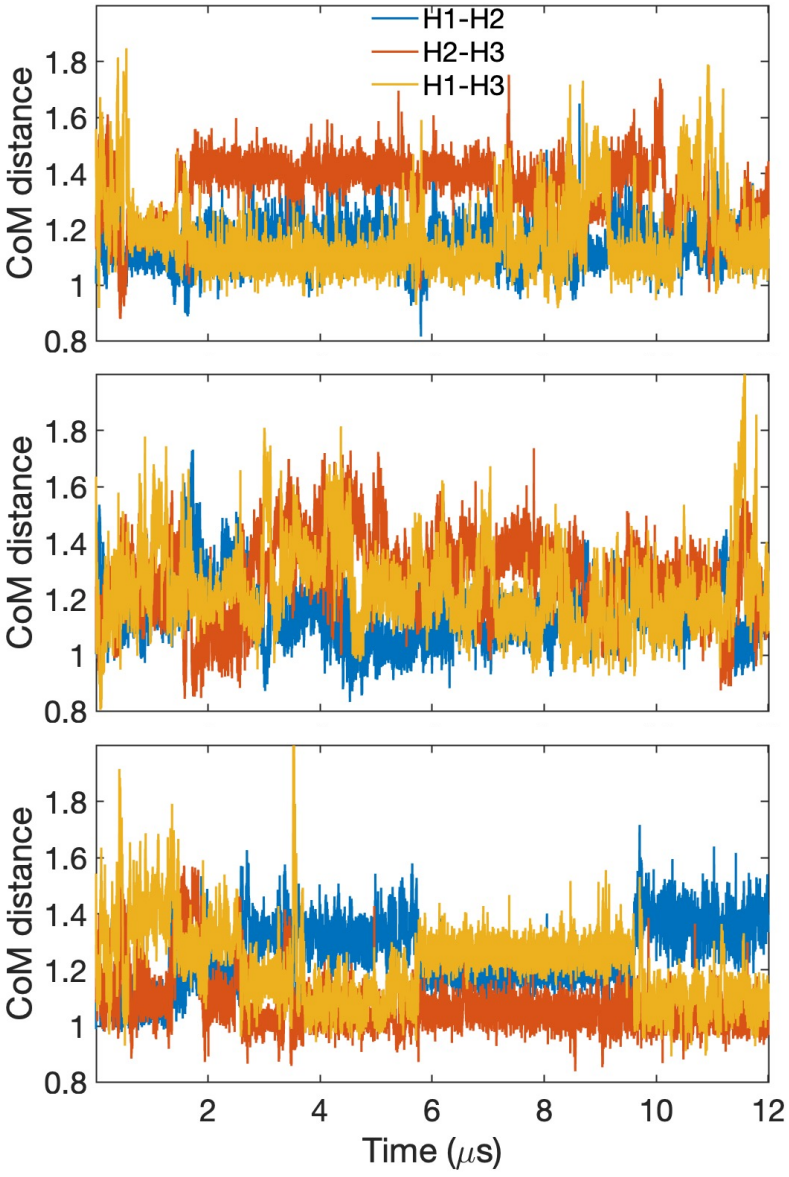

Figure A.6: Time evolution of Center-of-Mass distances between helix pairs in free form NCBD across three trajectories.

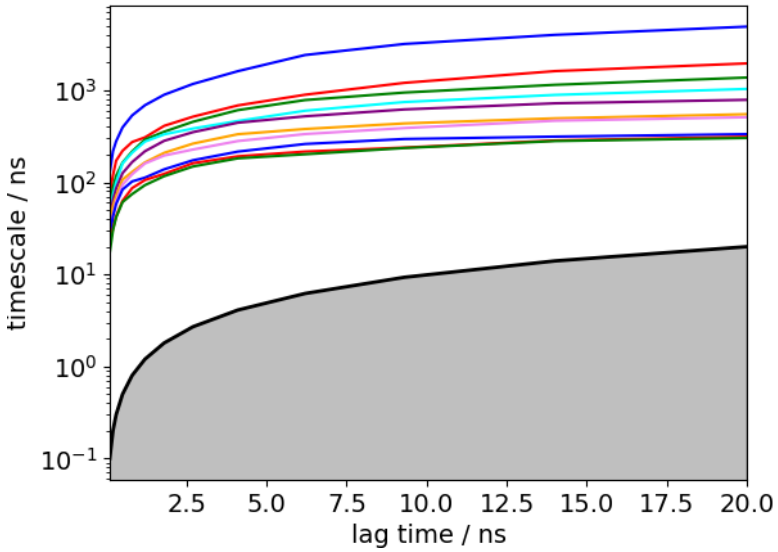

Figure A.7: Implied timescales of the NCBD Markov state model. The top 10 implied timescales of the MSMs calculated at a range of lag times are shown: The gray area signifies the region where timescales become equal to or smaller than the lag time and can no longer be resolved. The lag time of 7 ns is chosen for our models, as the timescales have approximately leveled off at that point.

A.1 Nucleobindin-1 Structural Modeling and Design

Designing 'on' activated and 'off' calcium unbound state

On State The Ca^{2+} bound state of the nucleobindin-1 spanning from 149 to 408 residues was modeled based on the NMR determined Ca^{2+} binding domain (calnuc) folded structure (PDB ID:1SNL). The annotated leucine zipper domain has a $\approx 90\%$ probability to form a coiled coil dimer evaluated from the Multicoil2 statistical tool. Based on the known structural and functional attributes, the DNA binding domain (DBD) is assumed to form an elongated helix resulting from the calcium bound folded state of calnuc which also triggers the formation of a parallel leucine zipper, forming an extended homodimer.

Structural modeling The DBD ranging from residue 180 to 236 was first modeled using CC Fold tool as a parallel coiled coil (homodimer) and using this

Figure A.8: Functional annotation of Nucleobindin-1

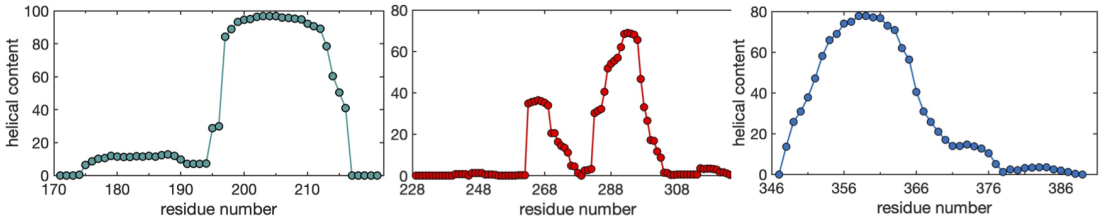

Figure A.9: Helical propensity calculated with AGADIR of different functional domains of Nucleobindin-1

template, comparative modeling was performed using Robetta web server. The output model was further used for modeling the 'on' state. Calcium binding domain (Calnuc) structure determined from NMR was retrieved from the protein data bank (PDB ID:1SNL). Calnuc spans from residue to 228 to 326. We removed the residue 228 to 245 (disordered N-term loop) and the 81 residue-structured calnuc (K246-F326) was used for subsequent modeling. The Leucine zipper (LZ) segment ranging from 339 to 408 was first modeled using CC Fold tool as a parallel coiled coil (homodimer) and then using the homodimer as a template, we performed comparative modeling using Robetta. The loop (G327-M338) connecting calnuc and LZ was built using USCF Chimera and refined with Modeller. Using the most probable loop conformation, we first created a peptide bond between the loop (M338) and one of the LZ monomer (H339) from the Robetta model and then joined the model with calnuc domain with a second peptide bond between calnuc (F326) and N-terminal (G327) of the connecting loop. The calnuc-LZ model (228-408) was then energy minimized and a copy was also saved as the second monomer, removing the free (non-bonded) LZ monomer. Next, the two calnuc-LZ monomers were dimerized based on the structural alignment (0.2 Angstrom RMSD cutoff) with reference to the original LZ homodimer model (not bonded to calnuc). Following the homodimer modeling, the two monomers

were slightly moved carefully with mouse movements (Y Z direction) to rectify for any apparent steric hindrances. Once the dimer model was visually optimized, the disordered segments (237-245) connecting the calnuc and DBD were built and each connected with one calnuc monomer (K246). Next, the connected segments were refined, the DBD coiled coil homodimer was connected to each of the segment-calnuc-LZ monomers. Finally, the N-terminal disordered segments (149-179) was built and connected to each of DBD-calnuc-LZ monomer model. The segments were refined and the whole dimer complex was then used as a template for comparative modeling with Robetta. The final 'on' (calcium bound) model is shown below in cartoon representation. The DNA binding domain helices have a 1.4 nm inter-helical distance as they start to open from C to N-terminal direction, sufficient to accommodate interactions with the DNA.

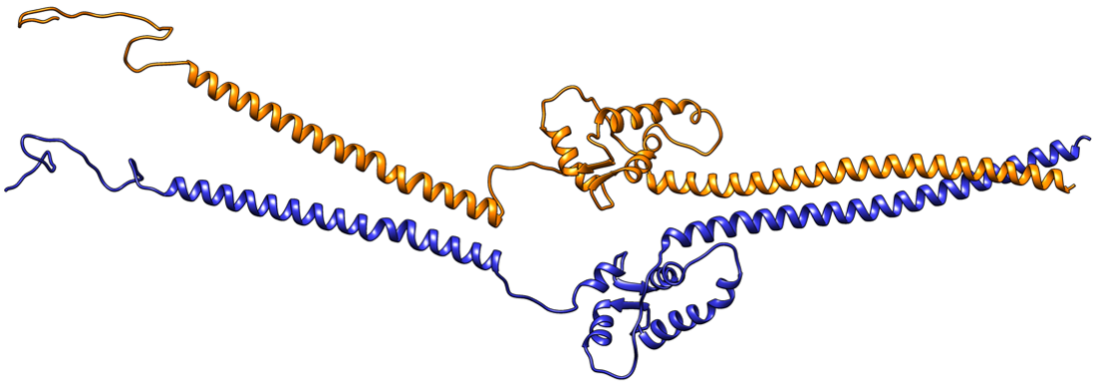

Figure A.10: Open state. Calcium bound state

Off State Modeling of Ca^{2+} unbound structure ranging from residue 149 to 408 was carried out based on the hypothesis that without Ca^{2+} calnuc domain remains conformationally flexible which allows for the overall structure to reorient, resulting in intra-protein coiled coil formation between the LZ domain and the DBD.

Structural modeling

LZ- H339 to G408, DBD- F149 to K218

First a parallel hetero coiled coil LZ – DBD was modeled using CC Fold web tool, followed by using the output model as the template for comparative

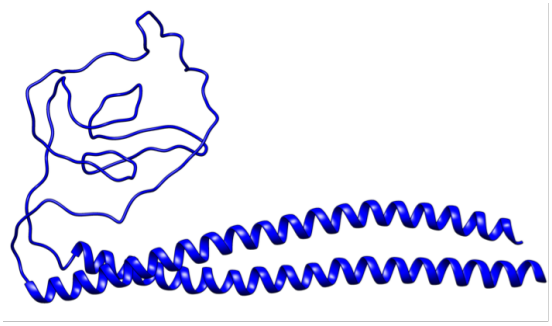

Figure A.11: Close state. Calcium unbound state

modeling. The Robetta model was used to construct the off state. The whole unstructured calnuc with extended disordered loops ranging from V219 to M338 residue was built using chimera as a 120 residue long random coil and then refined with Modeller. The conformation with minimal end to end (N- and C-term) distance was chosen and N-term (V219) was joined with a peptide bond to C-term of DBD (K218) and C-term of the unstructured calnuc (M338) was connected to the LZ (H339). The complete structure was then used as a template for comparative modeling with Robetta and the final structure is shown in Figure.

Both Ca^{2+} bound and unbound conformations have been modeled carefully to make sure most hydrophobic residues are buried in the core/interface and polar charged residues (mostly arginines) are in interfacial region of the DNA binding domain to account for interaction with DNA (Ca^{2+} bound state)

References

[1] Peter E Wright and H.Jane Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331, oct 1999.

[2] Rita Pancsa and Peter Tompa. Structural Disorder in Eukaryotes. *PLoS ONE*, 7(4):e34687, apr 2012.

[3] Vladimir N. Uversky. A decade and a half of protein intrinsic disorder: Biology still waits for physics. *Protein Science*, 22(6):693–724, jun 2013.

[4] Rebecca B. Berlow, H. Jane Dyson, and Peter E. Wright. Functional advantages of dynamic protein disorder. *FEBS Letters*, 589(19PartA):2433–2440, sep 2015.

[5] Peter Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533, oct 2002.

[6] Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29, jan 2015.

[7] Peter E Wright and H Jane Dyson. Linking folding and binding. *Current Opinion in Structural Biology*, 19(1):31–38, feb 2009.

[8] Kenji Sugase, H. Jane Dyson, and Peter E. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, 447(7147):1021–1025, jun 2007.

[9] H. Jane Dyson. Expanding the proteome: disordered and alternatively folded proteins. *Quarterly Reviews of Biophysics*, 44(4):467–518, nov 2011.

[10] M Madan Babu. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society transactions*, 44(5):1185–1200, 2016.

[11] Vladimir N. Uversky. What does it mean to be natively unfolded? *European Journal of Biochemistry*, 269(1):2–12, jan 2002.

[12] Christopher J. Oldfield and A. Keith Dunker. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annual Review of Biochemistry*, 83(1):553–584, jun 2014.

[13] Pedro Romero, Zoran Obradovic, Xiaohong Li, Ethan C. Garner, Celeste J. Brown, and A. Keith Dunker. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Genetics*, 42(1):38–48, jan 2001.

[14] M. M. Babu, R. W. Kriwacki, and R. V. Pappu. Versatility from Protein Disorder. *Science*, 337(6101):1460–1461, sep 2012.

[15] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends in Biochemical Sciences*, 33(1):2–8, jan 2008.

[16] Alessandro Borgia, Madeleine B Borgia, Katrine Bugge, Vera M Kissling, Pétur O Heidarsson, Catarina B Fernandes, Andrea Sottini, Andrea Soranno, Karin J Buholzer, Daniel Nettels, Birthe B Kragelund, Robert B Best, and Benjamin Schuler. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, 555(7694):61–66, 2018.

[17] Stephen J. Demarest, Maria Martinez-Yamout, John Chung, Hongwu Chen, Wei Xu, H. Jane Dyson, Ronald M. Evans, and Peter E. Wright. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*, 415(6871):549–553, jan 2002.

[18] Allan Chris M. Ferreon, Josephine C. Ferreon, Peter E. Wright, and Ashok A. Deniz. Modulation of allostery by protein intrinsic disorder. *Nature*, 498(7454):390–394, jun 2013.

[19] V. J. Hilser and E. B. Thompson. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proceedings of the National Academy of Sciences*, 104(20):8311–8315, may 2007.

[20] Hesam N. Motlagh, James O. Wrabl, Jing Li, and Vincent J. Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, apr 2014.

[21] Peter Tompa, Csilla Szász, and László Buday. Structural disorder throws new light on moonlighting. *Trends in Biochemical Sciences*, 30(9):484–489, sep 2005.

[22] Bin Y Qin, Cheng Liu, Hema Srinath, Suvana S Lam, John J Correia, Rik Derynck, and Kai Lin. Crystal structure of IRF-3 in complex with CBP. *Structure (London, England : 1993)*, 13(9):1269–77, sep 2005.

[23] C.W. Lee, M.A. Martinez-Yamout, H.J. Dyson, and P.E. Wright. Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator binding domain of creb binding protein. *Biochemistry*, 2010;49(46):9964-71. PubMed PMID: 20961098; PMCID: PMC2982890.

[24] Victor Muñoz. Conformational Dynamics and Ensembles in Protein Folding. *Annual Review of Biophysics and Biomolecular Structure*, 36(1):395–412, jun 2007.

[25] Athi N. Naganathan, Urmi Doshi, Adam Fung, Mourad Sadqi, and Victor Muñoz. Dynamics, Energetics, and Structure in Protein Folding †. *Biochemistry*, 45(28):8466–8475, jul 2006.

[26] P. Li, F. Y. Oliva, A. N. Naganathan, and V. Munoz. Dynamics of one-state downhill protein folding. *Proceedings of the National Academy of Sciences*, 106(1):103–108, jan 2009.

[27] Luis Alberto Campos, Mourad Sadqi, and Victor Muñoz. Lessons about Protein Folding and Binding from Archetypal Folds. *Accounts of Chemical Research*, 53(10):2180–2188, oct 2020.

[28] M. Cerminara, T. M. Desai, M. Sadqi, and V. Munoz. Downhill protein folding modules as scaffolds for broad-range ultrafast biosensors. *J. Am. Chem. Soc.*, 134(19):8010–3, 2012.

[29] Victor Muñoz, Luis A Campos, and Mourad Sadqi. Limited cooperativity in protein folding. *Current Opinion in Structural Biology*, 36:58–66, feb 2016.

[30] A.N. Naganathan and M. Orozco. The native ensemble and folding of a protein molten-globule: functional consequence of downhill folding. *J Am Chem Soc*, 2011;133(31):12154-61. PubMed PMID: 21732676.

[31] Y. Wang, X. Chu, S. Longhi, P. Roche, W. Han, E. Wang, and J. Wang. Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proceedings of the National Academy of Sciences*, 110(40):E3743–E3752, oct 2013.

[32] H.Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*, 12(1):54–60, feb 2002.

[33] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nature reviews. Molecular cell biology*, 6(3):197–208, mar 2005.

[34] Duy Phuoc Tran and Akio Kitao. Kinetic Selection and Relaxation of the Intrinsically Disordered Region of a Protein upon Binding. *Journal of Chemical Theory and Computation*, 16(4):2835–2845, apr 2020.

[35] Zhirong Liu and Yongqi Huang. Advantages of proteins being disordered. *Protein Science*, 23(5):539–550, may 2014.

[36] Yongqi Huang and Zhirong Liu. Kinetic Advantage of Intrinsically Disordered Proteins in Coupled Folding–Binding Process: A Critical Assessment of the "Fly-Casting" Mechanism. *Journal of Molecular Biology*, 393(5):1143–1159, nov 2009.

[37] Veronika Csizmok, Ariele Viacava Follis, Richard W. Kriwacki, and Julie D. Forman-Kay. Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chemical Reviews*, 116(11):6424–6462, jun 2016.

[38] S.L. Shammas, J.M. Rogers, S.A. Hill, and J. Clarke. Slow, Reversible, Coupled Folding and Binding of the Spectrin Tetramerization Domain. *Biophysical Journal*, 103(10):2203–2214, nov 2012.

[39] Yongdae Shin and Clifford P. Brangwynne. Liquid phase condensation in cell physiology and disease. *Science*, 357(6357):eaaf4382, sep 2017.

[40] Monika Fuxreiter, Peter Tompa, István Simon, Vladimir N Uversky, Jeffrey C Hansen, and Francisco J Asturias. Malleable machines take shape in eukaryotic transcriptional regulation. *Nature Chemical Biology*, 4(12):728–737, dec 2008.

[41] Rajendra Sharma, David De Sancho, and Victor Muñoz. Interplay between the folding mechanism and binding modes in folding coupled to binding processes. *Physical Chemistry Chemical Physics*, 19(42):28512–28516, 2017.

[42] Alexey G. Kikhney and Dmitri I. Svergun. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters*, 589(19PartA):2570–2577, sep 2015.

[43] Lauren Ann Metskas and Elizabeth Rhoades. Single-Molecule FRET of Intrinsically Disordered Proteins. *Annual Review of Physical Chemistry*, 71(1):391–414, apr 2020.

[44] Malene Ringkjøbing Jensen, Markus Zweckstetter, Jie-rong Huang, and Martin Blackledge. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chemical Reviews*, 114(13):6632–6660, jul 2014.

[45] Song-Ho Chong, Prathit Chatterjee, and Sihyun Ham. Computer Simulations of Intrinsically Disordered Proteins. *Annual review of physical chemistry*, 68:117–134, 2017.

[46] Christopher M. Baker and Robert B. Best. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(3):182–198, may 2014.

[47] Payel Das, Silvina Matysiak, and Jeetain Mittal. Looking at the Disordered Proteins through the Computational Microscope. *ACS Central Science*, 4(5):534–542, may 2018.

[48] Emiliano Brini, Carlos Simmerling, and Ken Dill. Protein storytelling through physics. *Science*, 370(6520):eaaz3041, nov 2020.

[49] Carsten Kutzner, Szilárd Páll, Martin Fechner, Ansgar Esztermann, Bert L. Groot, and Helmut Grubmüller. More bang for your buck: Improved use of GPU nodes for GROMACS 2018. *Journal of Computational Chemistry*, 40(27):2418–2431, oct 2019.

[50] Junjie Zou, Carlos Simmerling, and Daniel P. Raleigh. Dissecting the Energetics of Intrinsically Disordered Proteins via a Hybrid Experimental and Computational Approach. *The Journal of Physical Chemistry B*, 123(49):10394–10402, dec 2019.

[51] Sarah Rauscher, Vytautas Gapsys, Michal J. Gajda, Markus Zweckstetter, Bert L. de Groot, and Helmut Grubmüller. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *Journal of Chemical Theory and Computation*, 11(11):5513–5524, nov 2015.

[52] Stefano Piana, Kresten Lindorff-Larsen, and David E. Shaw. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal*, 100(9):L47–L49, may 2011.

[53] Elena Papaleo, Carlo Camilloni, Kaare Teilum, Michele Vendruscolo, and Kresten Lindorff-Larsen. Molecular dynamics ensemble refinement of the heterogeneous native state of NCBD using chemical shifts and NOEs. *PeerJ*, 6:e5125, 2018.

[54] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73, jan 2017.

[55] Stefano Piana, Alexander G. Donchev, Paul Robustelli, and David E. Shaw. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B*, 119(16):5113–5123, apr 2015.

[56] Paul Robustelli, Stefano Piana, and David E. Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21):E4758–E4766, may 2018.

[57] C Clementi, H Nymeyer, and J N Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *Journal of molecular biology*, 298(5):937–53, may 2000.

[58] Xiakun Chu and Victor Muñoz. Roles of conformational disorder and downhill folding in modulating protein–DNA recognition. *Physical Chemistry Chemical Physics*, 19(42):28527–28539, 2017.

[59] Adrian Gustavo Turjanski, J. Silvio Gutkind, Robert B. Best, and Gerhard Hummer. Binding-Induced Folding of a Natively Unstructured Transcription Factor. *PLoS Computational Biology*, 4(4):e1000060, apr 2008.

[60] Paul Robustelli, Stefano Piana, and David E. Shaw. Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein. *Journal of the American Chemical Society*, 142(25):11092–11101, jun 2020.

[61] Elin Karlsson, Cristina Paissoni, Amanda M. Erkelens, Zeinab A. Tehranizadeh, Frieda A. Sorgenfrei, Eva Andersson, Weihua Ye, Carlo Camilloni, and Per Jemth. Mapping the transition state for a binding reaction between ancient intrinsically disordered proteins. *Journal of Biological Chemistry*, 295(51):17698–17712, dec 2020.

[62] M. Kjaergaard, K. Teilum, and F.M. Poulsen. Conformational selection in the molten globule state of the nuclear coactivator binding domain of cbp. *Proc Natl Acad Sci U S A*, 2010;107(28):12535-40. PubMed PMID: 20616042; PMCID: PMC2906600.

[63] M. Knott and R.B. Best. A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: evidence from molecular simulations. *PLoS Comput Biol*, 2012;8(7):e1002605. PubMed PMID: 22829760; PMCID: PMC3400577.

[64] D. Ganguly, W. Zhang, and J. Chen. Electrostatically accelerated encounter and folding for facile recognition of intrinsically disordered proteins. *PLoS Comput Biol*, 2013;9(11):e1003363. PubMed PMID: 24278008; PMCID: PMC3836701.

[65] J.Y. Kim, F. Meng, J. Yoo, and H.S. Chung. Diffusion-limited association of disordered protein by non-native electrostatic interactions. *Nat Commun*, 2018;9(1):4707. PubMed PMID: 30413699; PMCID: PMC6226484.

[66] S. J. Demarest. Packing, specificity, and mutability at the binding interface between the p160 coactivator and CREB-binding protein. *Protein Science*, 13(1):203–210, jan 2004.

[67] A. Vallee-Belisle and K. W. Plaxco. Structure-switching biosensors: inspired by nature. *Curr. Opin. Struct. Biol.*, 20(4):518–26, 2010.

[68] Kevin W. Plaxco and H. Tom Soh. Switch-based biosensors: a new approach towards real-time, in vivo molecular detection. *Trends Biotechnol.*, 29(1):1–5, 2011.

[69] Jeung-Hoi Ha and Stewart N. Loh. Protein conformational switches: from nature to design. *Chemistry*, 18(26):7984–7999, 2012.

[70] Alexis Vallée-Bélisle, Francesco Ricci, and Kevin W. Plaxco. Thermodynamic basis for the optimization of binding-induced biomolecular switches and structure-switching biosensors. *Proc. Natl. Acad. Sci. USA*, 106(33):13802, 2009.

[71] D. De Sancho and V. Muñoz. Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys. Chem. Chem. Phys.*, 13:17030–17043, 2011.

[72] V. Muñoz. Conformational dynamics and ensembles in protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 36:395–412, 2007.

[73] Maria M. Garcia-Mira, Mourad Sadqi, Niels Fischer, Jose M. Sanchez-Ruiz, and Victor Muñoz. Experimental identification of downhill protein folding. *Science*, 298(5601):2191, 2002.

[74] M. Sadqi, D. Fushman, and V. Muñoz. Atom-by-atom analysis of global downhill protein folding. *Nature*, 442(7100):317–21, 2006.

[75] Athi N. Naganathan and V. Muñoz. Thermodynamics of downhill folding: Multi-probe analysis of pdd, a protein that folds over a marginal free energy barrier. *J. Phys. Chem. B*, 118(30):8982–8994, 2014.

[76] Victor Muñoz, Luis A. Campos, and Mourad Sadqi. Limited cooperativity in protein folding. *Curr. Opin. Struct. Biol.*, 36:58–66, 2016.

[77] P. Li, F. Y. Oliva, A. N. Naganathan, and V. Muñoz. Dynamics of one-state downhill protein folding. *Proc. Natl. Acad. Sci. USA*, 106:103–108., 2009.

[78] L. Sborgi, A. Verma, V. Munoz, and E. de Alba. Revisiting the nmr structure of the ultrafast downhill folding protein gpw from bacteriophage lambda. *PLoS One*, 6(11):e26409, 2011.

[79] Lorenzo Sborgi, Abhinav Verma, Stefano Piana, Kresten Lindorff-Larsen, Michele Cerminara, Clara M. Santiveri, David E. Shaw, Eva de Alba, and Victor Muñoz. Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations. *J. Am. Chem. Soc.*, 137(20):6506–6516, 2015.

[80] A. Fung, P. Li, R. Godoy-Ruiz, J. M. Sanchez-Ruiz, and V. Munoz. Expanding the realm of ultrafast protein folding: gpw, a midsize natural single-domain with alpha+beta topology that folds downhill. *J. Am. Chem. Soc.*, 130(23):7489–95, 2008.

[81] Katharine A. White, Diego Garrido Ruiz, Zachary A. Szpiech, Nicolas B. Strauli, Ryan D. Hernandez, Matthew P. Jacobson, and Diane L. Barber. Cancer-associated arginine-to-histidine mutations confer a gain in ph sensing to mutant proteins. *Sci. Signal.*, 10(495):eaam9931, 2017.

[82] Scott E. Boyken, Mark A. Benhaim, Florian Busch, Mengxuan Jia, Matthew J. Bick, Heejun Choi, Jason C. Klima, Zibo Chen, Carl Walkey, Alexander Mileant, Aniruddha Sahasrabuddhe, Kathy Y. Wei, Edgar A. Hodge, Sarah Byron, Alfredo Quijano-Rubio, Banumathi Sankaran, Neil P. King, Jennifer Lippincott-Schwartz, Vicki H. Wysocki, Kelly K. Lee, and David Baker. De novo design of tunable, ph-driven conformational changes. *Science*, 364(6441):658, 2019.

[83] Özkan Yildiz, Kutti R Vinothkumar, Panchali Goswami, and Werner Kühlbrandt. Structure of the monomeric outer-membrane porin ompg in the open and closed conformation. *EMBO J.*, 25(15):3702–3713, 2006.

[84] Stephen P. Edgcomb and Kenneth P. Murphy. Variability in the pka of histidine side-chains correlates with burial within proteins. *Proteins: Struct., Funct., Bioinf.*, 49(1):1–6, 2002.

[85] J. E. Crooks, F. Hibbert, and A. V. Willi. *Proton Transfer*. Comprehensive Chemical Kinetics. Elsevier, Amsterdam, 1977.

[86] Victor Munoz and Michele Cerminara. When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. *Biochem. J.*, 473(17):2545–2559, 2016.

[87] Douglas E. V. Pires, David B. Ascher, and Tom L. Blundell. Duet: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nuc. Acid. Res.*, 42(W1):W314–W319, 2014.

[88] Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E. Felberg, David H. Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W. Gohara, Todd Dolinsky, Robert Konecny, David R. Koes, Jens Erik Nielsen, Teresa Head-Gordon, Weihua Geng, Robert Krasny, Guo-Wei Wei, Michael J. Holst, J. Andrew McCammon, and Nathan A. Baker. Improvements to the apbs biomolecular solvation software suite. *Prot. Sci.*, 27(1):112–128, 2018.

[89] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. Gromacs: Fast, flexible, and free. *J. Comp. Chem.*, 26(16):1701–1718, 2005.

[90] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[91] J. L. F. Abascal and C. Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *J. Chem. Phys.*, 123(23):234505, 2005.

[92] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[93] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18(12):1463–1472, 1997.

[94] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[95] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. App. Phys.*, 52(12):7182–7190, 1981.

[96] Sidhartha Chaudhury, Daniel R. Ripoll, and Anders Wallqvist. Structure-based pka prediction provides a thermodynamic basis for the role of histidines in ph-induced conformational transitions in dengue virus. *Biochem. Biophys. Rep.*, 4:375–385, 2015.

[97] J. R. Casey, S. Grinstein, and J. Orlowski. Sensors and regulators of intracellular ph. *Nat. Rev. Mol. Cell Biol.*, 11(1):50–61, 2010.

[98] Jyoti Srivastava, Diane L Barber, and Matthew P Jacobson. Intracellular ph sensors: design principles and functional significance. *Physiology*, 2007.

[99] C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[100] P Tompa, N E Davey, T J Gibson, and M M Babu. A million peptide motifs for the molecular biologist. *Mol Cell*, 55(2):161–169, 2014.

[101] Maria M Garcia-Mira, Mourad Sadqi, Niels Fischer, Jose M Sanchez-Ruiz, and Victor Muñoz. Experimental Identification of Downhill Protein Folding. *Science*, 298(5601):2191, 2002.

[102] Vladimir N Uversky and A Keith Dunker. Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(6):1231–1264, 2010.

[103] P Csermely, R Palotai, and R Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci*, 35(10):539–546, 2010.

[104] Joachim Lätzer, Garegin A Papoian, Michael C Prentiss, Elizabeth A Komives, and Peter G Wolynes. Induced Fit, Folding, and Recognition of the NF-κB-Nuclear Localization Signals by IκBα and IκBβ. *Journal of Molecular Biology*, 367(1):262–274, 2007.

[105] Joseph M. Rogers, Vladimiras Oleinikovas, Sarah L. Shammas, Chi T. Wong, David De Sancho, Christopher M. Baker, and Jane Clarke. Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[106] Jakob Dogan, Stefano Gianni, and Per Jemth. The binding mechanisms of intrinsically disordered proteins. *Physical Chemistry Chemical Physics*, 16(14):6323–6331, 2014.

[107] Huan-Xiang Zhou. From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophysical journal*, 98(6):L15–L17, 2010.

[108] M R Jensen, R W Ruigrok, and M Blackledge. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr Opin Struct Biol*, 23(3):426–435, 2013.

[109] Alan R Fersht and Luis Serrano. Principles of protein stability derived from protein engineering experiments. *Current Opinion in Structural Biology*, 3(1):75–83, 1993.

[110] Athi N Naganathan and Victor Muñoz. Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proceedings of the National Academy of Sciences*, 107(19):8611, 2010.

[111] C Tanford. Protein denaturation. *Adv Protein Chem*, 23:121–282, 1968.

[112] H Jane Dyson and Peter E Wright. Peptide conformation and protein folding. *Current Opinion in Structural Biology*, 3(1):60–65, 1993.

[113] José L Neira, Laura S Itzhaki, Daniel E Otzen, Ben Davis, and Alan R Fersht. Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis11Edited by J.Karn. *Journal of Molecular Biology*, 270(1):99–110, 1997.

[114] Richard J Lindsay, Rachael A Mansbach, S Gnanakaran, and Tongye Shen. Effects of pH on an IDP conformational ensemble explored by molecular dynamics simulation. *Biophysical Chemistry*, 271:106552, 2021.

[115] W. Zhang, D. Ganguly, and J. Chen. Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. *PLoS Comput Biol*, 2012;8(1):e1002353. PubMed PMID: 22253588; PMCID: PMC3257294.

[116] F. Zosel, D. Mercadante, D. Nettels, and B. Schuler. A proline switch explains kinetic heterogeneity in a coupled folding and binding reaction. *Nat Commun*, 2018;9(1):3332. PubMed PMID: 30127362; PMCID: PMC6102232.

[117] Robert B. Best. Computational and theoretical advances in studies of intrinsically disordered proteins, 2017.

[118] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[119] Szilárd Páll, Mark James Abraham, Carsten Kutzner, Berk Hess, and Erik Lindahl. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. In Stefano Markidis and Erwin Laure, editors, *Solving Software Challenges for Exascale*, pages 3–27. Springer International Publishing.

[120] H Jane Dyson, Mark Rance, Richard A Houghten, Peter E Wright, and Richard A Lerner. Folding of immunogenic peptide fragments of proteins in water solution: II. The nascent helix. *Journal of Molecular Biology*, 201(1):201–217, 1988.

[121] Berk Hess, Henk Bekker, Herman J C Berendsen, and Johannes G E M Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18(12):1463–1472, 1997.

[122] H J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[123] Daniela Bonetti, Francesca Troilo, Angelo Toto, Maurizio Brunori, Sonia Longhi, and Stefano Gianni. Analyzing the Folding and Binding Steps of an Intrinsically Disordered Protein by Protein Engineering. *Biochemistry*, 56(29):3780–3786, 2017.

[124] Victor Muñoz and Luis Serrano. Elucidating the folding problem of helical peptides using empirical parameters. *Nature Structural Biology*, 1(6):399–409, 1994.

[125] M Buck. Trifluoroethanol and colleagues: cosolvents come of age. Recent studies with peptides and proteins. *Q Rev Biophys*, 31(3):297–355, 1998.

[126] Shneior Lifson and A Roig. On the Theory of Helix—Coil Transition in Polypeptides. *The Journal of Chemical Physics*, 34(6):1963–1974, 1961.

[127] B H Zimm and J K Bragg. Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. *The Journal of Chemical Physics*, 31(2):526–535, 1959.

[128] Urmi Doshi and Victor Muñoz. Kinetics of α-helix formation as diffusion on a one-dimensional free energy surface. *Chemical Physics*, 307(2):129–136, 2004.

[129] Yee-Hsiung Chen, Jen Tsi Yang, and Kue Hung Chau. Determination of the helix and β form of proteins in aqueous solution by circular dichroism. *Biochemistry*, 13(16):3350–3359, 1974.

[130] John P Hennessey and W Curtis Johnson. Information content in the circular dichroism of proteins. *Biochemistry*, 20(5):1085–1094, 1981.

[131] Kresten Lindorff-Larsen, Nikola Trbovic, Paul Maragakis, Stefano Piana, and David E. Shaw. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society*, 2012.

[132] J M Scholtz and R L Baldwin. The mechanism of alpha-helix formation by peptides. *Annu Rev Biophys Biomol Struct*, 21:95–118, 1992.

[133] L Sborgi, A Verma, S Piana, K Lindorff-Larsen, M Cerminara, C M Santiveri, D E Shaw, E de Alba, and V Muñoz. Interaction Networks in Protein Folding via Atomic-Resolution Experiments and Long-Time-Scale Molecular Dynamics Simulations. *J Am Chem Soc*, 137(20):6506–6516, 2015.

[134] Robert B Best and Gerhard Hummer. Optimized Molecular Dynamics Force Fields Applied to the HelixCoil Transition of Polypeptides. *The Journal of Physical Chemistry B*, 113(26):9004–9015, 2009.

[135] D De Sancho and V Muñoz. Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys. Chem. Chem. Phys.*, 13:17030–17043, 2011.

[136] V Ozenne, F Bauer, L Salmon, J R Huang, M R Jensen, S Segard, P Bernadó, C Charavay, and M Blackledge. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11):1463–1470, 2012.

[137] David H Brookes and Teresa Head-Gordon. Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *Journal of the American Chemical Society*, 138(13):4530–4538, 2016.

[138] Y. He, S. Nagpal, M. Sadqi, E. De Alba, and V. Munõz. Glutton: A tool for generating structural ensembles of partly disordered proteins from chemical shifts. *Bioinformatics*, 35(7), 2019.

[139] M.O. Ebert, S.H. Bae, H.J. Dyson, and P.E. Wright. Nmr relaxation study of the complex formed between cbp and the activation domain of the nuclear hormone receptor coactivator actr. *Biochemistry*, 2008;47(5):1299-308. PubMed PMID: 18177052.

[140] Jakob Dogan, Josefin Jonasson, Eva Andersson, and Per Jemth. Binding Rate Constants Reveal Distinct Features of Disordered Protein Domains. *Biochemistry*, 54(30):4741–4750, aug 2015.

[141] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12(1):7–8, jan 2015.

[142] John D. Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics, 2014.

[143] Brooke E. Husic and Vijay S. Pande. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, 140(7):2386–2396, feb 2018.

[144] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature chemistry*, 9(10):1005–1011, 2017.

[145] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A Software

Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11(11):5525–5542, nov 2015.

[146] Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, sep 2010.

[147] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, 139(1):015102, jul 2013.

[148] Frank Noé, Hao Wu, Jan Hendrik Prinz, and Nuria Plattner. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *Journal of Chemical Physics*, 2013.

[149] Alan Grossfield and Daniel M. Zuckerman. Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. pages 23–48. 2009.

[150] Alan Grossfield, Paul N. Patrone, Daniel R. Roe, Andrew J. Schultz, Daniel Siderius, and Daniel M. Zuckerman. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*, 1(1), 2019.

[151] Athi N. Naganathan, Urmi Doshi, and Victor Muñoz. Protein Folding Kinetics: Barrier Effects in Chemical and Thermal Denaturation Experiments. *Journal of the American Chemical Society*, 129(17):5673–5682, may 2007.

[152] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How Fast-Folding Proteins Fold. *Science*, 334(6055):517–520, oct 2011.

[153] Jae-Yeol Kim and Hoi Sung Chung. Disordered proteins follow diverse transition paths as they fold and bind to a partner. *Science*, 368(6496):1253–1257, jun 2020.

[154] D. Ganguly, W. Zhang, and J. Chen. Synergistic folding of two intrinsically disordered proteins: searching for conformational selection. *Mol Biosyst*, 2012;8(1):198-209. PubMed PMID: 21766125.

[155] F. Sturzenegger, F. Zosel, E.D. Holmstrom, K.J. Buholzer, D.E. Makarov, D. Nettels, and B. Schuler. Transition path times of coupled folding and binding reveal the formation of an encounter complex. *Nat Commun*, 2018;9(1):4708. PubMed PMID: 30413694; PMCID: PMC6226497.

[156] J. Dogan, A. Toto, E. Andersson, S. Gianni, and P. Jemth. Activation barrier-limited folding and conformational sampling of a dynamic protein domain. *Biochemistry*, 2016;55(37):5289-95:27542287.

[157] V. Iesmantavicius, J. Dogan, P. Jemth, K. Teilum, and M. Kjaergaard. Helical propensity in an intrinsically disordered protein accelerates ligand binding. *Angew Chem Int Ed Engl*, 2014;53(6):1548-51. PubMed PMID: 24449148.

[158] X. Liu, J. Chen, and J. Chen. Residual structure accelerates binding of intrinsically disordered actr by promoting efficient folding upon encounter. *J Mol Biol*, 2019;431(2):422-32. PubMed PMID: 30528464; PMCID: PMC6687458.

[159] M. Knott and R.B. Best. Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *J Chem Phys*, 2014;140(17):175102. PubMed PMID: 24811666; PMCID: PMC4032430.

[160] Q. Yu, W. Ye, W. Wang, and H.F. Chen. Global conformational selection and local induced fit for the recognition between intrinsic disordered p53 and cbp. *PLoS One*, 2013;8(3):e59627. PubMed PMID: 23555731; PMCID: PMC3608666.