Multi-scale Simulations of Dynamic Protein Structures and Interactions

Yumeng Zhang
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Biochemistry, Biophysics, and Structural Biology Commons, and the Computational Chemistry Commons

Recommended Citation
Zhang, Yumeng, "Multi-scale Simulations of Dynamic Protein Structures and Interactions" (2024). Doctoral Dissertations. 3104.
https://doi.org/10.7275/36463575 https://scholarworks.umass.edu/dissertations_2/3104

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Multi-scale Simulations of Dynamic Protein Structures and Interactions

A Dissertation Presented

by

Yumeng Zhang

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2024

Department of Chemistry
Multi-scale Simulations of Dynamic Protein Structures and Interactions

A Dissertation Presented

by

YUMENG ZHANG

Approved as to style and content by:

__________________________
Jianhan Chen, Chair

__________________________
Scott Auerbach, Member

__________________________
Min Chen, Member

__________________________
Gregory Grason, Member

Ricardo Metz, Department Head
Department of Chemistry
DEDICATION

To my parents, advisor, and other people who light the way forward for me in my academic journey, and to my cats whose warmth and companionship have been the constant source of comfort.
ACKNOWLEDGMENTS

As I reflect on the relatively short yet profoundly transformative journey that has brought me to my graduate studies at UMass Amherst, I find myself overwhelmed with gratitude and nostalgia. From the very first day I set foot on campus, walked the unfamiliar roads, plucked my first apple, met incredible people, to the myriad of experiences we collectively traversed, each moment etches an indelible mark in my memory.

At this juncture, I want to extend my deepest thanks to you, Professor Jianhan Chen, my research advisor. Your decision to welcome me into the research group was the initiation of the delightful and enjoyable exploration into the realm of computational modeling. Your boundless passion for research not only ignited my enthusiasm for the intricate world of biochemistry but also fostered a deep love for the research I have been privileged to work with. Your sagacity consistently guided me along the right path, and your unwavering patience and encouragement provided the fortitude to rise resiliently from failures and moments of uncertainty. Moreover, I am deeply grateful for your mentoring not only in academics but also in life. Your kind, generous, and continuous support, coupled with your warmth, has been the guiding light that directly lifted me from the depths of my lowest energy minima.

Meanwhile, I want to express my heartfelt appreciation to my dissertation committee members, Professor Scott Auerbach, Professor Min Chen, and Professor Greg Grason, for your dedicated assistance and encouragement throughout my research journey. A special gratitude goes to Professor Lynmarie Thompson, whose mentoring has played an integral role in my personal and academic growth.

I am equally indebted to my exceptional lab mates, who have contributed to fostering a conducive learning environment. Dr. Xiaorong Liu, your tireless efforts to guide me through my
research, help me resolve scientific problems, provide detailed tutorials, and carefully revise my manuscripts have been invaluable. Your unwavering support extends to my life, creating a life filled with warmth and care, resulting in unforgettable memories. Many thanks to Dr. Mahdieh Yazdani, who not only taught me how to navigate the Ph.D. career but also provided invaluable assistance during the 2022 internship application process. Dr. Zhiguang Jia, your countless efforts in helping me compile the computer, addressing computational challenges, and engaging in research discussions with you have benefited me so much. Dr. Shanlong Li, the collaboration with you in exploring research is truly unforgettable. Your knowledge and dedication were instrumental in completing our exciting project within one year. Working with you has been a rewarding experience. Dr. Erik Nordquist, your generous assistance in both my research and personal life has been invaluable. Our discussions about research and science have always been enlightening, and your encouragement during my moments of self-doubt has been greatly appreciated. I wish you a bright future and success in achieving your dreams in your new position. Dr. Xiping Gong, I want to express my heartfelt gratitude for your selfless assistance. Your guidance has played a pivotal role in refining my coding skills, patiently leading me through each step of the process. Your expertise has proven invaluable in resolving countless bugs throughout my Ph.D., and your adept problem-solving, encompassing both scientific and technical aspects, has been crucial in overcoming research challenges. Furthermore, during the most challenging times of my Ph.D., your support, alongside Professor Chen, was a beacon that helped me navigate difficulties and persist. I am deeply thankful to both of you for propelling me to this stage in my academic journey. Special thanks to Qianlan Jia, whose timely assistance was a bonfire during my moments of greatest need. His willingness to dedicate time to help me overcome obstacles has been truly appreciated. I am also grateful to Jian Huang, an intelligent, patient, and unique person, whose
presence has been like a shining moon, providing warmth and illuminating the path during dark nights. Heartfelt thanks to Samantha Schultz for the unforgettable moments we shared in the lab, at the BPS conference, and during our party time filled with laughter and camaraderie. I express my gratitude to Kairong Dong and Shrishti Barethiya, whose infectious smiles have been powerful and encouraging throughout our time together, and the cherished memories of fun outings. Many thanks to David Docoeur, for your extravagant feasts and delicious cakes. And sincere thanks to Juni Campbell, Aron Korsunsky, and Ryan Pham for your generous, kind, and unwavering support in the lab.

I also want to express my heartfelt appreciation to my collaborators, Spencer Shorkey, Jacqueline Sharp, and Kaitlyn Gilliam, for their invaluable contributions in delving into the fascinating realm of nanopore tweezer and proteases research. Their collaborative spirit and dedication have enriched my exploration in this captivating field.

I am grateful to my cohort and friends in the Chemistry Department, who have played an essential role in making my graduate life at UMass Chemistry enjoyable. Special thanks to Yi Du and Ruolan Fan, who have accompanied me, creating cherished memories of shared laughter and camaraderie.

Many thanks to my friends Shuang Wan, Lianjiao Zhou, and Han Wang, as well as Wei Zhang, Xinman Liu, Haochen Zhang, Zitian Zhang, and Chuanping Shi. Your encouragement throughout my Ph.D. career has been a source of strength and motivation, and I am grateful for the support and friendships we had.

My special appreciations go to my beloved cats, faithful companions throughout my entire Ph.D. journey. Their comforting presence and unique ways of expressing warmth have been a source of solace and joy, turning my home into a haven every day.
Last but certainly not least, I want to express my deep gratitude to my family, whose robust support has been a pillar of strength throughout my academic journey. To my family in China, your constant encouragement and mental support have motivated me to forge ahead in the face of challenges. A special and heartfelt acknowledgment goes to my mom, who, even amidst a global pandemic, traveled to the U.S. to provide invaluable assistance. Her mere presence lifted me from moments of sadness, and her culinary skills brought comfort through the creation of delicious meals. Her enduring optimism and the warm glow of her welcoming light have created a haven, always patiently waiting to embrace me upon my return. The steadfast support and encouragement from all my family members have been a source of inspiration, urging me to navigate my academic path with resilience and dedication.
ABSTRACT

Multi-scale Simulations of Dynamic Protein Structures and Interactions

FEBRUARY 2024

YUMENG ZHANG,
B.S., WUHAN UNIVERSITY, CHINA
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jianhan Chen

Intrinsically disordered proteins (IDPs) are functional proteins that lack stable tertiary structures in the unbound state. They frequently remain dynamic even within specific complexes and assemblies. IDPs are major components of cellular regulatory networks and have been associated with cancers, diabetes, neurodegenerative diseases, and other human diseases. Computer simulations are essential for deriving a molecular description of the disordered protein ensembles and dynamic interactions for mechanistic understanding of IDPs in biology, diseases, and therapeutics. However, accurate simulation of the heterogeneous ensembles and dynamic interactions of IDPs is extremely challenging because of both the prohibitive computational cost and demanding force field accuracy. In this dissertation, we developed a set of enhanced sampling and multi-scale simulation methods to overcome these limitations, and successfully applied them to study the structure, interaction and phase separation of IDPs.

We have first applied the state-of-the-art explicit solvent atomistic simulations to study the inhibitory mechanism of the disordered N-terminal domain of *Staphylococcal* peroxidase inhibitor (SPIN). We performed high-temperature simulations to study the coupled binding and folding process during SPIN inhibition of the host myeloperoxidase (MPO) enzyme. The results showed
that differences in dynamics may provide a physical basis of the ability of different SPIN homologs to inhibit innate immunity.

Recognizing the need for enhanced sampling methods for IDP simulation, we have developed a new replica-exchange with solute tempering (REST) protocol to achieve more efficient explicit solvent sampling of disordered protein ensembles. We proposed that the scaling of protein-water interactions in REST is a free parameter that could be optimized to better control how the protein conformational properties (e.g., chain expansion) at different effective temperatures and achieve more effective sampling. Specifically, we developed a REST3 protocol that rebalances the protein-protein and protein-water interactions and eliminates the unanticipated chain collapse at high temperature conditions in the previous REST2 protocol. Application to model IDPs demonstrated that REST3 prevented the conformational segregation during exchanges, leading to an effective temperature random walk across all conditions and accelerating the simulation of the protein conformational space.

Even with enhanced sampling, accurate description of disordered conformations at atomistic level remains extremely challenging for complex IDPs. Alternatively, coarse-grained simulations can provide an effective strategy for overcoming the length- and time-scale limitations. Here, we refined a hybrid-resolution coarse-grained model (HyRes) for accurate simulation of disordered protein ensembles and dynamic protein interactions. HyRes contains atomistic backbone and coarse-grained sidechain beads, to provide semi-quantitative description of residual secondary structures and long-range interactions. Specifically, we introduced a surface area-based implicit solvation energy term, and iteratively re-optimized the effective interaction strength potentials. The new model, referred as HyRes II, provides near quantitative descriptions of IDP long-range non-specific interactions and secondary structures, at a level comparable to the latest
atomistic protein force fields. Applied to the disordered N-terminal transactivation domain (TAD) of tumor suppressor p53, HyRes II faithfully recapitulates various nontrivial structural properties to a level of accuracy that is comparable to a99SB-disp, one of the best atomistic protein force fields. Moreover, we demonstrate HyRes II’s efficacy in accurately simulating the dynamic interaction between TAD and the DNA-binding domain of p53, generating structural ensembles that align closely with existing NMR data.

Encouraged by successes of HyRes II for probing dynamic interactions of IDPs, we further investigated its suitability for simulating IDP-mediated phase separation, which underlies the formation of biomolecular condensates and has attracted intensive interests. Compared to the popular single-bead models, HyRes has the potential to describe backbone-mediated interactions and capture the role of residual structures in phase separation. Reimplemented on GPU, our simulations showed that HyRes is efficient enough to directly simulate the spontaneous phase separation of IDPs and at the time balanced enough to capture the effects of mutational and structural perturbations. For peptide GY-23, HyRes simulations reveal increased $\beta$-structures in condensates, which are consistent with experimental observations. For the conserved region (CR) of TDP-43, HyRes simulations successfully recapitulate the apparent correlation between helical propensities and condensate stability. In depth analysis, however, revealed that residual helices did not directly mediate interpeptide interactions to stabilize the condensed phase. Instead, it is the balance between backbone and sidechain-mediated interactions, as modulated by residual structures, that actually determines phase separation propensity.

Finally, we have applied the HyRes II model to study the dynamic interaction of West Nile virus (WNV) NS2B/NS3 proteases with the ClyA protein nanopore. Nanopore tweezers provide a powerful approach for label-free detection of protein dynamics at the single-molecule level, by
capturing the protein analyte in the lumen of the nanopore. From the steered-MD and standard MD simulations, we discovered that the protease could bind dynamically to a middle region of the ClyA nanopore, mediated mainly by electrostatically interactions. In particular, we identified a key Glu residue within the ClyA lumen, mutation of which to Ala or Lys could further stabilize the protease/nanopore interaction. This led to the design a modified ClyA nanopore tweezer that can stably capture the protease and resolve the dynamics between NS2B/NS3 open and closed conformations.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>V</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>IX</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>XIII</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>XVI</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>XVII</td>
</tr>
<tr>
<td>CHAPTER 1</td>
<td>1</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Intrinsically disordered proteins and their functional activities</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Coupled binding and folding of IDPs</td>
<td>3</td>
</tr>
<tr>
<td>1.1.2 Fuzzy complexes of IDPs</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Multi-scale molecular dynamic simulations in studying IDPs</td>
<td>7</td>
</tr>
<tr>
<td>1.2.1 Atomistic simulation of IDP structure, dynamics, and interactions</td>
<td>8</td>
</tr>
<tr>
<td>1.2.2 Enhanced sampling methods for efficient simulation of IDPs</td>
<td>10</td>
</tr>
<tr>
<td>1.2.3 Coarse-grained simulations in studying long timescale IDP dynamics and interactions</td>
<td>13</td>
</tr>
<tr>
<td>1.3 Dissertation outline</td>
<td>15</td>
</tr>
<tr>
<td>CHAPTER 2</td>
<td>17</td>
</tr>
<tr>
<td>STRUCTURAL STABILITY OF DISORDERED SPIN N-TERMINAL REGION IN MYELOPEROXIDASE INHIBITION</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>18</td>
</tr>
<tr>
<td>2.2 Methods</td>
<td>44</td>
</tr>
<tr>
<td>2.2.1 High Temperature Simulations of SPIN/MPO Complexes</td>
<td>22</td>
</tr>
<tr>
<td>2.2.2 Room Temperature Simulations of Free NTDs</td>
<td>23</td>
</tr>
<tr>
<td>2.2.3 Analysis</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Results and Discussion</td>
<td>25</td>
</tr>
<tr>
<td>2.3.1 High Temperature Simulations Reveal Step-wise Binding of SPIN NTD and CTD</td>
<td>25</td>
</tr>
<tr>
<td>2.3.2 Cooperative Binding and Folding of SPIN-aureus NTD</td>
<td>26</td>
</tr>
<tr>
<td>2.3.3 Conformational Selection-like Mechanism for SPIN-delphini NTD</td>
<td>29</td>
</tr>
<tr>
<td>2.3.4 Elevated Pre-folding in Unbound SPIN-delphini NTD</td>
<td>31</td>
</tr>
<tr>
<td>2.4 Conclusions</td>
<td>32</td>
</tr>
<tr>
<td>2.5 Supporting Material</td>
<td>34</td>
</tr>
<tr>
<td>CHAPTER 3</td>
<td>40</td>
</tr>
<tr>
<td>RE-BALANCING REPLICA EXCHANGE WITH SOLUTE TEMPERING FOR SAMPLING SYANMIC PROTEIN CONFORMATIONS</td>
<td>40</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>41</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>44</td>
</tr>
</tbody>
</table>
3.2.1 REST Algorithm ................................................................. 44
3.2.2 System setup and simulation protocols ................................. 45
3.2.3 Analysis ............................................................................. 47

3.3 Results and Discussion ......................................................... 48
3.3.1 Severe over-compaction at high temperatures in REST2 ....... 48
3.3.2 Over-compaction at high temperatures can lead to replica segregation in REST2 ................................................................. 50
3.3.3 Re-balancing the solute-solvent interactions in REST3 .......... 52
3.3.4 REST3 eliminates conformational trapping at high temperatures and replica segregation ................................................................. 55
3.3.5 Conformational sampling and convergence of REST2 and REST3 ... 57
3.3.6 REST3 can further reduce the required number of replicas ....... 62

3.4 Conclusions ............................................................................ 63
3.5 Supporting Material .................................................................. 66

CHAPTER 4 ................................................................................................. 76
TOWARDS ACCURATE COARSE-GRAINED SIMULATIONS OF DISORDERED PROTEINS AND THEIR DYNAMIC INTERACTIONS ................................................................. 76
4.1 Introduction ............................................................................. 77
4.2 Methods ................................................................................... 81
4.2.1 Original HyRes Protein Model ............................................... 81
4.2.2 HyRes II: design and optimization ........................................ 83
4.2.3 Folded and disordered proteins for training and testing ..... 84
4.2.4 CG simulation protocols ....................................................... 85
4.2.5 Atomistic simulations ......................................................... 86
4.2.6 Analysis ............................................................................. 88
4.3 Results and Discussion ............................................................. 89
4.3.1 Balance of IDP chain dimensions and secondary structures .... 89
4.3.2 Balance between short-range and electrostatic interactions .... 93
4.3.3 Ability of HyRes II in describing nontrivial local and long-range structures of p53 TAD ................................................................. 95
4.3.4 Ability of HyRes II to distinguish compactness of structured and disordered proteins ....................................................... 98
4.3.5 Dynamic interactions of p53 TAD/DBD .................................. 99
4.4 Conclusions ............................................................................ 104
4.5 Supporting Material ............................................................... 106

CHAPTER 5 ................................................................................................. 117
TOWARDS ACCURATE SIMULATION OF COUPLING BETWEEN PROTEIN SECONDARY STRUCTURE AND PHASE SEPARATION ................................................................. 117
5.1 Introduction ............................................................................. 118
5.2 Methods ................................................................................... 124
5.2.1 HyRes simulations of IDP monomers .................................... 124
5.2.2 Simulation of phase separation and equilibrium .................. 125
5.2.3 Phase diagram and properties of the condensates ............... 126
5.3 Results and Discussion ............................................................. 126
5.3.1 GPU acceleration of the HyRes protein model ................. 126
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1 Native intermolecular contacts between SPIN and MPO</td>
<td>34</td>
</tr>
<tr>
<td>Table 2.2 Native intramolecular contacts within NTDs of SPIN-<em>aureus</em> and SPIN-<em>delphini</em></td>
<td>35</td>
</tr>
<tr>
<td>Table 2.3 Representative dissociation transition segments used in contact probability calculations</td>
<td>36</td>
</tr>
<tr>
<td>Table 3.1 Summary of REST2 and REST3 simulations</td>
<td>66</td>
</tr>
<tr>
<td>Table 3.2 Parameters for REST2 and REST3 protocols</td>
<td>66</td>
</tr>
<tr>
<td>Table 4.1 HyRes II vdW interaction strength parameters for all side chain beads</td>
<td>92</td>
</tr>
<tr>
<td>Table 4.2 Ensemble averaged properties of model peptides calculated from CG and AT simulations</td>
<td>93</td>
</tr>
<tr>
<td>Table 4.3 Ensemble averaged radius of gyration (Rg) of ordered and disordered peptides simulated using HyRes and HyRes II (in Å)</td>
<td>99</td>
</tr>
<tr>
<td>Table 4.4 Sequences of training set, test set peptides used in the work</td>
<td>106</td>
</tr>
<tr>
<td>Table 4.5 Ensemble averaged conformational properties of poly-glycine simulated using HyRes, HyRes II, and ABSINTH</td>
<td>108</td>
</tr>
<tr>
<td>Table 5.1 Lennard-Jones parameters for all backbone atoms in HyRes-GPU</td>
<td>148</td>
</tr>
<tr>
<td>Table 5.2 Sequences of all IDPs studied in the work</td>
<td>149</td>
</tr>
<tr>
<td>Table 5.3 Experimental Rg and the corresponding measurement conditions</td>
<td>151</td>
</tr>
<tr>
<td>Table 5.4 IDPs with available experimental data on residue helicity profiles and the corresponding measurement conditions</td>
<td>152</td>
</tr>
<tr>
<td>Table 5.5 Experimental and simulated C_{sur} (in μM) for TDP-43 CR variants</td>
<td>152</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Figure 1.1</td>
<td>Illustration of disordered protein structures and interactions</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Illustration of coupled binding and binding and co-folding of IDPs</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Representative structure of SPIN/MPO complexes</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Fraction of native SPIN/MPO intermolecular interactions</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Cooperative binding and folding of SPIN-\textit{aureus}</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Cooperative binding and folding of SPIN-\textit{delphini}</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Fraction of native intra-molecular contacts of SPIN-NTDs</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Interactions between SPIN and MPO in truncated system</td>
</tr>
<tr>
<td>Figure 2.7</td>
<td>Time evolution of NTD native contact fractions</td>
</tr>
<tr>
<td>Figure 2.8</td>
<td>Time evolution of NTD native contact fractions</td>
</tr>
<tr>
<td>Figure 2.9</td>
<td>Overlay of the structures of SPIN-\textit{delphini} in complex with MPO</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Overall protein chain expansion of p53-NTD as a function of temperature in REST2 simulations in six explicit solvent protein force fields</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Conformational properties of p53-NTD using unscaled (MD) and scaled Hamiltonians</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Efficiency of replica exchange in REST2 simulations of p53-NTD</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Conformational properties as a function of temperature from REST3 simulations of p53-NTD and KID in a99SB-disp</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Efficiency of replica exchange in REST2 simulations of p53-NTD using 16 or 8 replicas</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>2D probability distributions of the radius of gyration and end-to-end distance of KID at 298 K</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Distribution of conformational ensembles at 298 K generated by REST2 and REST3</td>
</tr>
</tbody>
</table>
Figure 3.8 Evolution of the conformational ensemble of p53-NTD at 298 K ........................................ 62
Figure 3.9 The initial conformations of KID and p53-NTD for folding and control simulations 67
Figure 3.10 Distributions of conformational properties of p53-NTD .................................................. 68
Figure 3.11 Distributions of conformational properties of KID ............................................................ 68
Figure 3.12 Evolution of temperature and radius of gyration for each of the 16 replicas of the
REST2 control simulation of p53-NTD in a99SB-disp ......................................................................... 69
Figure 3.13 Evolution of temperature and radius of gyration for each of the 16 replicas of the
REST2 folding simulation of p53-NTD in a99SB-disp ......................................................................... 70
Figure 3.14 Evolution of temperature and radius of gyration for each of the 16 replicas of the
REST3 folding simulation of p53-NTD in a99SB-disp ......................................................................... 71
Figure 3.15 Efficiency of replica exchange of KID simulations with REST2 and REST3 control
and folding simulations ......................................................................................................................... 72
Figure 3.16 Conformational distributions of KID and p53-NTD from REST2 and REST3
simulations in a99SB-disp .................................................................................................................. 73
Figure 3.17 2D probability distributions of the starting position and length of partial helices of
KID at 298 K as a function of REST2 and REST3 simulation time ..................................................... 74
Figure 3.18 Evolution of temperature and radius of gyration for all replicas of the REST3 folding
simulation of p53-NTD in a99SB-disp using 8 replicas ........................................................................ 75
Figure 3.19 Properties of conformational ensembles as a function of effective temperature
derived from RSET2 and REST3 simulations of p53-NTD (folding simulation) in a99SB-disp 75
Figure 4.1 HyRes representation of amino acids .............................................................................. 81
Figure 4.2 Residue helicity profiles of (AAQAA)$_3$ and KID calculated using HyRes and HyRes
II models ............................................................................................................................................... 92
Figure 4.3 Averaged $R_g$ as a function of the charge patterning parameter $\kappa$ for 30 variants of the
(EK)$_{25}$ peptide .................................................................................................................................. 94
Figure 4.4 Probability distribution p53 TAD from HyRes and HyRes II at 300 K ............................... 95
Figure 4.5 PRE results of p53 TAD from experiments, AT simulations, and CG simulations .... 96
Figure 4.6 PCA analysis of the conformational space sampled by p53 TAD in dynamic
interactions with DBD ....................................................................................................................... 101
Figure 4.7 PRE results of p53 TAD/DBD interactions from experiments and HyRes II model. 103

Figure 4.8 Initial conformations of peptides/proteins used in all simulations ......................... 108

Figure 4.9 Probability distribution of the $R_g$ and end-to-end distance of Gly$_{10}$ and Gly$_{20}$ .... 109

Figure 4.10 Probability distribution of the $R_g$ and end-to-end distance of (AAQAA)$_3$ and KID 110

Figure 4.11 2D histogram of p53 TAD geometric properties based on the $R_g$ and end-to-end distance simulated in HyRes and HyRes II ................................................................. 111

Figure 4.12 Stability and dimension of folded proteins in HyRes and HyRes II ....................... 112

Figure 4.13 Major states of dynamic interactions of p53 TAD/DBD sampled in each of the eight independent HyRes II simulations ................................................................. 113

Figure 4.14 PCA analysis of conformational space sampled by HyRes simulations of the disordered p53 TAD with respect to DBD ................................................................. 114

Figure 4.15 Major states of dynamic interactions of p53 TAD/DBD sampled in each of the eight independent HyRes simulations ................................................................. 115

Figure 4.16 TAD interaction hot spots on DBD ......................................................... 116

Figure 5.1 Representation, efficiency, and accuracy of HyRes-GPU ................................. 128

Figure 5.2 Monomeric conformational properties of IDPs in HyRes-GPU .............................. 131

Figure 5.3 Spontaneous phase separation of GY-23 in HyRes-GPU ................................. 133

Figure 5.4 Phase separation of GY-23 peptide variants ............................................... 135

Figure 5.5 Conformational properties of WT GY-23 at phase coexistence at 300 K ............... 137

Figure 5.6 Residue helicity profiles and saturation concentrations of TDP-43 CR variants.... 140

Figure 5.7 Contact map and the relationship between helicity and phase separation behavior of TDP-43 CR ................................................................. 144

Figure 5.8 The probability distributions of $R_g$, $R_{ee}$, and residue helicity of model peptides ...... 153

Figure 5.9 Correlation of residue helicities between HyRes-GPU and NMR SSP results ...... 153

Figure 5.10 Preparation of the initial conformations for phase separation simulations ............ 154

Figure 5.11 The density of the largest droplets during six simulations of WT GY-23 at 300 K 154
Figure 5.12 Finite-size effects on the simulation of GY-23 condensates ........................................... 155
Figure 5.13 GY-23 phase separation properties ................................................................................. 155
Figure 5.14 Phase separation of WT and mutant GY-23 peptides ......................................................... 156
Figure 5.15 Conformational properties of monomeric GY-23 peptides .............................................. 157
Figure 5.16 $\beta$-structure properties in the condensed phase of WT GY-23 at 280 K and 260 K. 158
Figure 5.17 Residue contacts of GY-23 variants ................................................................................. 159
Figure 5.18 Helicity of TDP-43 CTD variants ...................................................................................... 160
Figure 5.19 $C_{\text{sat}}$ of WT, G335A, and G338A TDP-43 CTD ................................................................. 161
Figure 5.20 Helix-helix interactions in condensed phase ............................................................................. 162
Figure 5.21 Probability differences of intermolecular backbone-backbone contacts between WT and $\text{WT}^{+\Delta \text{Hel}}$ TDP-43 CR condensates .......................................................... 163
Figure 5.22 The effect of Ala on intermolecular residue-residue contact probabilities .................... 164
Figure 5.23 Probability differences of inter-molecular residue-residue contacts between WT and M337P ................................................................................................................................. 165
Figure 6.1 The pulling traces of proteases in $S.\ typhi$ pore and $E.\ coli$ pore, and representative snapshots for the mid-binding events ......................................................................................... 173
Figure 6.2 Mid-binding configurations and properties of proteases in $E.\ coli$ and $S.\ typhi$ pores 175
Figure 6.2 Mid-binding states and properties of proteases in $S.\ typhi$ E57A and $S.\ typhi$ E57K pores ...................................................................................................................................................... 178
Figure 6.4 Illustration of steered MD simulations pulling WNV NS2B/NS3 proteases through $S.\ typhi$ (green) and $E.\ coli$ (grey) ClyA ................................................................................................. 179
Figure 6.5 Illustration of standard MD simulations to explore protease dynamics in the middle of the ClyA (highlighted region) in ClyA .................................................................................. 180
CHAPTER 1
INTRODUCTION

1.1 Intrinsically disordered proteins and their functional activities

Intrinsically disordered proteins (IDPs) represent a unique class of functional proteins that remain fully or partially unstructured in the unbound state under physiological conditions [1-8]. Unlike the well-folded global proteins, IDPs are generally enriched with charged and hydrophilic residues but lack hydrophobic residues, rendering them structural heterogeneous and more exposed to the solvent [9-13]. The inherent thermodynamic instability of IDP conformations enables them to sensitively respond to numerous stimuli, positioning them as pivotal players in crucial biological activities, such as cellular regulation, post-translation, and signaling [14-19]. IDPs actively engage in one-to-many or many-to-one modulations, acting as dynamic hubs in protein-protein interaction networks and serving as scaffolds during biological condensation [19-23]. Their versatility extends them to be highly involved in self-regulation, drug bindings, assembles, and phase separations through dynamically low-affinity or high-specific interactions [24-29].

Despite their indispensability, the inherent flexibility of IDPs make them susceptible to promiscuous interactions in vivo [30-32]. For example, IDPs are prone to engage in unwanted interactions due to mis-expression, dysregulation, and pathological gelation, associating them with a spectrum of diseases, including diabetes, amyloidosis, and cancer [28, 33-37]. Specifically, functional IDPs in infectious viruses and bacteria consequently lead to severe abnormalities of host proteins by maliciously inhibitions [38-41]. While it has been established that many IDPs undergo disorder-to-order conformational transitions upon specific interactions [42-47], emerging evidence suggests that IDPs can persist in an unstructured state even within specific complexes.
and functional assemblies [48-53]. The intricate interplay between IDP dynamics and functions remains ambiguous. Therefore, understanding IDP conformations, dynamics and possible mechanisms in cellular interactions is pivotal for developing of pathological therapies and antibacterial/antiviral drug design.

Figure 1.1 Illustration of disordered protein structures and interactions. (A) The heterogeneous structures of intrinsically disordered proteins, with the secondary structure colored yellow, purple, cyan, and white for β-sheet, α-helix, turn, and coil properties, respectively. The structure of β-sheet and α-helix is shown in cartoon. (B) IDP (pKID) in complex with folded protein (KIX) (pKID/KIX complex, PDBID: 1KDX), while the disordered pKID folds into helix upon binding. The folded protein is shown in surface style and colored white. The IDP is shown in cartoon style and colored cyan. (C) The ensemble of IDP in complex with folded protein (exemplified by p53-TAD/CypD complex), while IDP p53-TAD remains disordered. The folded protein is shown in surface style and colored silver. Each structure of p53-TAD in the ensemble is shown in cartoon and colored gradient from blue to red (following the BWR color scheme). (D) The disorder-disorder protein complex (exemplified by GY-23 condensate), while IDPs remain highly disordered. Each chain in the complex is shown in
cartoon and colored by chain IDs. While (A) represents the structured complex, (C) and (D) together represents the fuzzy complex.

1.1.1 Coupled binding and folding of IDPs

IDPs/IDRs (intrinsically disordered region) are widely recognized for their prevalent in cells and critical roles in cellular regulation and signal transduction [14-19, 54]. During cellular activities, IDPs/IDRs often undergo disorder-to-order conformational transitions, adopting helix, β-structures, or disordered structures upon binding to their biological partners [42-47, 55, 56]. Noteworthy examples include the disordered kinase inducible activation domain (KID) of the transcription factor CREB, which will fold into a helix upon binding to the KIX domain of the CREB binding protein for transcription activation [57-62], and the disordered N-terminal domain of bacterium *staphylococcus* secreted protein (SPIN), which will form a hairpin by binding to the oxidative enzyme myeloperoxidase (MPO) to survive from the immune system [56, 63, 64]. Additionally, several IDPs exhibit diverse conformations upon binding to different targets, exemplified by the C-terminus of the tumor suppressor p53, which will form a helix in S100ββ-p53 complex, a sheet in, and coils in CBP-p53 and cyclin A2-p53 complex [65-67]. Such structural plasticity benefits IDPs to be favored by the target proteins and can effectively avoid the topology frustration during the active pore fitting.

In principle, there are three prevailing theories in explaining the folding-upon-binding activity adopted by IDPs (Figure 1.2A). On one hand, the residual structure may facilitate the effective binding process, referred to as conformational selection-like mechanism, in which the disordered proteins would approximately pre-form the structures that preferred by targets (Figure 1.2A, a) [68-70]. On the other hand, the structural plasticity plays critic roles in binding and can decrease the rigidity of residual conformation, where IDPs/IDRs can undergo rapid folding upon encountering the specific recognition efficiently, referred as induced folding-like mechanism.
Besides, the folding and binding can occur more cooperatively with comparable rates, a prevalent phenomenon easily given the delicate balance between the cost of folding entropy and the gain of the binding enthalpy (Figure 1.2A, trace c) [78]. Naturally, diverse systems employ varying mechanisms. To unravel the hidden mechanism of coupled binding and folding process, numerous experimental techniques such as nuclear magnetic resonance (NMR) and small-angle X-ray scattering (SAXS) provide valuable insights. Sugase et al characterized the mechanism adopted by the phosphorylated KID to bind to KIX using NMR, showing that pKID remain largely disordered in the unbound free state but would fold into helix and become stabilized by contacting with the hydrophobic residues on the surface of KIX folded domain (Figure 1.1B) [79]. On the other hand, ubiquitin presents “bound” conformations in the free state as characterized by X-ray, which indicate the conformational-selection mechanism are critical for different partner recognitions in ubiquitin-involved activities [80].
and $Q_{\text{fold}}$ stand for binding and folding property with values from 0 to 1 referring to unbound/unfold to bound/fold, respectively. The left panel draws the rough conformational states of IDP (colored orange) in binding. (B) The co-folding exemplified using NS2B and NS3 binding, with NS2B colored in red and NS3 in cyan, respectively. The N’- and C’- of NS2B are labeled out. The catalytic triad on NS3 is circled by green with the sidechain of three residues colored green.

Except for following the theories discussed above, the corresponding binding and folding have been frequently observed in many IDP functional activities [81, 82]. One well-known example is the co-folding of the flavivirus proteases NS3/NS2B, which are essential for viral replication by cleaving apart the virus polypeptides. NS2B/NS3 consists of two core components, the 184 amino acid NS3 protease and a 47 amino acid co-factor peptide NS2B, which are both known to exhibit significant intrinsic structural dynamics in the free state as being identified by NMR [82]. However, titration NS3 with NS2B showed dramatic conformational change for both two proteins reflected by the secondary chemical shift ($\Delta \delta$), indicating the formation of two $\beta$-barrel on NS3 and the co-folded $\beta$-structured NS2B N-terminal. The C-terminal of NS2B will fold into a hairpin and bind closely to the catalytic center of NS3 (illustrated in Figure 1.2B). After the co-folding, NS2B CTD still can dynamically fluctuate from the active site, representing the “open” and “closed” two states in reference to the ligand unbound/bound conditions, with a predomination of the “closed” conformation in the solution as indicated by the small-angle X-ray scatter and NMR studies [83-85]. Clear understandings on how NS2B/NS3 co-fold are necessary for discovering novel targeting strategies that may yield new lead compounds with drastically improved drug-like properties to treat flavivirus infections.

1.1.2 Fuzzy complexes of IDPs

Beyond undergoing the disorder-to-order conformational changes, IDPs can maintain a high degree of disorder upon binding to their targets. Such features have been observed in various interactions, including small molecule-disorder, disordered-ordered and disordered-disordered...
protein interactions. The non-specific dynamic interactions enable IDPs to properly regulate cellular activities, like augmenting the selectivity of the hots-guest bindings. One essential IDP involved cellular regulations is performed by the N-terminal transaction domain of tumor suppressor p53 (p53-TAD) [86]. p53-TAD orchestrates interactions with diverse signaling pathways to control cellular processes while remaining highly dynamic in specific complexes. For instance, the small molecule epigallocatechin gallate (EGCG) from green tea, can dynamically bind to p53-TAD in μM. By outcompeting the regulatory E3 ligase MDM2, EGCG induces the conformational collapse of p53-TAD, disrupting the MDM2-mediated degradation and stabilizing p53 for anti-tumor activity [87]. Recent nuclear magnetic resonance (NMR) studies and simulations reveal dynamic interactions between p53-TAD and its DNA binding domain (DBD, residues 95-290), which will compete against the non-specific DNA bindings, enhancing the target site discrimination of p53-DBD [88, 89]. Importantly, no obvious changes in secondary structure propensities of p53-TAD are observed between the bound and unbound states, as characterized both by experiments and simulations [89, 90]. Furthermore, p53-TAD participates in the regulation of mitochondrial permeability transition pore’s (mPTP) opening by dynamically bind to the cyclophilin D, the only established component of mPTP at present (Figure 1.1C) [53]. Overall, it is obvious that p53-TAD dynamic cellular interactions are critical in numerous biological activities.

IDPs also frequently serve as the central components in driving biomolecular phase separation [22, 91]. It has been suggested that IDPs can establish multiple non-specific interactions, like dynamically binding to folded proteins and nucleic acids, to serve as the scaffolds of the condensates [92, 93]. This ability, known as multivalency, arises from diverse interactions, including electrostatic and polar interactions, π-interactions, hydrogen bond interactions, and hydrophobic interactions, among others [94-98]. Despite their subtle strength, these multivalent
interactions are potent enough to facilitate IDPs association without triggering excessive folding or aggregation. The less stiffness compared to the folded proteins also renders them to dynamically exchanges within the biological condensates much easier. Experiments have demonstrated that IDPs can undergo self-assembly, and mutations in IDPs can alter their phase behavior, potentially leading to pathologies such as neurodegenerative diseases. For example, mutations in RNA-binding proteins FUS and TDP-43 are associated with amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) [99-101], while Tau protein dysfunction is prone to have liquid-phase to solid-phase transitions, resulting in pathogenic aggregation within cellular condensates that associated with neurodegenerative disorders like Alzheimer's disease (AD) [102-104]. Moreover, growing experiments indicate that the structural changes of certain IDPs substantially impact the phase behavior. For instance, many IDPs and IDR s adopt more extended conformations in the condensate state and exhibit higher propensity for $\beta$-structures [105-107]. Some studies suggest that phase separation is able to induce folding of poly-(PR) peptides into helical conformations, which is a hallmark of aging-related phase transition pathologies [108]. Recent studies have also shown that the C-terminal domain (CTD) of TDP-43 undergoes transient conformational changes during phase transitions, forming helices during phase separation and subsequently aging into amyloid structures [109, 110]. Therefore, understanding the conformational dynamics of IDPs within phase separation is crucial for comprehending this biologically significant condensation process.

1.2 Multi-scale molecular dynamic simulations in studying IDPs

Studying the conformation and dynamics of IDP poses a significant challenge when relying on traditional ensemble-based techniques like NMR [111, 112], small-angle X-ray scattering (SAXS) [113-115] and Förster resonance energy transfer (FRET) [116, 117]. These methods along
struggle to describe the detailed conformational fluctuations due to their limitation in measuring only averaged properties. This limitation impedes the resolution of the heterogeneity and transient structural features inherent in IDPs. Therefore, molecular dynamics (MD) simulations using transferable energy functions emerge as a particularly powerful tool on offering dynamic insights into IDPs at multi-scale levels.

1.2.1 Atomistic simulation of IDP structure, dynamics, and interactions

Recent breakthroughs in hardware, including GPU-enabled MD algorithms that provide over a 100-fold acceleration compared to traditional CPU-based approaches [118, 119], and the specialized supercomputers like Anton that is capable to achieve over 17 μs simulation time per day for biosystem consisting of ~ 24000 atoms [120, 121], have significantly advanced atomistic simulations of proteins in the explicit solvent. These advances, coupled with the developments of the accurate general force fields, which have been extensively rebalanced for describing folded as well as disordered proteins, have enabled in-depth investigations into biological system of interests. Particularly, the a99SB-disp and CHARMM36m (referred to as C36m) are two most popular force fields that have been widely applied due to their reliability in presenting detailed conformational ensembles for diverse folded and disordered proteins, aligning well with the structures measured by experiments [122, 123]. The extensions of small molecular force fields, such as GAFF and CGenFF, further facilitate the study of interactions between small-molecules and proteins [124-128]. As being discussed, simulations of EGCG/p53-TAD interaction in a99SB-disp force field provide significant insights into the binding surface and conformational changes of p53-TAD [87].

Atomistic simulations have played crucial roles in unraveling the kinetics and thermodynamics of coupled binding and folding, providing valuable insights into essential various sub-states. For example, the Shawn group conducted a study on the interaction mechanism
between an α-helical molecular recognition element and two disordered peptides. Utilizing a99SB-disp force field, these simulations not only re-capitulate the structural propensity of IDPs, but also provide the free-energy landscape in the binding events to help characterize the mechanisms of folding-upon-binding [129]. Notably, for some homologous IDP, which share high sequence identity and conformational similarity, exhibit diverse functional capacities towards different biological targets, which dramatically improves their specificity towards different host species. Zhang et al underlies the inhibitory mechanism between Staphylococcus Peroxidase Inhibitor (SPIN) and human oxidative enzyme MPO by resolving the dynamic interactions between SPIN N-terminal domain (NTD) and MPO using the C36m force field [130]. In contrast to the well-populated disorder-to-α-helices formation events, where the conformational-selection and induced-folding theories mostly work for, thermodynamic results indicated that the formation of β-hairpin, being adopted by SPIN-NTD, generally led to broader binding states, considerably slower folding rates (in micro-second), and complex temperature dependent behaviors (anti-Arrhenius behavior) [131]. The atomistic simulations provide a predictive understanding of how the sequence, residual structures, or conformational propensities contribute to the SPIN homologs’ inhibitory capacities (Please see CHATPTER 2 for more detailed discussions).

Atomistic simulations also provide significant insights into protein backbone involved interactions and potential the effects of transient structures in phase separation of IDPs. However, due to the prohibitive computational cost for condensation simulations, the application of all-atom MD simulations in studying phase separation is usually restricted to small model peptides. For example, Rauscher et al. simulated the model condensate containing 27 copies of elastin-like peptides, (GVPGV)7, showing that peptide backbone would form transient and sparse hydrogen-bonded turns, but individual chains were maximally disordered [132]. Galvanetto et al. performed
6 \( \mu \)s-long condensate simulations containing 96 prothymosin-\( \alpha \) and 80 histone H1 molecules, corresponding to roughly 4 million atoms, which obtained enough sampling for a meaningful comparison with experiments [133]. Sarthak et al. even enabled condensate simulations of tens of millions of atoms and achieved over 150 ns/day with their NAMD implementation on the supercomputer [134]. It is worth noting that atomistic simulations stand out as one of the most powerful and straightforward approaches on probing detailed IDP dynamics, especially focusing on the backbone-related interactions, at present.

1.2.2 Enhanced sampling methods for efficient simulation of IDPs

As one of the most accurate and powerful computational method, direct atomistic simulations of extensive conformational fluctuations and transitions of proteins remain extremely challenging. For example, re-analysis of a 30-\( \mu \)s simulation of a relatively small 40-residue A\( \beta \)40 peptide generated using special purpose supercomputer ANTON 2 revealed a surprisingly limited level of convergence at both the secondary and tertiary structure levels [135]. This underscores the critical need for so-called enhanced sampling techniques for generating statistically representative structural ensembles of proteins. This challenge becomes even more important in the studies of IDPs, whose structures and functions need to be described using heterogeneous conformational ensembles [59, 136-139]. Additionally, owing to IDPs more solvent-exposed nature, the larger number of solvent molecules will further increase the computational cost. Arguably, the functional mechanism of an IDP lies in how its disordered ensembles respond to various cellular stimuli and signals, such as the binding of ligands and cofactors, changes in cellular environments, and post-translational modifications. Therefore, there is a critical need to generate well-converged disordered ensembles of an IDP and resolve their differences in various states in order to establish the molecular basis of its function.
Implicit solvent models offer an alternative efficient approach by reducing system size via excluding the solvent molecules and enabling faster sampling of solute conformations. These models are designed to achieve accurate descriptions of the solvation free energies that govern solvent effects on solute conformations, including the Generalized Born and surface area continuum solvation model (GBSA) [140], Generalized Born and molecular volume model (GBMV) [141-144], and the Poisson-Boltzmann and surface area continuum solvation model (PBSA) [145]. The application of implicit solvent models provides faster energy calculations and conformational sampling, with up to a 60-fold speedup compared to explicit solvent models in folding simulations of small systems and resolving protein interactions [146]. For example, Ganguly et al identified the multistep sequential induced folding mechanism of the pKID/KIX recognition by performing atomistic simulations with GBSA model, which is largely consistent with the NMR characterizations but provides more detailed insights towards the intermediate binding stats [59]. However, accurate descriptions of the solvation free energy are still challenging for implicit solvent models. This can come from an inaccurate description of the non-polar solvation energy due to the inadequate capture of the solvent accessible surface area towards protein folded inner cavity.

The emergence of enhanced sampling techniques helps to efficiently sample the protein conformational space by either adding a bias force or potential to the system to cross the high-energy barriers between local energy minima during the simulations [147]. Such techniques can be divided into methods that either the collective variables (CV) dependent or independent. The CV-based methods, like umbrella sampling [148], metadynamics [149], and steered MD [150], will require a pre-selection of the reaction coordinates that can be used to describe the process of interest. The speed-up is thus realized by reducing the phase space to the space of the CVs. The
CV-independent methods, on the other hand, are more attractive to study the complex biosystems that are difficult to pre-define the reaction variables, like IDPs. Along with the breakthroughs of many force fields that greatly extend the physics-based atomistic simulations [29, 122, 123, 151, 152], the temperature replica exchange Molecular Dynamics (T-RE) [153] has emerged as one most powerful CV-free protocols for practical simulations of biomolecules in explicit solvent [154-156]. By exchanging the temperatures with the neighbor replicas, T-RE executes the random walk on the energy surface without being trapped at the local energy minima. However, the number of replicas required for sufficient samplings scales as $O(f^{1/2})$, where $f$ is the degrees of freedom (DOFs) of the whole system [157]. To expand the reachable system size, several variants [157-163] of RE have been developed to either reduce the replicas needed or promote faster conformational transitions. Particularly, the replica-exchange with solute tempering [159, 160] (REST) method, proposed by the Bern group, has proven to improve the exchange efficiency in protein folding simulations. In REST, only the selected “solute” region (always referring to biomolecules) is subjected to tempering (or Hamiltonian scaling) while the rest of the system (“solvent”) will maintain unscaled at all temperature conditions. Since the solvent-solvent Hamiltonian is untouched, their DOFs are excluded from the exchange process, which reduces the replicas needed for large biosystems dramatically comparing to the traditional REX method. In the 20000 atoms alanine dipeptide system, REST was found to be 7-10 times more efficient than T-REX that requiring 22 replicas to cover the temperatures ranging from 300 to 600 K [159]. The “ignoring” of the solvent DOFs becomes more attractive in IDP studies, where the larger solvent box is required due to the generally more extended biomolecular globular structures. REST has been identified to be able to cover the conformational space and derive the structural heterogeneity
of the disordered Bcl-2 protein more efficiently and sufficiently than the standard MD simulations [164].

1.2.3 **Coarse-grained simulations in studying long timescale IDP dynamics and interactions**

Although enhanced sampling methods have alleviated some computational cost challenges in atomistic simulations, the inherent complexity of biological processes and micro-second to second level functional timescale often necessitates extensive simulation times for a comprehensive understanding. To date, sampling at the coarse-grained (CG) level is one of the most optimal approaches for studying biological systems. By grouping atoms into beads, CG models decrease the system’s degree of freedom and smoother the energy surface. Meanwhile, by omitting specific atom bond vibrations, the timestep used in CG simulation can be largely increased. With careful parametrizations guided by atomistic simulations or experiments, the state-of-the-art CG models present high potency on investigating the focused biological properties with efficiency and accuracy simultaneously. For example, the topology-based (Gō-like) models have also proven effective for determining the mechanism and kinetics of IDP interactions, particularly the coupled binding and folding process. The implication is that binding and folding are governed by similar principles that require minimal frustration for efficiency. Note that Gō-like models generally require additional calibrations to provide a more quantitative description of the balance between intermolecular interactions and intrinsic conformational propensities [165, 166]. Ganguly et al developed the calibrated sequence-flavored Gō-like model to study the coupled binding and folding. Calibrations though the inclusion of the knowledge-based MJ statistic potentials, the sequence-flavored Gō-like model avoids the over-estimation of the native-contacts in the original Gō-model [167]. This calibrated model successfully simulated the co-folding between the p160
steroid receptor co-activator (ACTR) and nuclear coactivator binding domain (NCBD) of transcription coactivator CREB-binding protein with residue-level insights, revealing that the NCBD/ACTR recognition is mainly initiated by forming a mini folded core and how the conformational selection might contribute to efficient recognition of IDPs [81].

One limitation of the Gō-like models is their dependence on native contacts, the nature that may not always be applicable to highly disordered proteins. Therefore, CG models that specifically designed to study IDP conformations, dynamics, and interactions have been developed. Baul et al developed the self-organized polymer coarse-grained model (SOP-IDP) based on the united residue CG model, which is powerful in delineating IDP sequence information with structures with accessible chain length up to 441 residues [168]. Wu et al developed the associative memory, water-mediated, structure and energy model (AWSEM-IDP), which can efficiently sample large conformational spaces for IDPs as well with a significant separation to globular protein conformational geometries [169]. Liu et al developed a hybrid resolution coarse-grained (HyRes) model with atomistic level description of the backbone atoms and the semi-qualitive representation of the residue sidechains [170]. Notably, HyRes model can qualitatively capture the long-range non-specific interactions and semi-quantitatively describe IDP residual structures.

As of now, the most successful biological condensate simulations relied on CG models significantly advanced by the lower degree of freedom and a smoother free energy landscape. Various CG models have been developed to study the phase separation of IDPs, including the HPS model and its variants [171-177], MOFF [178], MARTINI [179], CALVADOS [180, 181], and Mpiipi model [182], where each residue is represented using one or several CG beads. Simulations with those models can reach the corresponding time and length scale for sufficient sampling of condensate dynamics, which have obtained great success on capturing the phase separation and
the mutational effects. However, due to the lack of explicit description of backbones, most CG models are not yet well suited to capture the residual secondary structure properties of IDPs, as well as the backbone-mediated interactions [183], which have shown to regulated the phase separation in poly(PR), TDP-43 CTD, and elastin-like peptide condensation as reported [107-109]. The HyRes model developed by our group and being carefully calibrated presents a good performance in reproducing transient secondary structures of IDPs and binding with folded proteins. The model has been successfully applied to study the coupling between secondary structure and phase separation.

1.3 Dissertation outline

Recognizing the significance of IDP conformations, dynamics, and interactions on regulating cellular activities, this dissertation will include applying multi-scale simulations to investigate IDP dynamics and functions, as well as development of new enhanced sampling methods for more efficient sampling of IDPs. Specifically, chapter 2 [130] discusses how to use atomistic simulations to study the coupled binding and folding of SPIN/MPO, aiming to investigating why different SPIN species present different inhibition ability although sharing high sequence identity and conformational similarity. Followed by chapter 2, chapter 3 [184] discussed the necessity of enhanced sampling method on driving more efficient atomistic simulation of IDPs with the explicit solvent. In this chapter, we calibrate the REST2 method by re-balance the protein-protein and protein-solvent interactions. We illustrate how this balance will influence the obtained ensembles, which will further influence the exchange efficiencies. We also exam how this balance will affect the conformational sampling of IDPs by carefully comparing the performance between calibrated REST2 (here referred as REST3) and REST2 protocol. In chapter 4 [90], we discussed the necessity of using coarse-grained models to study long time-scale protein activities and
introduced an optimized HyRes model. Particularly, we aim to solve the over-compaction problem in the original HyRes model, which is coming from the lack of descriptions on the solvation energy. Meanwhile, we focus on preserving the ability of HyRes on describing IDP non-specific interactions and residual structures. In chapter 5 [185], we discussed one application of HyRes model on studying the homotypic IDP phase separation. Particularly, we first test if HyRes is able to describe the spontaneous phase separation of model peptides and whether the model is capable to capture the mutational effects. We then investigate how the residual structural properties affect the phase separation for GY-23 and TDP-43 conserved region (CR). In chapter 6, we showed another application of HyRes model that was applied to investigate the dynamic interactions between WNV NS2B/NS3 proteases in the ClyA nanopore. We first explored if there are specific binding that favored by NS2B/NS3 proteases and ClyA nanopore. We then identified the significant residue on ClyA, which can be engineered and help to stabilize the proteases/pore interactions. We aimed to figure out how to resolve different conformational states of WNV NS2B/NS3 proteases by specifically engineering the ClyA nanopore. Finally, a brief summary and directions for further research are discussed in chapter 7.
CHAPTER 2

STRUCTURAL STABILITY OF DISORDERED SPIN N-TERMINAL REGION IN MYELOPEROXIDASE INHIBITION

Gram-positive pathogenic bacteria \textit{Staphylococcus} express and secret staphylococcal peroxidase inhibitor (SPIN) proteins to help evade neutrophil-mediated immunity by inhibiting the activity of the main oxidative-defense player myeloperoxidase (MPO) enzyme. SPIN contains a structured 3-helix bundle C-terminal domain, which can specifically bind to MPO with high affinity, and an intrinsically disordered N-terminal domain (NTD), which folds into a structured \(\beta\)-hairpin and inserts itself into the active site of MPO for inhibition. Mechanistic insights of the coupled folding and binding process are needed in order to better understand how residual structures and/or conformational flexibility of NTD contribute to the different strengths of inhibition of SPIN homologs. In this work, we applied atomistic molecular dynamics simulations on two SPIN homologs, from \textit{S. aureus} and \textit{S. delphini}, respectively, which share high sequence identity and similarity, to explore the possible mechanistic basis for their different inhibition efficacies on human MPO. Direct simulations of the unfolding and unbinding processes at 450 K reveal that these two SPIN/MPO complexes systems follow surprisingly different mechanisms of coupled binding and folding. While coupled binding and folding of SPIN-\textit{aureus} NTD is highly cooperative, SPIN-\textit{delphini} NTD appears to mainly utilize a conformational selection-like mechanism. These observations are in contrast to an overwhelming prevalence of induced folding-like mechanisms for intrinsically disordered proteins that fold into helical structures upon binding. Further simulations of unbound SPIN NTDs at room temperature reveal that SPIN-\textit{delphini} NTD

has a much stronger propensity of forming β-hairpin like structures, consistent with its preference to fold and then bind. These may help explain why the inhibition strength is not well correlated with binding affinity for different SPIN homologs. Altogether, our work establishes the relationship between the residual conformational stability of SPIN-NTD and their inhibitory function, which can help us develop new strategies towards treating Staphylococcal infections.

2.1 Introduction

*Staphylococcus* is a group of gram-positive pathogenic bacteria that can lead to a broad range of infections including pneumonia and toxic shock syndrome [186, 187]. Staphylococcal infections are becoming an increasingly severe threat to public health, with an estimate of ~3 million cases in the United States every year and expanding incidence of antibiotic resistance [187]. To defend against the invasions of *staphylococcus*, neutrophils are critical innate immune response components in hosts and serve as the first defensive line by releasing the anti-bacterium hypochlorous acid [63] and other reactive oxidant species (ROS) [64, 188]. Particularly, myeloperoxidase (MPO) is one of the most abundant granule enzymes in neutrophils that can catalyze the production of ROSs from hydrogen peroxide (H₂O₂) to help kill the bacterium. However, *Staphylococcus* has been found to be able to evade the neutrophil-mediated innate immune defense and sometimes turn host cells into “Trojan Horses” for bacterial dissemination in vivo [56, 189-192]. In particular, the bacterium can secret Staphylococcal Peroxidase INhibitor (SPIN) proteins, which bind MPO with nanomolar affinity and inhibit its enzymatic activity [63, 188]. SPIN consists of an intrinsically disordered N-terminal domain (NTD) and a structured 3-helix bundle C-terminal domain (CTD) [63, 188, 193]. The inhibitory activity requires the disordered SPIN NTD and can be largely abolished with deletion or certain mutations of the NTD region [193]. Structural studies have revealed that SPIN NTD folds into a β-hairpin and inserts
itself into MPO’s active site in the complex [193], which prevents the substrate H₂O₂ from accessing the catalytic heme in MPO’s active pocket. As a result, the enzyme becomes incapable of producing ROSs, thus protecting *Staphylococcus* from killing by neutrophils [188].

Recently, multiple SPIN homologs that share high sequence identity and conformational similarity have been identified with various inhibitory capacities towards human MPO [63]. Interestingly, their inhibitory capacities show little correlation with their binding affinities to MPO [63]. For example, while SPIN-agnetis binds human MPO with a $K_D$ of ~42 nM, it has little measurable inhibitory effect on MPO activity. The implication is that, the folded SPIN CTD largely determines the binding affinity to MPO, while the disordered NTD dictates the inhibitory efficacy. Furthermore, structural studies suggest that all SPN NTD homologs likely fold into essentially the same β-hairpin conformation in the bound state [63]. Therefore, functional differences between SPIN homologs may be directly related to the disordered unbound state and/or the coupled binding and folding processes themselves. Specifically, two key questions are: (1) how residual structures or conformational plasticity contribute to the facile folding and binding of SPIN NTD, thus potentially impacting the inhibition strength, and (2) whether SPIN homologs show different mechanisms of coupled binding and folding.

Intrinsically disordered proteins/regions (IDPs/IDRs) like SPIN NTD are prevalent in biology and frequently play key roles in cellular regulation and signal transduction [1-3, 5, 7, 68]. IDPs also frequently undergo coupled binding and folding for function [194-199]. Two classes of mechanisms have been generally invoked in studies of IDP coupled binding and folding. In so-called conformational selection-like mechanisms [200-204], residual structures in unbound state of an IDP may resemble the folded complex and serve as initial binding sites to facilitate efficient molecular recognition (that is, fold and then bind). On the other hand, an IDP could undergo rapid
folding upon nonspecific encountering with its target, following the so-called induced folding-like mechanism [205-207]. Here, structural plasticity plays a more important role, such as to enable facile IDP folding on the target surface [74, 81, 208-216]. For the cases where the binding pocket is deep and rugged, induced fitting can direct the peptide to reach the spot and then fold to the energetically favored states [43, 46, 75, 217-219]. It should be noted that existing mechanistic studies have mainly involved IDPs that fold into α-helices, ordered loops or a single β-strand upon binding and that induced folding has been found to be prevalent in these IDPs [135, 139, 196]. SPIN NTD is notably different from these existing studies; it represents the first case study of coupled binding and folding of an IDP into a β-hairpin. Folding of β-hairpin structures involves cooperative formation of long-range contacts and has been shown to be much slower than helix-coil transitions with substantial entropy-dominant free energy barriers [220-222]. It remains unclear if SPIN-NTD will display similar mechanistic features to IDPs with simple folded structures.

In this work, we focus on two SPIN homologs, SPIN-aureus and SPIN-delphini. They share 53% sequence identity and 80% sequence similarity, and both bind to human MPO with nanomolar affinities and fold into essentially identical β-hairpin structures [56, 189-192]. Interestingly, although SPIN-delphini binds to MPO ~19 times weaker than SPIN-aureus, its half maximal inhibitory concentration (IC₅₀) is only ~6 times higher. We will mainly utilize atomistic simulations in explicit solvent to probe the conformational properties of unbound NTDs from SPIN-aureus and SPIN-delphini and to investigate their coupled binding and folding processes. Such simulations have significantly benefited from recent advances in both GPU-enabled MD algorithms [119, 223-227], which can provide over 100-fold acceleration compared to traditional CPU-based approaches, and accurate general-purposed protein force fields [122, 123, 228-231],
which have been extensively rebalanced for describing both folded and disordered proteins. Simulations of temperature-driven dissociation process of two SPIN/MPO complexes at 450 K recapitulate that SPIN CTD dominates specific binding to MPO and further reveal surprising differences in coupled binding and folding of NTD of these two SPIN homologs. The binding and folding are highly cooperative for SPIN-\textit{aureus} NTD, while SPIN-\textit{delphini} NTD prefers to be partially folded before binding to the MPO active site. Further simulations at the room temperature show that unbound SPIN-\textit{delphini} NTD is much more structured. These results suggest an important role of residual structures of SPIN NTD in its facile recognition and inhibition of MPO, which may help us better understand the sequence-structure-function relationship of SPIN.
**Figure 2.1 Representative structure of SPIN/MPO complexes.** (A) Overlay of structures SPIN-*aureus* and SPIN-*delphini* in complex with human MPO. The structures were taken from PDB 5UZU and 6BMT for SPIN-*aureus* and SPIN-*delphini*, respectively. MPO is colored grey and SPIN-*aureus* and SPIN-*delphini* are colored red and green, respectively. Heme was shown in sticks. (B) NTD sequences of two SPIN homologs with conserved residues are highlighted in blue and similar residues in yellow. The sequence alignment is calculated using BLAST [232] that shows 53% identity and 80% similarity. The secondary structures in the bound state are marked with arrows.

### 2.2 Methods

#### 2.2.1 High Temperature Simulations of SPIN/MPO Complexes

All simulations were performed with the GPU accelerated CHARMM/OpenMM interface [223, 233, 234] in CHARMM36m force field [122], one of the most accurate transferable models that is suitable for IDP simulations, especially those favoring the β-sheet conformational profiles. The initial structures of SPIN-*aureus* and SPIN-*delphini* in complex with human MPO were taken from the crystal structures (PDB 5UZU SPIN-*aureus* [193] and 6BMT for SPIN-*delphini* [63]) (see Figure 2.1 A). To reduce the computational cost, only segments of MPO that are within 12 Å of SPIN are included in the current simulations, which consist of residues 167-200, 255-444, 490-506, 526-540, and 566-596 for MPO (Figure 2.6). To prevent the unfolding of MPO, all backbone heavy atoms of structured MPO segments (excluding loop residues 268-288, 380-395, and 317-328) and the bound heme group were restrained by harmonic potentials with a force constant of 1.0 kcal/(mol Å²) in all simulations. Proper amount of Na⁺ and Cl⁻ ions were added to neutralize the systems and to reach a NaCl concentration of 50 mM in accord with the experimental conditions [63]. The final solvated box contains about ~30000 TIP3P water molecules and has a dimension of ~9.2 x 9.4 x 11.3 nm³.

Each solvated system was first energetically minimized for 500 steps using steepest decent and another 500 steps using the adopted basis Newton-Raphson algorithm. The system was then slowly heated up from 100 K to 300 K in 10 ps under the constant volume condition. Equilibration
Simulations were then performed at 300 K and 1 atm for a total of 1 ns, during which all protein heavy atoms were additionally restrained using harmonic potentials with force constants slowly decreasing from 5.0 kcal/(mol Å²) to 0.0 kcal/(mol Å²). Langevin thermostat was used to control the temperature and Monte Carlo barostat with volume move attempt every 25 steps was used to control the pressure. Lengths of all bonds involving hydrogen atoms were constrained using the SHAKE algorithm [235] to allow for an integration time step of 2 fs. Long-range electrostatic interactions were treated using the particle mesh Ewald method [236], and the short-range van der Waals (vdW) interactions were treated with the twin-range cutoff at 12 and 14 Å.

To identify the optimal temperatures for unbinding/unfolding simulations, a series of pilot simulations were performed at temperatures ranging from 400 K to 500 K at 1 atm. Once an optimal temperature was identified (450 K), two sets of simulations were performed for each complex to probe temperature-induced SPIN unfolding and unbinding process. In one set, three additional simulations were performed at 450 K with different initial velocities to better characterize the dissociation of SPIN from MPO. These simulations were run until the NTD dissociated from the active pocket (i.e., with the fraction of native contacts between two molecules \( Q_{\text{inter}} < 0.3 \)), which all occurred within 400 ns. In the second set, 40 independent replicas were performed for each complex at 450 K for 250 ns each, with the helical region of SPIN CTD (Figure 2.1B) harmonically restrained with a force constant of 1.0 kcal/(mol Å²). The purpose of the second set is to directly examine the unfolding and unbinding of the NTDs.

2.2.2 Room Temperature Simulations of Free NTDs

The initial folded hairpin structures of SPIN-NTDs were taken from the same complex structures (Figure 2.1A). Both systems contain a 13-residue fragment (SPIN-aureus residues 33-45 and SPIN-delphini residues 28-40; see Figure 2.1B for sequences). 20 replicas were used to
simulate the unfolding events for two SPIN-NTDs at 300K. The solvated systems contain ~3500 TIP3P waters and have dimensions of ~4.2 x 4.3 x 5.4 nm³. Similar protocols as described above were applied to minimize and equilibrate the system. For each system, 20 independent production simulations were performed for 50 ns each at 300 K, which was sufficient to observe spontaneous unfolding of the β-hairpin structure.

2.2.3 Analysis

All the analyses were carried out using CHARMM and additional in-house scripts. All molecular visualizations were prepared using VMD [237]. The fractions of intermolecular and intramolecular native contacts, $Q_{\text{inter}}$ and $Q_{\text{intra}}$, are calculated to monitor the unfolding and unbinding process. The native contacts are first identified from the crystal structure of two complexes if the minimum heavy atom distance between two residues is no greater than 4.2 Å (Table 2.1 and 2.2). Note that for intramolecular native contacts, we exclude residue pairs that are close in sequence space and only consider those whose residue IDs are different by at least 3. The contacts in simulation trajectories were then calculated using the same criterion. Based on protein folding funnel theory, native interactions dominate the overall pathway [238-241]. Therefore, only native contacts were considered here. The unbinding and unfolding kinetics were analyzed using a double exponential approximation of the decay of $Q_{\text{inter}}$ and $Q_{\text{intra}}$ averaged over all replica runs (40 for the complexes and 20 for free NTDs). The first 50 ns trajectories were considered in unfolding and unbinding kinetic analysis, which were sufficient to capture the dissociation events. Pseudo free energy surfaces were also calculated to better characterize the baseline mechanisms of coupled binding and folding, derived directly from two-dimensional (2D) probability distributions along $Q_{\text{inter}}$ and $Q_{\text{intra}}$. For the data used to construct contact probabilities, we specifically focused on short segments of the trajectories where actual dissociation transitions
occurred. For example, only the first 15 ns trajectory in replica 1 of SPIN-aureus/MPO simulation was considered, which included the entire unbinding and unfolding transition (see Figure 2.7). By doing this, the results will not be interfered by the transient refolding events observed after complete dissociation (see Figure 2.7 replica 40 at 200 ns for example). The segments for each trajectory that were selected to calculate the contact maps can be found in Table 2.3. Note that for replicas where NTD remains bound and folded at the end of the 250 ns-simulation, we only selected the first 50 ns of trajectories to compute contact maps. In this way, we could avoid masking important details about the transition pathways by over-representing data of the bound and folded state.

2.3 Results and Discussion

2.3.1 High Temperature Simulations Reveal Step-wise Binding of SPIN NTD and CTD

High temperature simulations have been shown to be capable of providing reliable mechanistic insights into folding of structured proteins as well as coupled binding and folding of IDPs [242-246]. The assumption here is that unfolding and unbinding is largely a reverse of coupled binding and folding. However, it is also known that the most probable transition pathways may depend on the temperature [247]. Therefore, it is important to find the lowest temperature to drive the unfolding and unbinding process within a given simulation timeframe. The pilot simulations suggest that the NTD of SPIN-aureus only starts to dissociate from the active pocket of MPO at 450 K within ~100 ns timescale, which becomes much faster at higher temperatures (Figure 2.6B). Note that rapid dissociation (e.g., at 475K) is not always preferred due to the risks of missing important details under non-physiological conditions and activating pathways not generally accessible under the physiological conditions. For example, the three-helix bundle of SPIN CTD would melt rapidly at 475 K and above, leading to premature disassociation from MPO.
within 10s of ns. This is consistent with the experimental observation that SPIN CTD largely dictates MPO binding [193]. Instead, simulations at 450 K seem to depict a more realistic dissociation process, where NTD unbinds first while the CTD remains largely fold and bound (Figure 2.2). The apparent decoupling and step-wise nature of the binding of SPIN CTD and NTD could explain why there is little correlation between the inhibition strength and binding affinity for different SPIN homologs. It’s likely that two domains of SPIN bind and function almost independently when interacting with MPO. As such, some SPIN homologs, e.g., SPIN-agnetis, show comparable nanomolar binding affinity as SPIN-aureus, but have no detectable inhibitory ability to human MPO [63]. Based on these observations, we will focus on the coupled binding and folding of SPIN NTD while the CTD is harmonically restrained to the bound state in subsequent simulation and analysis.

![Figure 2.2 Fraction of native SPIN/MPO intermolecular interactions. Fractions of native intermolecular contacts between the NTD (blue) and CTD (red) of SPIN during four independent simulations of the SPIN-aureus/MPO complex at 450 K.](image-url)

2.3.2 Cooperative Binding and Folding of SPIN-aureus NTD
A total of 40 independent 250-ns simulations were performed at 450 K to explore the conformational fluctuations, dynamic interactions and dissociation pathways of SPIN-*aureus* NTD with human MPO. As summarized in Figure 2.7, SPIN-*aureus* NTD tends to dissociate rapidly and its unfolding and unbinding often happen simultaneously. For example, in 36 out of 40 replicas (except for replicas 6, 8, 16 and 25), NTD fully dissociated (with $Q_{\text{inter}} < 0.2$) within 200 ns. Particularly, among 30 out of the 36 runs (except for replicas 10, 11, 17, 23, 29 and 35) unbind/unfold occurred within the first 50 ns, or sometimes even more rapidly within 15 ns. To quantitively describe the dissociation process and probe the mechanisms of coupled binding and folding, we calculated the average fractions of intermolecular and intramolecular native contacts formed by NTD, denoted $Q_{\text{inter}}$ and $Q_{\text{intra}}$, respectively, from all replicas. The results were then fitted with a double exponential function (Figure 2.3 A). Not surprisingly, the unbinding and unfolding kinetics of SPIN-*aureus* NTD are similar, consistent with the observation that they appear highly correlated. As shown in Figure 2.3A, the initial fast phase $\tau_1$ for unbinding and unfolding are 0.12 and 0.28 ns, respectively, followed by a slow phase unbinding ($\tau_2$ of 11.40 ns) and unfolding ($\tau_2$ of 13.45 ns). We further constructed the pseudo 2D free energy surface as a function of NTD $Q_{\text{inter}}$ and $Q_{\text{intra}}$, derived from the dissociation transition segments (see Methods for details). The result, shown in Figure 2.3 B, confirms a highly cooperative mechanism of SPIN-*aureus* NTD coupled binding and folding with NTD $Q_{\text{inter}}$ and $Q_{\text{intra}}$ increasing simultaneously in a highly correlated fashion. The minimum free energy path (dashed line) largely follows the diagonal line expected for an ideally cooperative mechanism.
Figure 2.3 Cooperative binding and folding of SPIN-aureus. (A) Average intramolecular and intermolecular native contact fractions ($Q_{\text{inter}}$ and $Q_{\text{intra}}$) as a function of simulation time at 450 K. The double exponential fits are plot using dotted lines, with the actual parameters also shown. (B) Pseudo free energy surface as a function of $Q_{\text{inter}}$ and $Q_{\text{intra}}$ derived from the transition paths (see Methods). The dashed line indicates the minimum free energy pathway. Key states (U, I and B) are also labeled. (C) Representative conformations for key states along the minimum free energy path, with SPIN-aureus and MPO shown in red and light grey, respectively.

The free energy surface also reveals three major conformational states of NTD folding and binding to MPO. State B (bound), with both $Q_{\text{inter}}$ and $Q_{\text{intra}}$ above 0.8, is the fully folded and bound state, and State U (unbound), with both $Q_{\text{inter}}$ and $Q_{\text{intra}}$ below 0.2, is the fully unfolded and unbound state. In addition, there is a partially bound and folded substate, I (intermediate state), where the values of $Q_{\text{inter}}$ and $Q_{\text{intra}}$ are around 0.4. Representative conformations for the three states of the complex (Figure 2.3C) illustrate that SPIN-aureus NTD does not tend to pre-fold into some ‘native-like’ β-hairpins conformations prior to binding to the active site of MPO, and vice versa. The cooperative nature of SPIN-aureus NTD is in contrast to previous experimental and computational studies of coupled binding and folding of IDPs into non-β-hairpin structures [248], where induced folding-like mechanisms are prevalent. However, this may not be surprising given the cooperative nature of folding of isolated β-hairpins [220-222]. In particular, the “speed-limit”
of β-hairpin folding usually is \(~1\mu s^{-1}\), much slower compared to helix-coil transitions (~100 ns), due to the requirement of forming long-range interactions and the presence of entropy-dominant barriers. Therefore, once SPIN-\textit{aureus} CTD is tightly bound, native-like interactions with the MPO surface play a direct role to facilitate the rapid folding of NTD and achieve a facile blockage of the MPO active site for inhibition.

2.3.3 Conformational Selection-like Mechanism for SPIN-\textit{delphini} NTD

Compared to SPIN-\textit{aureus}, which is secreted by \textit{S. aureus} that appears to be particularly adapted to survive the neutrophil-mediated immunity with the highest binding affinity (K$_D$ = 15.9 nM) and inhibition strength (IC$_{50}$ = 4.6 nM) to human MPO, SPIN-\textit{delphini} has a moderate binding affinity (K$_D$ = 310 nM) but the 2nd strong inhibitory ability (IC$_{50}$ = 29.7 nM) among nine SPIN homologs previously analyzed [63]. A possible explanation is that SPIN-\textit{delphini} NTD may have evolved to be less dependent on the tight binding of CTD. Interestingly, high-temperature simulations indeed reveal significant differences between coupled binding and folding of NTDs from SPIN-\textit{aureus} and SPIN-\textit{delphini}. As summarized in Figure 2.8, \(~40\%\) of the 40 (17/40) replicas failed to observe full dissociation of SPIN-\textit{delphini} NTD’s during the 250 ns simulations, which is about 3-fold of \(~10\%\) for SPIN-\textit{aureus}. The implication is that SPIN-\textit{delphini} NTD fits the active site of MPO tighter than SPIN-aureus NTD, which would be consistent with disproportionally strong inhibitory function of SPIN-\textit{delphini} despite weakened overall binding affinity.

Further analysis of unfolding and unbinding kinetics and free energy surface reveal that SPIN-\textit{delphini} NTD mainly follow a distinct conformational selection-like mechanism (Figure 2.4), where the NTD tends to gain substantial native β-hairpin structures prior to forming intermolecular interactions with MPO. This is well reflected in unbinding and unfolding kinetics.
On average, the unbinding rates of SPIN-\textit{delphini} NTD (\(\tau_1 = 0.06\) ns, \(\tau_2 = 18.4\) ns, Figure 2.4A) are similar to those of SPIN-\textit{aureus} NTD (\(\tau_1 = 0.12\) ns, \(\tau_2 = 11.4\) ns, Figure 2.3A). However, the unfolding rates of SPIN-\textit{delphini} NTD (\(\tau_1 = 1.31\) ns, \(\tau_2 = 196.97\) ns) are over 10-fold slower than unbinding rates. In addition, SPIN-\textit{delphini} NTD is considerably more folded at 15 ns, with \(Q_{\text{intra}}\sim 0.6\) compared to \(\sim 0.3\) for SPIN-\textit{aureus} NTD. That is, while SPIN-\textit{aureus} NTD unbinds and unfolds to similar levels at a given time (Figure 2.3A), SPIN-\textit{delphini} NTD tends to retain much higher residual structures while it unbinds.

\textbf{Figure 2.4 Cooperative binding and folding of SPIN-\textit{delphini}.} (A) Average intramolecular and intermolecular native contact fractions (\(Q_{\text{inter}}\) and \(Q_{\text{intra}}\)) as a function of simulation time at 450 K. The double exponential fits are plot using dotted lines, with the actual parameters also shown. (B) Pseudo free energy surface as a function of \(Q_{\text{inter}}\) and \(Q_{\text{intra}}\) derived from the transition paths (see Methods). The dashed lines indicate the minimum free energy pathways, with key states labeled. (C) Two major parallel dissociation pathways and key intermediate states for coupled binding and folding of SPIN-\textit{delphini} NTD to human MPO, with SPIN and MPO shown in red and light grey, respectively.
2.3.4 Elevated Pre-folding in Unbound SPIN-delphini NTD

For conformational selection to be an efficient mechanism for coupled binding and folding, there should be high levels residual structures in unbound IDPs [208]. Since SPIN-aureus NTD follows cooperative binding and folding while SPIN-delphini NTD prefers a conformational selection-like mechanism, we further characterized the stability of hairpin-like structures in their unbound states under the physiological conditions. As shown in Figure 2.5, although SPIN-delphini NTD showed slightly faster unfolding rates, it remained more structured than SPIN-aureus NTD. The limiting NTD $Q_{\text{intra}}$ decayed to 0.57 and 0.40 for SPIN-delphini and SPIN-aureus, respectively (see Figure 2.5A). Importantly, the probability distributions of $Q_{\text{intra}}$ show that there is a very high probability for SPIN-delphini NTD to remain partially folded ($Q_{\text{intra}} > 0.5$). Such an elevated residual β-hairpin structures in unbound SPIN-delphini NTD is consistent with the observation of conformational selection-like mechanism of its coupled binding and folding (see above). The more dynamic nature of SPIN-aureus NTD suggests that it depends on specific MPO interactions to facilitate its folding into the β-hairpin structure, thus following a cooperative binding and folding mechanism (Figure 2.3B).

**Figure 2.5 Fraction of native intra-molecular contacts of SPIN-NTDs.** (A) Decay of NTD $Q_{\text{intra}}$ for SPIN-aureus (blue) and SPIN-delphini (red) at 300 K starting with the fully folded
conformation, averaged over 20 replicas of 50-ns simulations. (B) Distributions of NTD $Q_{\text{intra}}$ for two SPIN homologs.

2.4 Conclusions

Extensive atomistic simulations have been performed in explicit solvent to gain a deeper understanding of the structural basis of how SPIN, a protein secreted by *Staphylococcus*, inhibits the activity of human MPO to help evade the neutrophil-mediated host innate immunity. It has been shown that the folded SPIN CTD can bind to MPO even in the absence of the disordered NTD, but the latter is required for the MPO inhibition function. Structural studies further revealed that SPIN NTDs folded into similar β-hairpins upon binding and inserted into the MPO active site for inhibition. Curiously, there is a poor correlation between the MPO binding affinity and inhibition efficacy among different SPIN analogs. The implication is that the conformational properties of unbound SPIN NTDs and their coupled binding and folding likely play central roles in their MPO inhibitory activity.

To further address these questions, we carried out extensive atomistic simulations in explicit solvent using the CHARMM36m force field and studied the structures and interactions of two SPIN homologs, namely, SPIN-\textit{aureus} and SPIN-\textit{delphini}. At an optimal temperature of 450 K, high-temperature simulations reveal that SPIN CTD and NTD binding to MPO follows a decoupled step-wise mechanism, consistent with the experimental observation that CTD is mainly responsible for specific MPO binding [193]. Further 450 K simulations of the unbinding and unfolding of SPIN NTD with CTD restrained in the bound state revealed striking difference in SPIN-\textit{aureus} and SPIN-\textit{delphini}. While coupled binding and folding SPIN-aureus NTD during interaction with MPO is highly cooperative, that of SPIN-\textit{delphini} mainly follows a conformational selection-like mechanism. Both are in contrast to a prevalence of induced folding-
like mechanism previously observed in experimental and computational studies of IDPs that fold into relatively simple structures such as helices and ordered loops [248]. This is an important new insight on coupled binding and folding of IDPs that is likely applicable to other IDPs that require the formation of long-range interactions for specific binding.

We further demonstrate that the mechanistic difference between SPIN-\textit{aureus} and SPIN-\textit{delphini} may be related to the intrinsic conformational properties of their NTDs in the unbound state. Specifically, SPIN-\textit{aureus} NTD is more dynamic and less structured, requiring MPO binding to facilitate its folding and thus a cooperative binding and folding mechanism. On the other hand, SPIN-\textit{delphini} NTD has a much higher propensity to adopt pre-folded hairpin-like conformations, allowing it to follow a conformational selection-like mechanism. As such, SPIN-\textit{delphini} NTD is less dependent on CTD binding to MPO for specific interaction and MPO inhibition. These structural and mechanistic differences could explain why SPIN-\textit{delphini} binds to MPO \textasciitilde19 times weaker than SPIN-\textit{aureus}, but its IC$_{50}$ is only \textasciitilde6 times higher. Taken together, the current atomistic simulations do not only provide new mechanistic principles on coupled binding and folding of IDPs into nontrivial $\beta$-hairpins, but also help to establish the structure-dynamics-function relationship of SPIN homologs. Moreover, it may suggest a new strategy to combating \textit{Staphylococcus} infection, such as by designing drug molecules that could destabilize residual structures in SPIN NTD.
2.5 Supporting Material

Table 2.1 Native intermolecular contacts between SPIN and MPO. *SPIN-aureus* in complex with MPO are shown in col 1\textsuperscript{st}-4\textsuperscript{th} and *SPIN-delphini* in complex with MPO are shown in col 5\textsuperscript{th}-8\textsuperscript{th}. The intermolecular native contacts between SPIN-NTD and MPO are highlighted in grey.

<table>
<thead>
<tr>
<th>SPIN-</th>
<th>MPO</th>
<th>SPIN-</th>
<th>MPO</th>
<th>SPIN-</th>
<th>MPO</th>
<th>SPIN-</th>
<th>MPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>aureus</td>
<td></td>
<td>aureus</td>
<td></td>
<td>delphini</td>
<td></td>
<td>delphini</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>192</td>
<td>47</td>
<td>381</td>
<td>30</td>
<td>265</td>
<td>42</td>
<td>381</td>
</tr>
<tr>
<td>33</td>
<td>346</td>
<td>47</td>
<td>383</td>
<td>30</td>
<td>404</td>
<td>42</td>
<td>382</td>
</tr>
<tr>
<td>34</td>
<td>270</td>
<td>48</td>
<td>381</td>
<td>31</td>
<td>268</td>
<td>42</td>
<td>383</td>
</tr>
<tr>
<td>34</td>
<td>272</td>
<td>49</td>
<td>352</td>
<td>31</td>
<td>269</td>
<td>43</td>
<td>381</td>
</tr>
<tr>
<td>35</td>
<td>265</td>
<td>49</td>
<td>380</td>
<td>31</td>
<td>270</td>
<td>44</td>
<td>352</td>
</tr>
<tr>
<td>35</td>
<td>346</td>
<td>49</td>
<td>381</td>
<td>32</td>
<td>265</td>
<td>44</td>
<td>379</td>
</tr>
<tr>
<td>35</td>
<td>404</td>
<td>49</td>
<td>400</td>
<td>32</td>
<td>266</td>
<td>44</td>
<td>380</td>
</tr>
<tr>
<td>36</td>
<td>268</td>
<td>50</td>
<td>368</td>
<td>32</td>
<td>268</td>
<td>44</td>
<td>381</td>
</tr>
<tr>
<td>36</td>
<td>269</td>
<td>50</td>
<td>379</td>
<td>33</td>
<td>268</td>
<td>44</td>
<td>400</td>
</tr>
<tr>
<td>36</td>
<td>270</td>
<td>51</td>
<td>352</td>
<td>33</td>
<td>282</td>
<td>45</td>
<td>368</td>
</tr>
<tr>
<td>37</td>
<td>265</td>
<td>51</td>
<td>364</td>
<td>33</td>
<td>313</td>
<td>46</td>
<td>352</td>
</tr>
<tr>
<td>37</td>
<td>266</td>
<td>51</td>
<td>365</td>
<td>33</td>
<td>573</td>
<td>46</td>
<td>364</td>
</tr>
<tr>
<td>37</td>
<td>268</td>
<td>51</td>
<td>367</td>
<td>33</td>
<td>577</td>
<td>46</td>
<td>378</td>
</tr>
<tr>
<td>37</td>
<td>405</td>
<td>51</td>
<td>378</td>
<td>33</td>
<td>581</td>
<td>46</td>
<td>400</td>
</tr>
<tr>
<td>38</td>
<td>268</td>
<td>51</td>
<td>400</td>
<td>33</td>
<td>586</td>
<td>47</td>
<td>352</td>
</tr>
<tr>
<td>38</td>
<td>313</td>
<td>52</td>
<td>352</td>
<td>34</td>
<td>268</td>
<td>50</td>
<td>354</td>
</tr>
<tr>
<td>38</td>
<td>577</td>
<td>54</td>
<td>367</td>
<td>34</td>
<td>282</td>
<td>50</td>
<td>355</td>
</tr>
<tr>
<td>38</td>
<td>586</td>
<td>55</td>
<td>354</td>
<td>34</td>
<td>313</td>
<td>50</td>
<td>356</td>
</tr>
<tr>
<td>39</td>
<td>271</td>
<td>55</td>
<td>355</td>
<td>35</td>
<td>386</td>
<td>50</td>
<td>357</td>
</tr>
<tr>
<td>39</td>
<td>282</td>
<td>55</td>
<td>356</td>
<td>35</td>
<td>532</td>
<td>62</td>
<td>359</td>
</tr>
<tr>
<td>39</td>
<td>313</td>
<td>55</td>
<td>357</td>
<td>35</td>
<td>573</td>
<td>63</td>
<td>356</td>
</tr>
<tr>
<td>40</td>
<td>532</td>
<td>68</td>
<td>357</td>
<td>35</td>
<td>576</td>
<td>63</td>
<td>357</td>
</tr>
<tr>
<td>40</td>
<td>576</td>
<td>71</td>
<td>359</td>
<td>35</td>
<td>577</td>
<td>67</td>
<td>356</td>
</tr>
<tr>
<td>40</td>
<td>577</td>
<td>72</td>
<td>356</td>
<td>36</td>
<td>270</td>
<td>70</td>
<td>352</td>
</tr>
<tr>
<td>41</td>
<td>382</td>
<td>72</td>
<td>357</td>
<td>36</td>
<td>382</td>
<td>70</td>
<td>356</td>
</tr>
</tbody>
</table>
Table 2.2 Native intramolecular contacts within NTDs of SPIN-*aureus* and SPIN-*delphini*.

<table>
<thead>
<tr>
<th></th>
<th>SPIN-<em>aureus</em> (NTD)</th>
<th>SPIN-<em>delphini</em> (NTD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>33 - 43</td>
<td>29 - 36</td>
<td></td>
</tr>
<tr>
<td>33 - 44</td>
<td>29 - 38</td>
<td></td>
</tr>
<tr>
<td>33 - 45</td>
<td>29 - 39</td>
<td></td>
</tr>
<tr>
<td>34 - 41</td>
<td>30 - 36</td>
<td></td>
</tr>
<tr>
<td>34 - 43</td>
<td>30 - 37</td>
<td></td>
</tr>
<tr>
<td>34 - 44</td>
<td>30 - 38</td>
<td></td>
</tr>
<tr>
<td>35 - 40</td>
<td>31 - 35</td>
<td></td>
</tr>
<tr>
<td>35 - 41</td>
<td>31 - 36</td>
<td></td>
</tr>
<tr>
<td>35 - 42</td>
<td>32 - 37</td>
<td></td>
</tr>
<tr>
<td>35 - 43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35 - 45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36 - 40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37 - 42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3 Representative dissociation transition segments used in contact probability calculations. Native contacts between SPIN-\textit{aureus}/MPO and SPIN-\textit{delphini}/MPO are selected from the whole 250 ns trajectories and shown in left and right, respectively. Cases where NTD remains bound and folded at the end of simulations are colored in blue, where only the first 50 ns are selected for contact map calculations.

<table>
<thead>
<tr>
<th>Time (ns)</th>
<th>SPIN-\textit{aureus}/MPO</th>
<th>SPIN-\textit{delphini}/MPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-15 ns</td>
<td>Rep</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,4,5,7,9,13,14,15,18,21,22,24,2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7,28,30,31,32,33,34,36,37,38,40</td>
<td></td>
</tr>
<tr>
<td>0-25 ns</td>
<td>Rep</td>
<td>Rep</td>
</tr>
<tr>
<td></td>
<td>2,3,6,8,12,16,19,20,25,26,39</td>
<td></td>
</tr>
<tr>
<td>0-50 ns</td>
<td>Rep</td>
<td>Rep</td>
</tr>
<tr>
<td></td>
<td>3,7,9,10,11,12,13,14,15,16,18,20,2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,2,2,25,27,28,29,32,34,37,38,39,40</td>
<td></td>
</tr>
<tr>
<td>0-100 ns</td>
<td>Rep 10,11,17,23,29,35</td>
<td></td>
</tr>
<tr>
<td>0-80 ns</td>
<td>Rep 1,17,33,35</td>
<td></td>
</tr>
<tr>
<td>0-150 ns</td>
<td>Rep 19,24,36</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.6 Interactions between SPIN and MPO in truncated system. (A) Truncated SPIN/MPO complex system, with included regions colored in grey opaquely and omitted parts colored transparently for MPO. SPIN is colored in red. The MPO loop regions that are not restrained during simulations are colored in cyan. And HEME (not contained in the truncated system) is colored by name and drawn by Bonds. The terminal residues of the selected segments of MPO were labeled out. (B) Dissociation of SPIN-NTD measured by $Q_{inter}$ during pilot of simulations at different temperatures from 400 K to 500 K.
Figure 2.7 Time evolution of NTD native contact fractions. The intermolecular contacts $Q_{\text{inter}}$ (blue) and intramolecular contacts $Q_{\text{intra}}$ (red) are obtained from the 40 replicas of 250 ns simulations of the SPIN-aureus/MPO complex at 450 K.
Figure 2.8 Time evolution of NTD native contact fractions. The intermolecular contacts $Q_{\text{inter}}$ (blue) and intramolecular contacts $Q_{\text{intra}}$ (red) are obtained from the 40 replicas of 250 ns simulations of the SPIN-$\text{delphini}/\text{MPO}$ complex at 450 K.
Figure 2.9 Overlay of the structures of SPIN-delphini in complex with MPO. State I1, \( Q_{\text{inter}} = 1.0, Q_{\text{inter}} \approx 0.6 \), with NTD ensemble transparently drawn (purple) compared to the initial bound and fold structure (green, explicitly drawn). MPO is colored grey and heme is shown in sticks.
CHAPTER 3

RE-BALANCING REPLICA EXCHANGE WITH SOLUTE TEMPERING FOR SAMPLING SYANMIC PROTEIN CONFORMATIONS²

Replica exchange with solute tempering (REST) is a highly effective variant of replica exchange for enhanced sampling in explicit solvent simulations of biomolecules. By scaling the Hamiltonian for a selected “solute” region of the system, REST effectively applies tempering only to the degrees of freedom of interest but not the rest of the system (“solvent”), allowing fewer replicas for covering the same temperature range. A key consideration of REST is how the solute-solvent interactions are scaled together with the solute-solute interactions. Here, we critically evaluate the performance of the latest REST2 protocol for sampling large-scale conformational fluctuations of intrinsically disordered proteins (IDPs). The results show that REST2 promotes artificial protein conformational collapse at high effective temperatures, which seems to be a designed feature originally to promote the sampling of reversible folding of small proteins. The collapse is particularly severe with larger IDPs, leading to replica segregation in the effective temperature space and hindering effective sampling of large-scale conformational changes. We propose that the scaling of the solute-solvent interactions can be treated as free parameters in REST, which can be tuned to control the solute conformational properties (e.g., chain expansion) at different effective temperatures and achieve more effective sampling. To this end, we derive a new REST3 protocol, where the strengths of the solute-solvent van der Waals interactions are recalibrated to reproduce the levels of protein chain expansion at high effective temperatures. The efficiency of REST3 is examined using two IDPs with nontrivial local and long-range structural

features, including the p53 N-terminal domain and the kinase inducible transactivation domain of transcription factor CREB. The results suggest that REST3 leads much more efficient temperature random walk and improved sampling efficiency, which also further reduces the number of replicas required. Nonetheless, our analysis also reveals significant challenges of relying on tempering alone for sampling large-scale conformational fluctuations of disordered proteins. It is likely that more efficient sampling protocols will require incorporating more sophisticated Hamiltonian replica exchange schemes in addition to tempering.

3.1 Introduction

Atomistic simulations of proteins in explicit solvent have greatly benefited from significant recent advances in both GPU-enabled molecular dynamics (MD) algorithms [119, 223-227], which can provide over 100-fold acceleration compared to traditional CPU-based approaches, and accurate general-purposed force fields [122, 123, 228-231], which have been extensively rebalanced for describing folded as well as unfolded proteins. Nonetheless, direct atomistic simulations of large-scale conformational fluctuations and transitions of proteins remain extremely challenging due to the prohibitive computational cost. Therefore, using enhanced sampling techniques are always required to generate statistically representative structural ensembles of proteins [155, 159, 249-255]. This is particularly true in studies of intrinsically disordered proteins (IDPs) [18, 248, 256, 257], an important class of proteins that rely on structural plasticity for function and need to be described using heterogeneous conformational ensembles [136, 258-261]. Arguably, the structural versatile allows IDPs to function diversely and specifically in different cellular activities, like the disorder-to-order conformational change upon binding to biological host or become dynamically folded to compete over non-specific target’s bindings. Generating well-
converged disordered ensembles and understanding their distinctions in various states is essential for establishing the molecular basis of an IDP's function.

For high-dimensional and diffusive processes such as IDP conformational fluctuations, temperature-based replica exchange (T-RE) [153, 155, 262] has proven to be one of the most powerful enhanced sampling techniques in general. In T-RE, multiple copies of the system ("replicas") are simulated in parallel at different temperatures and periodically attempt to exchange simulation temperatures according to Metropolis criteria that maintain the detailed balance. As a result, replicas can undergo random walk in the temperature space to promote barrier crossing and conformational sampling. T-RE has played an instrumental role in recent implicit and explicit solvent force field optimizations, by helping to provide converged conformational distributions of model peptides [122, 263-265]. An important limitation of T-RE for explicit solvent simulations is that the number of replicas required for a given temperature range scales as $\sqrt{N}$, where $N$ is the total number of atoms [155]. The number of replicas required can become prohibitive for explicit solvent simulation of even modestly sized IDPs that require very large water boxes to accommodate the conformational flexibility. Replica exchange with solute tempering (REST) is a powerful variant of T-RE designed to dramatically reduce the number of replicas required [159, 160, 162, 163]. The basic idea is to use Hamiltonian rescaling to achieve effective tempering, such that different regions of the system can be simulated under different (effective) temperatures [159]. This allows only the selected "solute" region (e.g., the protein) to be subjected to tempering while the rest of the system ("solvent") is maintained at the same temperature for all replicas. As a result, only interactions related to the "solute" contribute to the Metropolis criteria of replica exchange and a much smaller number (by 3- to 10-fold) of replicas is required to cover the desired temperature range. For example, only 16 replicas were required in REST simulations of the
disordered N-terminal domain of p53 (p53-NTD, residues 1-61) [37, 231, 266], achieving ~25% acceptance rates from 298 to 500 K. A comparable T-RE simulation of the same solvated system (~72,000 atoms) would require over 100 replicas to ensure sufficient exchange acceptance rates (e.g., ~20%). Importantly, the solute region in REST simulations does not need to include the whole protein, allowing flexible tempering of selected protein regions of interest [163, 164, 267]. REST can be further generalized to only include selected energy terms for tempering to further reduce the number of replicas required [161].

Despite the great flexibility of REST in targeted tempering, how well it translates into better efficiency in sampling large scale protein conformational transitions is not always clear. Disrupting the delicate balance between protein-protein, protein-water and water-water interactions due to the re-scaling of different components of the Hamiltonian can result in elevated energy barriers and actually end up hindering the sampling [268]. In the improved REST2 protocol, the solute-solvent interactions were increasingly weakened at higher effective temperature conditions compared to the original protocol. As a result, the protein remains more compact at higher temperatures, which apparently could allow REST2 to generate more reversible folding transitions of beta-hairpin peptides [160]. In this work, we critically analyzed the efficacy of REST2 for the simulation of disordered protein conformational ensembles using two model IDPs with nontrivial local and long-range structural features. The results reveal an alarmingly strong tendency of REST2 to generate increasingly compact ensembles at higher temperatures, especially for larger and more flexible IDPs. Importantly, the overly compact conformations at high temperatures could rarely be exchanged to low temperatures, leading to the segregation of replicas and greatly reducing the efficiency of random walk in the temperature space. We propose that the scaling factors of solute-solvent van der Waals (vdW) interactions can be treated as free
parameters to control the solute conformational properties at different temperatures for maximizing sampling efficiency. To this end, we design a new REST3 protocol that aims to reproduce the levels of protein chain expansion at high temperatures. The new protocol eliminates the exchange bottleneck and greatly improves the efficiency of random walk in the temperature space, realizing similar conformational convergence with a smaller number of replicas. Nonetheless, our analysis also reveals that tempering has important limitations in driving cooperative local and global conformational transitions of proteins, due to the entropic nature of the associated free energy barriers [269]. It is likely that more efficient sampling protocols will require incorporating more sophisticated Hamiltonian replica exchange schemes in addition to effective tempering [165, 269-272].

3.2 Methods

3.2.1 REST Algorithm

REST allows tempering on only a selected region of interest (e.g., “solute”) while keeping the rest of the system (e.g., “solvent”) at a single temperature. This can significantly reduce the number of DOFs that contribute to the Metropolis criteria of replica exchange, and thus a much smaller number of replicas is needed to cover the same temperature range [159, 160]. The energy of the whole system can be divided into three parts: the solute-solute energy $E_{pp}$, the interaction energy between solute and solvent $E_{pw}$, and the self-interaction energy between the solvent molecules $E_{ww}$. The scaled Hamiltonian at condition $m$ is then given as,

$$E_{m}^{\text{REST}}(\chi) = \lambda_{m}^{pp} E_{pp}(\chi) + \lambda_{m}^{pw} E_{pw}(\chi) + \lambda_{m}^{ww} E_{ww}(\chi),$$ (3.1)

where $\chi$ denotes the system coordinates and $\lambda$’s are the scaling factors of the three components. The scaling factor of solvent-solvent interactions $\lambda_{ww}$ will be kept constant (e.g., 1.0) at all conditions in REST. The solute-solute interactions scaling factor ($\lambda_{m}^{pp}$), on the other hand, is
adjusted for each condition to achieve the desired effective temperature of $T_m$, which is usually exponentially spaced between $T_0$ (the temperature of interest) and $T_{\text{max}}$ with $M$ total replicas,

$$T_m = T_0 \left( \frac{T_{\text{max}}}{T_0} \right)^{m-1}, \quad m = 0, 1, \ldots, M - 1. \quad (3.2)$$

The two existing variants of REST protocols differ in how the solute-solvent energy term is scaled. The original REST protocol [159] is equivalent to having:

$$\lambda_{m}^{\text{pp}} = \beta_m / \beta_0, \quad \lambda_{m}^{\text{pw}} = \frac{(\beta_0 + \beta_m)}{2\beta_0}, \quad \lambda_{m}^{\text{ww}} = 1, \quad (3.3)$$

where $\beta_m = 1/k_BT_m$, $\beta_0 = 1/k_BT_0$ and $k_B$ is the Boltzmann constant. However, it has found that REST showed limited efficiency in driving large conformational transitions and led to exchange bottlenecks between low and high temperature conditions [268]. In the revised REST2 protocol [160],

$$\lambda_{m}^{\text{pp}} = \beta_m / \beta_0, \quad \lambda_{m}^{\text{pw}} = \sqrt{\beta_m / \beta_0}, \quad \lambda_{m}^{\text{ww}} = 1. \quad (3.4)$$

The geometric averaging scaling scheme for setting $\lambda_{m}^{\text{pw}}$ in REST2 weakens the solute-solvent interactions compared to the original REST protocol, and the weakening effect is maximal at condition $T_{\text{max}}$. Note that weakening solute-solvent interactions in REST2 was intentional, for artificially maintaining more compact conformations at high temperatures to facilitate refolding and potentially drive faster reversible folding transitions of beta-hairpins and mini-proteins [160]. It remains unclear whether this is an optimal scaling scheme to simulate the dynamic and potentially more extended ensemble of IDPs. In this work, we propose a new REST3 protocol by introducing an additional calibration factor $\kappa_m$ for vdW interactions between the solute and solvent, and more details can be found in the third session of Results and Discussion.

### 3.2.2 System setup and simulation protocols

Two well-studied IDPs were used to evaluate the REST protocols in this work. The kinase inducible transactivation domain (KID) of transcription factor CREB includes residues 119–146
p53-NTD includes residues 1-61 of p53 (MEEPQ SDPSV EPPLS QETFS DLWKL LPENN VLSPL PSQAM DDLML SPDDI EQWFT EDPGP D). The N- and C- termini were capped with acetyl and N-methyl amide groups, respectively. Both proteins have been extensively characterized by MD, NMR and other biophysical approaches [59, 61, 231, 273-279]. Unless otherwise noted, we ran two independent simulations (control and folding) for each protocol starting from two distinct conformations to evaluate simulation convergence. The control simulation was initiated from a helical state and the folding simulation from an extended state (see Figure 3.9). Except for the control simulations of KID, the proteins were solvated in truncated octahedron boxes, where the shortest distance between two opposite hexagons was ~10.0 nm for KID (folding, with ~25000 water molecules) and ~9.8 nm for p53-NTD (both control and folding, with ~23500 water molecules), respectively. In KID control simulations, the folded initial conformation was solvated in an ~8.5 nm cubic box using about 20000 water molecules. Na+ or Cl- ions were added to neutralize the systems. The periodic boundary conditions were utilized in all simulations, and simulation boxes were large enough to prevent the protein from significant interactions with the periodic images (e.g., see Figure 3.10).

All initial conformations were first energy minimized using a steepest descent algorithm for 6000 steps in GROMACS [280, 281]. 100 ps NVT simulations at 298 K followed by 1 ns of NPT simulations at 298 K and 1 atm were then performed with protein heavy atoms harmonically restrained with a 1000 kJ/mol/nm² force constant. The systems were equilibrated under the same NPT condition for another 1 ns without restraints before the standard MD or REST production simulations. All new simulations in this work were carried out in the a99SB-disp protein force field [123], which is arguably one of the best for simulating dynamic protein conformations [231].
The vdW interactions were smoothly switched off at 1.0 nm, and the long-range electrostatic interactions were treated using the particle mesh Ewald (PME) method [236]. The lengths of hydrogen-containing bonds were constrained using the LINCS algorithm [282].

The REST2 and REST3 simulations were carried out using GROMACS [280, 281] patched with PLUMED 2.3.0 [283-285]. As summarized in Table 3.1, a total of eight REST simulations were performed. These simulations include either 16 replicas spanning an effective temperature of 298 K to 500 K or 8 replicas spanning 298 K to 450 K. Detail configurations of all REST2 and REST3 protocols are given in Table 3.2. Replica exchange was attempted every 2 ps. The total lengths of simulations are 1 µs per replica for KID and 2 µs per replica for p53-NTD. Only folding REST3 simulations were performed for p53-NTD due to the longer simulation length. We have previously performed REST2 simulations of p53-NTD in six force fields [231], where each simulation was ~ 1 µs per replica. Here, the p53-NTD simulations in a99SB-disp with REST2 were extended to 2 µs/replica for both control and folding runs. In addition, two 500 ns standard MD simulations were performed at 500 K for p53-NTD to guide our design of the REST3 protocol.

3.2.3 Analysis

Unless otherwise noted, the first 200 ns of all REST2 and REST3 simulations were excluded in subsequent analysis, which was performed using a combination of GROMACS, in-house scripts, and the MDTraj package [286]. The helical propensities were identified using the standard Dictionary of Secondary Structure of Proteins (DSSP) protocol [287]. For KID simulations, the error bars were estimated from the difference between the folding and control runs. For p53-NTD simulations, the 2 µs folding trajectories were divided into two parts (0.9 µs each excluding the first 200 ns) for error bars calculation. Principal component analysis (PCA) was performed using the python SciKit-learn package [288] to evaluate the sampling efficiency as
well as to visualize the simulated ensembles. For this, snapshots were taken every 100 ps from the entire trajectories to collect all conformations sampled at 298 K (including the first 200 ns). For each peptide, the ensembles generated from REST2 and REST3 simulations were combined and aligned using the backbone atoms before performing the PCA analysis to derive a common set of principal components. The PCA analysis was performed based on the coordinates of Cα atoms. The free energy surfaces were derived directly from the 2D probability distributions along the first two principal components (PCs). All molecular visualizations were done using VMD [237].

3.3 Results and Discussion

3.3.1 Severe over-compaction at high temperatures in REST2

We previously utilized REST2 to evaluate the performance of six latest protein force fields for their ability to describe various local and long-range structures of p53-NTD [231]. These force fields include CHARMM36m (C36m) [122] and its variant C36mw [122], CHARMM22* (C22*) [289-291], Amber ff99SB-ILDN [292, 293] and ff99SB-ILDN-TIP4P-D [292, 294], and a99SB-disp [123]. Curiously, the level of convergence depends strongly on the force field even with ~1 μs/replica simulation time. It was further found that some helical states appeared to be long-lived in C36m and C36mw in REST2, even though these states would readily unfold during standard MD at 298K within 1 μs [231]. The implication is that the rescaling scheme of protein-protein and protein-water interactions in REST2 may result in elevated energy barriers for helix-coil transitions. Here, we further analyze the conformational properties of p53-NTD as a function of effective temperature from these REST2 simulations in all six force fields. As summarized in Figure 3.1 and 3.10, p53-NTD adopts spuriously compact ensembles under high temperature conditions in all six force fields, leading to smaller radius of gyration and end-to-end distance. The helicity level also often increases together with an elevated level of compaction at high temperatures (e.g., in a99SB-
disp, Figure 3.10C). Ensembles for p53-NTD at 500 K are 1.5-2 times more compact than the ensembles generated at the lowest temperature of 298 K, except for the ff99SB-ILDN and C22* that lead to severe over compaction at 298 K. Even for these two later force fields, the scaling scheme of REST2 (Eq. 3.4) drives further compaction above 400 K after a modest chain expansion between 298 to 400 K. Similar observations can also be made on REST2 simulation of the smaller and less flexible KID in a99SB-disp (Figure 3.11), demonstrating that this behavior is force field and protein independent.

**Figure 3.1 Overall protein chain expansion of p53-NTD as a function of temperature in REST2 simulations in six explicit solvent protein force fields.** (A) Average radius of gyration and (B) end-to-end distance. The error bars were calculated as the differences between results from the control and folding REST2 simulations. The first 200 ns was excluded from both the control and folding runs.

Importantly, the collapsed ensembles at high effective temperatures in REST2 are artificial. In Figure 3.2, we compare the REST2 ensemble in a99SB-disp at an effective temperature of 500 K to the one generated by standard MD simulations at 500 K using the unscaled force field. Clearly, the protein does not become severely compact (Figure 3.2A and B, blue vs green trace) or highly helical (Figure 3.2C, blue vs green trace) in standard 500 K MD simulations with the original unscaled Hamiltonian. These results suggest that the scaling scheme $\lambda_m^{pw} = \sqrt{\beta_m/\beta_0}$ in REST2
weakens solute-solvent interactions too much compared to solute-solute interactions and that the level of imbalance becomes more severe at higher temperatures. The observation that IDPs are driven to undergo severe compaction at high effective temperatures likely contributes to the challenges of REST2 protocol in achieving convergence during 1 μs/replica simulations of p53-NTD using C36m or C36mw in the previous study [231].

![Figure 3.2 Conformational properties of p53-NTD using unscaled (MD) and scaled Hamiltonians. In REST2 and REST3 κ_m = 1.0 or 1.06, respectively; see Table 3.2. (A) radius of gyration, (B) end-to-end distance, and (C) residue helicity profile. The distributions at 500 K with the unscaled Hamiltonian were calculated from the last 400 ns of two independent 500-ns runs and the error bars show the difference between these two runs. Results for REST2 and REST3 at 500 K were calculated from the last 1.8 μs of the ensembles and the error bars were estimated as the difference between the first and second 0.9 μs of these ensembles.](image)

3.3.2 Over-compaction at high temperatures can lead to replica segregation in REST2

To evaluate the consequence of compaction at high temperatures, we analyzed the efficiency of replica random walk in the temperature space in both control and folding REST2 simulations of p53-NTD in a99SB-disp (Table 3.1). Inspection of the temperature evolution of all replicas (Figures 3.12 and 3.13) reveals substantial segregation in the temperature space. For example, several replicas including 11, 12, and 15 occupy the high temperature conditions exclusively throughout the folding REST2 simulation (Figure 3.13). We calculated a set of metrics to further evaluate the consequence of segregation in the temperature space, including the average occupancies at the lowest temperature, average effective temperatures and total number of round
trips between the lowest and highest conditions for each replica. In an ideal replica exchange simulation, all replicas undergo efficient random walk in the temperature space and should be equivalent to each other. However, the results, summarized in Figure 3.3, show that all metrics are highly non-uniform for replicas in both control and folding REST2 simulations of p53-NTD. Only subsets of replicas contribute to the ensemble at 298 K and several replicas take few round trips between the lowest and highest temperature conditions. The segregation is particularly severe in the control REST2 simulation. Overall, there are only about 8.82/µs and 9.93/µs round trips per replica in the control and folding runs, respectively. Further analysis showed that inefficient temperature random walk in REST2 hampered conformational samplings, which will be discussed in the following sections. Importantly, the observed replica segregation in the temperature space is not due to inefficiency in neighboring replica exchanges; the acceptance rates are relatively uniform at ~25% in both control and folding REST2 runs. Instead, it should be attributed to the conformational trapping of replicas, particularly at high temperatures. For example, the three replicas trapped at high temperature regions in the folding REST2 simulations (11, 12 and 15) are all trapped in compact conformational states with very small radius of gyration (Figure 3.13, blue traces). With worse temperature mixing and strong potential bias introduced by the folded initial conformation during the control REST2 run, we will mainly focus on the analysis of folding simulations of p53-NTD in the following sessions.
Figure 3.3 Efficiency of replica exchange in REST2 simulations of p53-NTD. (A) occupancy at the lowest temperature ($T_0$, 298 K), (B) average effective temperature, and (C) the number of temperature round trip ($N_{trans}$) per $\mu$s for each replica in control (left) and folding (right) runs. The dashed lines in A and B mark the expected values if perfect random walk occurs ($T_0\% = 0.06$; $<T> = 390.9$ K). The dashed line in C indicates the average $N_{trans}$ per $\mu$s of all 16 replicas, which is 8.82/$\mu$s and 9.93/$\mu$s in the control and folding runs, respectively.

3.3.3 Re-balancing the solute-solvent interactions in REST3

The artificial conformational collapse at high effective temperatures and resulting replica segregation in the temperature space are a direct consequence of the scaling scheme in the REST2 protocol (Eq. 3.4), which is intentional and designed to drive more reversible folding transitions of mini-proteins and beta-hairpins. This limitation may thus be addressed by re-balancing of solute-solvent and solute-solute interactions in the REST protocol (Eq. 3.1). In particular, we propose that the scaling of the solute-solvent interactions in REST does not need to follow prescribed mixing rules such as Eq. 3.2 and Eq. 3.3. Instead, $\lambda_{m}^{PW}$ may be treated as free parameters that can be carefully tuned to control the solute conformational properties (e.g., chain expansion) at different effective temperatures for more efficient sampling. We note that scaling of solute-solute electrostatic interactions is achieved by scaling of all solute partial charges, such that the use of PME requires the solute-solvent electrostatic interactions to follow the geometric mixing rule of
Eq. 3.3. However, the solute-solvent van der Waals (vdW) interactions could be independently tuned. As such, we recast Eq. 3.1 in REST3 as,

\[ E_{\text{REST3}}^m(\chi) = \lambda_{\text{pp}}^m E_{\text{pp}}^m(\chi) + \lambda_{\text{pw}}^m E_{\text{pw}}^m(\chi) + \kappa_m \lambda_{\text{pw}}^m E_{\text{vdw}}^m(\chi) + \lambda_{\text{ww}}^m E_{\text{ww}}^m(\chi), \]  

(3.5)

where \( \kappa_m \) is the additional scaling factor for the solute-solvent vdW interactions at condition \( m \). Setting \( \kappa_m = 1 \) for all conditions recovers the REST2 protocol.

As a proof of concept, we recalibrated the solute-solvent relative interaction strength by matching the conformational properties of p53-NTD at high temperatures, where the artificial compaction is particularly severe in REST2 simulations (Figures 3.10, 3.11, and 3.4). The results show that a modest increase in the solute-solvent vdW interactions with \( \kappa_m = 1.06 \) is sufficient to largely recover both local and global structural properties of p53-NTD at 500 K (Figure 3.2), as compared to results from standard MD using the original unscaled s99SB-disp force field. Based on this, we simply use a linear relationship to set \( \kappa_m \) for other conditions in REST3,

\[ \kappa_m = \begin{cases} 1, & m < 4 \\ (1 + 0.005 \times (m - 3)), & m \geq 4 \end{cases} \]  

(3.6)

Note that \( \kappa_m = 1 \) for the first four conditions, because the artificial collapse of p53-NTD appears to be minimal at low effective temperatures (see Figure 3.1). Detailed values of all scaling factors of the REST3 protocol tested below can be found in Table 3.2.
Figure 3.4 Conformational properties as a function of temperature from REST3 simulations of p53-NTD and KID in a99SB-disp. p53-TAD was shown in left panel calculated from folding simulations, and KID was shown in right panel, calculated from the control and folding simulations. (A) average radius of gyration, (B) distributions of radius of gyration, (C) distribution of end-to-end distance, and (D) average residue helicity profiles. Results from REST2 run are also shown in panel (A) for comparison. See Methods for calculation of error bars.
3.3.4 REST3 eliminates conformational trapping at high temperatures and replica segregation

We first performed one folding REST3 simulation of p53-NTD in a99SB-disp to evaluate the new protocol’s efficiency in driving replica random walk in the temperature space and conformational sampling. As summarized in Figure 3.4 (left column), rescaling of the solute-solvent vdW interactions in REST3 (Eq. 5 and 6) eliminates the artificially increasing compactness at higher temperatures. Note that p53-NTD is a highly dynamic IDP with a large radius of gyration at room temperature (experimental value ~ 2.39 nm [275]); it apparently does not undergo further chain expansion as the temperature is increased from 298 K to 500 K (Figure 3.4A, top left). This is also confirmed in standard MD simulations of p53-NTD at 500 K using the original, unscaled a99SB-disp force field (Figure 3.2). However, the residual helicity does gradually decrease at increasing temperature in REST3 (Figure 3.4D, left). This is in contrast to artificially elevate helicity at high temperatures in REST2 (Figures 3.2C and 3.10C). Therefore, the new REST3 protocol seems to be able to largely recapitulate the expected conformational properties of p53-NTD within the simulation temperature range.

Removing artificial compaction at high temperatures leads to much more efficient replica random walk in REST3. As summarized in Figure 3.14, all REST3 replicas undergo rapid random diffusion throughout the temperature range. Importantly, all replicas are completely free of trapping in compact states and appear to efficiently sample a wide range of conformational states with different levels of compaction (Figure 3.14, blue traces for radius of gyration). Reflecting much more efficient mixing, all replicas in REST3 have similar average temperatures, number of round trips between the lowest and highest temperatures, and contribute similarly to the lowest temperature ensemble (Figure 3.5), which are hallmarks of well-converged replica exchange.
simulations. In particular, the average number of temperature round trips increase to 30.3/µs per replica, compared to 9-10/µs per replica in REST2 runs (see Figures 3.3). Note that the average exchange acceptance rates are essentially identical (~25%) in these REST2 and REST3 runs. The dramatic improvement in the efficiency of random walk in the temperature space is a direct result of eliminating the artificial conformational compaction at higher temperatures (Figures 3.12 and 3.13).

**Figure 3.5 Efficiency of replica exchange in REST2 simulations of p53-NTD using 16 or 8 replicas.** (A) occupancy at the lowest temperature ($T_0$, 298 K), (B) average effective temperature, and (C) the number of temperature round trip ($N_{trans}$) per µs for each replica in control (left) and folding (right) runs. The dashed lines in A and B mark the expected values if perfect random walk occurs ($T_0% = 0.06$; $<T> = 390.9$ K). The dashed line in C indicates the average $N_{trans}$ per µs of all 16 replicas, which is 8.82/µs and 9.93/µs in the control and folding runs, respectively.

We further evaluate the efficacy and transferability of REST3 in preventing conformational trapping and replica segregation using KID, which is a smaller and more structured IDP [59]. As summarized in Figure 3.4 (right column), REST2 also leads to modest artificial conformational compaction and increased structural level for KID. REST3 effectively eliminates the artificial compaction and generates ensembles that show appropriate temperature-dependent
conformational properties at both secondary and global levels. Interestingly, the efficiency of replica exchange in REST2 does not appear to deteriorate as much compared to the case of p53-NTD (Figure 3.3 vs 3.15). As such, the replica exchange efficiency does not benefit significantly from the new REST3 protocol. This observation should not be surprising, considering that KID is a smaller IDP and adopts much more compact conformational ensembles even at 298 K (Figure 3.4A). The artificial over compaction is much less severe than the highly flexible p53-NTD. Nonetheless, as will be discussed later, ensembles derived from the folding and control REST3 runs appear to be better converged at all temperatures. Taken together, REST3 is an effective protocol free of artificial compaction at high effective temperatures and much more suitable for the simulation of dynamic protein conformational ensembles compared to REST2 in general.

3.3.5 Conformational sampling and convergence of REST2 and REST3

Figure 3.6 2D probability distributions of the radius of gyration and end-to-end distance of KID at 298 K. Properties were derived from independent REST2 and REST3 control and
folding simulations (Table 3.1). The first 200 ns of all trajectories were excluded as the equilibrium phase. All probability distributions were first converted into the free energy surface before plotting.

Examination of the evolution of the conformations of replicas in REST2 vs. REST3 as a function of simulation time clearly demonstrates more efficient exploration of different conformational spaces in REST3. For example, all REST3 replicas reversibly sample diverse conformations with a wide range of radius of gyration throughout the 2 µs simulation time (Figure 3.14, blue traces). In contrast, many REST2 replicas rarely sample conformations of different sizes (Figure 3.12 and 3.13, blue traces) and the lack of sampling by individual replicas is particularly prominent in the control run. Nonetheless, it is also known that the apparent convergence of the lowest temperature ensemble of a replica exchange simulation can arise due to replica mixing even in the presence of substantial replica segregation and conformational trapping [269, 295]. Indeed, with 1 or 2 µs sampling per replica, the current REST2 and REST3 simulations generated highly consistent disordered ensembles at 298 K for both KID and p53-NTD in the a99SB-disp force field, with similar local and global structural properties such as radius of gyration distributions and residue helicity profiles (Figure 3.16). Closer inspection of the structural ensembles generated by independent folding and control runs, however, suggests that certain structural properties are better converged in REST3 simulations. For example, substantial differences persist between the 2D distributions of the radius of gyration and end-to-end distance derived from folding and control REST2 runs (Figure 3.6 A vs B), while the distributions are far more consistent from REST2 runs (Figure 3.6 C vs D).

We further assess the sampling and convergence of REST2 and REST3 by performing the PCA analysis of the disordered ensembles at 298 K, which allows the heterogenous ensembles to be projected in principal axes with the largest variances (see Methods). The results confirm a
higher level of consistency between ensembles of KID generated by folding and control runs of REST3 compared to those of REST2, but at the same time reveal substantial residual differences even for REST3 simulations (Figure 3.7A). The later observation is somewhat surprising, considering the small size of KID and the apparent convergence of various 1D distributions with 1 µs per replica sampling time (e.g., Figure 3.16). On the other hand, this really illustrates the critical challenge of sampling disordered protein conformational ensembles in explicit solvent even for modest-sized IDPs [261]. For p53-NTD, REST3 appears to be capable to sample broader metastable states (e.g., areas with free energy < 0.5 kcal/mol, Figure 3.15B), but the ensembles generated by the two REST protocols are much more similar compared to those of KID. Besides the longer sampling time of 2 µs per replica, the better convergence of p53-NTD ensembles may also be attributed to its more extended and less structured nature despite the longer sequence (Figure 3.16).
Figure 3.7 Distribution of conformational ensembles at 298 K generated by REST2 and REST3. Properties were derived from the control and folding simulations for KID (A) and folding simulation for p53-NTD (B) in the a99SB-disp force field. The ensembles were projected along the same first two principal components derived from PCA analysis of all conformations sampled from both REST2 and REST3 runs of each individual protein. All probability distributions were first converted into the free energy surface before plotting.

To directly compare the convergence rates, we further analyze the evolution of the disordered ensembles at 298 K as a function of REST2 and REST3 simulation time. For KID, we focus on the sampling of transient partial helices, as analysis of the global structural properties
already reveals improved convergence with REST3 (e.g., Figure 3.6 and 3.7). For this, we calculated the probabilities of partial helices as identified by their starting positions and lengths. The results, summarized in Figure 3.17, again highlight the challenge of sampling disordered protein conformations, as the helical substate distributions remain significantly different between 1-μs control and folding runs for both REST2 and REST3. Even longer sampling time is likely required for achieving better convergence at the helical substate level. Nonetheless, there appears to be an overall improvement in the convergence of the populations of helical substates in REST3 compared to REST2. For example, the populations of partial helices starting around residue 1 (red arrows in Figure 3.17 B and D) reach plateaus by ~600 ns during the REST3 folding run, but continue to increase and only reach similar (converged) levels near the end of the 1 μs REST2 folding run. Similar observations can be made for some helical states in the control runs (e.g., purple arrows in Figure 3.17 A and C). For p53-NTD, the 2 μs REST2 and REST3 simulations yield highly similar ensembles, as reflected in various 1D distributions (Figure 3.16) and PCA analysis (Figure 3.7B). Examination of the evolution of the disordered ensemble at 298 K as a function of simulation time suggests that REST2 and REST3 are quite similar in achieving the overall convergence (Figure 3.8 A vs B), despite substantial differences in conformational coverage of individual replicas (Figure 3.12 and 3.13 vs 3.14, blue traces). This again highlights how replica mixing itself can lead to apparent convergence in replica exchange simulations. Nonetheless, it is also clear that REST3 is able to sample broader free energy basin with shorter simulation times (e.g., regions with free energy < 0.5 kcal/mol). Taken together, the detailed conformational analyses above supports that improved replica exchange efficiency and conformational sampling of individual replicas in REST3 can indeed further improve the convergence of the disordered ensembles generated at the lowest temperature condition.
Figure 3.8 Evolution of the conformational ensemble of p53-NTD at 298 K. (A) REST2, (B) REST3 (16-replica), and (C) REST3 (8-replica) folding simulations in a99SB-disp. The ensembles were projected along the same first two principal components derived from PCA analysis of all conformations sampled from REST2 and REST3 runs. All probability distributions were first converted into free energy surface before plotting.

3.3.6 REST3 can further reduce the required number of replicas

The observation that REST2 does not appear to suffer significantly from conformational trapping and replica segregation at high effective temperatures suggests that smaller number of replicas may be sufficient for the simulation of p53-NTD using the REST3 protocol. Here, we
evaluate a REST3 protocol with only 8 replicas spanning 298 K to 450 K (see Table 3.2) and perform a folding simulation of p53-NTD for 2 µs per replica. Although the larger temperature spacing leads to smaller average replica exchange acceptance (~6.5%), all replicas can efficiently explore the whole temperature space and are free of conformational trapping (Figure 3.18). As a result, all replicas contribute significantly to the lowest temperature ensemble and have similar average temperatures (Figure 3.5B), which are hallmarks of well-converged replica exchange simulations. The average temperature round-trip transition rate is 26.7/µs per replica, compared to that of ~30.3/µs for the 16-replica REST3 run. Importantly, the resulting disordered ensemble at 298 K is highly similar to those generated by both REST2 and REST3 protocols (Figure 3.19). Comparison of the 298 K ensembles as a function of simulation time (Figure 3.8C) further demonstrates that the convergence rate of the 8-replica REST3 protocol is similar to that of the 16-replica REST3 protocol and superior to that of the 16-replica REST2 protocol. Taken together, it is evident that the better-balanced solute-solute and solute-solvent interactions in REST3 allow it to effectively sample disordered protein conformational ensembles with much less computational sources.

3.4 Conclusions

REST is one of the most effective enhanced sampling approaches for biomolecular simulations that is particularly suitable for explicit solvent simulations by dramatically reducing the number of replicas required. In this work, we critically evaluated the replica exchange and sampling efficiency of the latest REST2 protocol for simulation of disordered protein ensembles. Our results reveal that REST2 leads to artificial conformational compaction at high effective temperatures. This is due to the imbalanced scaling of solute-solute and solute-solvent interactions, which was originally designed to promote reversible folding of mini-proteins and beta-hairpins.
Importantly, the artificial over compaction at high temperatures leads to conformational trapping and segregation of replicas in the temperature space during REST2 simulations. These problems can be particularly severe for highly flexible IDPs such as p53-NTD, where some replicas can completely fail to contribute to the lowest temperature ensemble.

We propose that the scaling of solute-solvent vDW interactions can be treated as a free parameter in REST protocols, which can be optimized to carefully control the conformational properties of the protein solute at various temperature conditions and achieve more efficient conformational sampling. To this end, we describe a new REST3 protocol that has been tuned to generate conformational distributions throughout the temperature range similar to those expected using the unscaled Hamiltonian. With a better balance between solute-solute and solute-solvent interactions, REST3 completely eliminates conformational trapping at high temperatures and the resulting replica segregation problem as observed in REST2. All REST3 replicas can undergo highly efficient random walk in the temperature space and sample a broad range of conformational space. As a result, REST3 can further improve the convergence of the disordered protein conformational ensemble at both local and global structure levels. Importantly, REST2 often relies on replica mixing for achieving the apparent convergence instead of true conformational sampling of individual replicas. We also show that REST3 allows one to further reduce the number of replicas required for sufficient sampling. Our tests suggest that at least about half of the computational sources can be saved compared to REST2 when simulating the moderate-sized IDPs such as p53-NTD.

Our critical analyses of REST2 and REST3 protocols further reveal significant challenges of relying on tempering alone for sampling large-scale conformational fluctuations of disordered proteins. Sampling of many detailed conformational substates has limited convergence even with
1 to 2 $\mu$s per replica sampling and significant differences often persist between independent runs initiated from drastically different initial conformations. This challenge can be attributed to the entropic nature of the barriers involved in folding transitions of local and/or global (transient) structures [269], such that the transition rates will only depend weakly on the simulation temperature. As such, further improvement of the REST protocol will likely require incorporating of more advanced Hamiltonian replica exchange in addition to tempering.
### 3.5 Supporting Material

#### Table 3.1. Summary of REST2 and REST3 simulations.

<table>
<thead>
<tr>
<th>Protocols</th>
<th>REST2 (298K-500K, 16-rep)</th>
<th>REST3 (298K-500K, 16-rep)</th>
<th>REST3 (298K-450K, 8-rep)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Folding</td>
<td>Control</td>
</tr>
<tr>
<td>KID</td>
<td>1 µs</td>
<td>1 µs</td>
<td>1 µs</td>
</tr>
<tr>
<td>p53-NTD</td>
<td>2 µs</td>
<td>2 µs</td>
<td>2 µs</td>
</tr>
</tbody>
</table>

#### Table 3.2 Parameters for REST2 and REST3 protocols. Note that for REST3, $\lambda_{m}^{pw,vdw} = K_{m} \lambda_{m}^{pw}$

<table>
<thead>
<tr>
<th>Rep</th>
<th>Temp (K)</th>
<th>REST2</th>
<th>$\kappa_{m}$</th>
<th>Rep</th>
<th>Temp (K)</th>
<th>REST3</th>
<th>$\kappa_{m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>298</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>392</td>
<td>0.759</td>
<td>0.871</td>
</tr>
<tr>
<td>1</td>
<td>308</td>
<td>0.966</td>
<td>0.983</td>
<td>9</td>
<td>406</td>
<td>0.733</td>
<td>0.856</td>
</tr>
<tr>
<td>2</td>
<td>319</td>
<td>0.933</td>
<td>0.966</td>
<td>10</td>
<td>420</td>
<td>0.708</td>
<td>0.841</td>
</tr>
<tr>
<td>3</td>
<td>330</td>
<td>0.902</td>
<td>0.950</td>
<td>11</td>
<td>435</td>
<td>0.684</td>
<td>0.827</td>
</tr>
<tr>
<td>4</td>
<td>342</td>
<td>0.871</td>
<td>0.933</td>
<td>12</td>
<td>450</td>
<td>0.661</td>
<td>0.813</td>
</tr>
<tr>
<td>5</td>
<td>354</td>
<td>0.841</td>
<td>0.917</td>
<td>13</td>
<td>466</td>
<td>0.639</td>
<td>0.799</td>
</tr>
<tr>
<td>6</td>
<td>366</td>
<td>0.813</td>
<td>0.902</td>
<td>14</td>
<td>483</td>
<td>0.617</td>
<td>0.785</td>
</tr>
<tr>
<td>7</td>
<td>379</td>
<td>0.785</td>
<td>0.886</td>
<td>15</td>
<td>500</td>
<td>0.596</td>
<td>0.772</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rep</th>
<th>Temp (K)</th>
<th>REST3</th>
<th>$\kappa_{m}$</th>
<th>Rep</th>
<th>Temp (K)</th>
<th>REST3</th>
<th>$\kappa_{m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>298</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>392</td>
<td>0.759</td>
<td>0.823 1.025</td>
</tr>
<tr>
<td>1</td>
<td>308</td>
<td>0.966</td>
<td>0.983</td>
<td>9</td>
<td>406</td>
<td>0.733</td>
<td>0.882 1.030</td>
</tr>
<tr>
<td>2</td>
<td>319</td>
<td>0.933</td>
<td>0.966</td>
<td>10</td>
<td>420</td>
<td>0.708</td>
<td>0.871 1.035</td>
</tr>
<tr>
<td>3</td>
<td>330</td>
<td>0.902</td>
<td>0.950</td>
<td>11</td>
<td>435</td>
<td>0.684</td>
<td>0.860 1.040</td>
</tr>
<tr>
<td>4</td>
<td>342</td>
<td>0.871</td>
<td>0.938</td>
<td>12</td>
<td>450</td>
<td>0.661</td>
<td>0.850 1.045</td>
</tr>
<tr>
<td>5</td>
<td>354</td>
<td>0.841</td>
<td>0.926</td>
<td>13</td>
<td>466</td>
<td>0.639</td>
<td>0.839 1.050</td>
</tr>
<tr>
<td>6</td>
<td>366</td>
<td>0.813</td>
<td>0.915</td>
<td>14</td>
<td>483</td>
<td>0.617</td>
<td>0.829 1.055</td>
</tr>
<tr>
<td>7</td>
<td>379</td>
<td>0.785</td>
<td>0.904</td>
<td>15</td>
<td>500</td>
<td>0.596</td>
<td>0.818 1.060</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rep</th>
<th>Temp (K)</th>
<th>REST3</th>
<th>$\kappa_{m}$</th>
<th>Rep</th>
<th>Temp (K)</th>
<th>REST3</th>
<th>$\kappa_{m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>298</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>377</td>
<td>0.790</td>
<td>0.907 1.020</td>
</tr>
<tr>
<td>1</td>
<td>316</td>
<td>0.943</td>
<td>0.971</td>
<td>5</td>
<td>400</td>
<td>0.745</td>
<td>0.886 1.027</td>
</tr>
<tr>
<td>2</td>
<td>335</td>
<td>0.889</td>
<td>0.946</td>
<td>6</td>
<td>424</td>
<td>0.702</td>
<td>0.867 1.035</td>
</tr>
<tr>
<td>3</td>
<td>355</td>
<td>0.838</td>
<td>0.924</td>
<td>7</td>
<td>450</td>
<td>0.662</td>
<td>0.850 1.045</td>
</tr>
<tr>
<td>Model Peptides</td>
<td>KID</td>
<td>p53 TAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>![Image]</td>
<td>![Image]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Folding</td>
<td>![Image]</td>
<td>![Image]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.9 The initial conformations of KID and p53-NTD for folding and control simulations. The N-terminal region of the peptide is colored red and blue for the C-terminal region (following the red-white-blue, RWB color scheme).
Figure 3.10 Distributions of conformational properties of p53-NTD. (A) Radius of gyration and (B) end-to-end distance and (C) residue helicity profiles under all 16 conditions derived from REST2 folding and control simulations in four selected protein force fields. The colors for different effective temperature conditions are shown at bottom.

Figure 3.11 Distributions of conformational properties of KID. (A) Radius of gyration and (B) end-to-end distance and (C) residue helicity profiles under all 16 conditions derived from REST2 control and folding simulations in a99sb-disp. Coloring of different effective temperature conditions is the same as Figure 3.10.
Figure 3.12 Evolution of temperature and radius of gyration for each of the 16 replicas of the REST2 control simulation of p53-NTD in a99SB-disp. The temperature and radius of gyration are colored in grey and blue, respectively. The grey dashed line marks the radius of gyration from SAXS (2.39 nm).
Figure 3.13 Evolution of temperature and radius of gyration for each of the 16 replicas of the REST2 folding simulation of p53-NTD in a99SB-disp. The temperature and radius of gyration are colored in grey and blue, respectively. The grey dashed line marks the radius of gyration from SAXS (2.39 nm).
Figure 3.14 Evolution of temperature and radius of gyration for each of the 16 replicas of the REST3 folding simulation of p53-NTD in a99SB-disp. The temperature and radius of gyration are colored in grey and blue, respectively. The grey dashed line marks the radius of gyration from SAXS (2.39 nm).
Figure 3.15 Efficiency of replica exchange of KID simulations with REST2 and REST3 control and folding simulations. As reflected in occupancy at the lowest temperature (top, $T_0$, 298 K), average effective temperature, and the number of temperature round trip per $\mu$s for each replica. The average $N_{\text{trans}}/\mu$s of all 16 replicas are 59.2/\mu s (A), 79.2/\mu s (B), 63.8/\mu s (C), and 62.9/\mu s (D) in REST2 and REST3 control and folding runs, respectively.
Figure 3.16 Conformational distributions of KID and p53-NTD from REST2 and REST3 simulations in a99SB-disp. (A) Radius of gyration, (B) end-to-end distance, (C) and average residue helicity profile. The error bars were calculated from the differences between the folding and control runs for KID. The folding trajectories for p53-NTD (excluding the first 200 ns) were divided into two parts for error bar calculations.
Figure 3.17 2D probability distributions of the starting position and length of partial helices of KID at 298 K as a function of REST2 and REST3 simulation time. All probability distributions were first converted into free energy surface before plotting. Purple and red arrows highlight two helical substates with significant difference convergence rates in REST2 and REST3.
Figure 3.18 Evolution of temperature and radius of gyration for all replicas of the REST3 folding simulation of p53-NTD in a99SB-disp using 8 replicas. The grey dashed line marks the radius of gyration from SAXS (2.39 nm).

Figure 3.19 Properties of conformational ensembles as a function of effective temperature derived from RSET2 and REST3 simulations of p53-NTD (folding simulation) in a99SB-disp. These properties include average radius of gyration (A) and end-to-end distance (B). (C) Average residue helicity profiles at 298 K for p53-NTD calculated by three protocols. The 2 µs folding simulation excluding the first 200 ns trajectory was divided into two parts for error bar calculations.
Intrinsically disordered proteins (IDPs) play crucial roles in cellular regulatory networks and are now recognized to often remain highly dynamic even in specific interactions and assemblies. Accurate description of these dynamic interactions is extremely challenging using atomistic simulations because of the prohibitive computational cost. Efficient coarse-grained approaches could offer an effective solution to overcome this bottleneck if they could provide an accurate description of key local and global properties of IDPs in both unbound and bound states. The recently developed hybrid resolution (HyRes) protein model has been shown to be capable of providing semi-quantitative description of the secondary structure propensities of IDPs. Here, we show that greatly improved description of global structures and transient interactions can be achieved by introducing a solvent-accessible surface area-based implicit solvent term followed with re-optimization of effective interaction strengths. The new model, termed HyRes II, can semi-quantitatively reproduce a wide range of local and global structure properties of a set of IDPs of various lengths and complexities. It can also distinguish the level of compaction between folded proteins and IDPs. In particular, applied to the disordered N-terminal transactivation domain (TAD) of tumor suppressor p53, HyRes II is able to recapitulate various nontrivial structural properties compared to experimental results, some of them to a level of accuracy that is almost comparable to results from atomistic explicit solvent simulations. Furthermore, we demonstrate

---

that HyRes II can be used to simulate the dynamic interactions of TAD with the DNA-binding domain of p53, generating structural ensembles that are highly consistent with existing NMR data. We anticipate that HyRes II will provide an efficient and relatively reliable tool towards accurate coarse-grained simulations of dynamic protein interactions.

4.1 Introduction

Intrinsically disordered proteins (IDPs) remain highly dynamic under physiological conditions and rely on a lack of stably folded tertiary structure to support versatile functions [1, 4, 7, 18, 66, 296-299]. Sequence analyses reveal that near one-half of cellular signaling-associated proteins contain substantial disordered regions [296]. By existing at the boundary of marginal stability and instability, the disordered conformational ensembles of IDPs are poised to respond sensitively to various signals, including post-translational modifications and bindings of small and macromolecules [2, 66, 298, 300, 301]. As such, these proteins are key components of the regulatory networks that dictate virtually all aspects of cellular decision-making. Examples are also emerging to show that IDPs can remain unstructured even in specific complexes and functional assemblies [302-308]. Such a dynamic mode of specific protein interactions seems much more prevalent than previously thought [261, 309-311]. Not surprisingly, mis-signaling and mis-regulation of IDPs are frequently involved in human diseases, including cancers, diabetes, and neurodegenerative diseases [296, 312-315].

A particularly important prototypical IDP is the tumor suppressor p53, which plays crucial roles in the cellular response to various genotoxic stress conditions [316, 317]. p53 is one of the most mutated genes in cancers [318-320]. The intrinsically disordered N-terminal transactivation domain of p53 [321, 322] (residues 1-61, p53 TAD) mediates its interactions with numerous signaling pathways that control the p53 activities [323-325]. Mutations, phosphorylation, and other
post-translational modifications on p53 TAD can directly influence the homeostasis of p53 [37, 88, 319, 326-331]. For example, phosphorylation on T55 and S46 augments p53’s binding to human transcription factor TFIIH to upregulate gene expression [88]. Mutations of L22Q and W23S prevent transactivation of many genes by inhibiting the interaction between p53 TAD and Med17 [332], a key coactivator in general transcriptions [333]. We recently showed that some cancer-associated mutations in p53 TAD, such as N29K/N30D, significantly perturbed the balance of p53 interactions with key activation and degradation regulators, likely by direct modulation of the disordered ensemble of TAD itself [37]. Two short amphipathic motifs within p53 TAD, known as AD1 (residues 18 to 26) and AD2 (residues 40 to 55), are believed to play key roles in p53 interactions [305, 321]. Yet, TAD can remain highly dynamic even in specific complexes [305, 334, 335]. Recent nuclear magnetic resonance (NMR) studies further show that p53 TAD transiently interacts with its DNA binding domain (DBD, residues 95-291) and competes with non-specific DNA interactions [89], which enhances target site discrimination and is further regulated by phosphorylation of TAD [88]. Therefore, there is a critical need to better understand the TAD conformational dynamics in both bound and unbound states, in order to establish a deeper understanding on how the disordered ensemble may contribute to various physiological and pathophysiological processes.

A critical challenge in studying IDP conformation and dynamics is that traditional ensemble-based techniques, such as NMR, small-angle X-ray scattering (SAXS) and Förster resonance energy transfer (FRET), are inadequate in describing detailed conformational fluctuations [336]. The reason is that only averaged properties can be measured, which alone cannot resolve the heterogeneous ensemble and transient structural features [258, 260, 337, 338]. There is thus an urgent need for molecular modeling and simulation to help provide these missing
details. In particular, physics-based atomistic simulations can probe the conformational states and dynamics of proteins with high temporal and spatial resolutions [136, 248, 259, 264, 339]. A major limitation, however, for atomistic simulations of IDPs can be the prohibitive computational cost, especially for larger IDPs or for describing their dynamic interactions. This has motivated significant interests in developing coarse-grained (CG) models that can dramatically reduce the computational cost and extend the accessible length and time scales [339].

Traditional CG models developed for folded proteins tend to generate overly compact ensembles for IDPs, and they require significant modification and reparameterization to accurately capture the more extended and dynamic conformational features [339]. By modifying the nonbonded Lennard-Jones potential in the MARTINI force field [340, 341] to match the transfer free energy of phase separation measured by experiments, Zakarya et al developed a more general CG force field that can be applied to simulate reversible liquid-liquid phase transitions (LLPS) [342]. The self-organized polymer coarse-grained model for IDPs (SOP-IDP) was parameterized to reproduce the experimental SAXS data and radii of gyration ($R_g$) of IDPs [168]. AWSEM-IDP, developed by Papoian and co-workers, directly includes a new $R_g$ potential term to control the peptide chain collapse in IDP simulations [169]. An important limitation, however, is that these models are not transferrable for both structured and disordered proteins. For this, the Zhang group recently developed a maximum entropy optimized force field (MOFF), which rebalanced the $C\alpha$-only model originally described by Hummer and co-workers for the dynamic protein complexes [343] to reproduce $R_g$ of a set of IDPs [178, 344, 345]. MOFF is able to correctly distinguish the level of compaction of folded and disordered proteins and successfully applied to simulate the LLPS properties of protein HP1 [178]. Similar $C\alpha$-only CG models have found important success in simulation of phase behavior of IDPs [346-348]. Nevertheless, a key challenge of these CG
models, particularly Cα-only ones, is to describe sequence-dependent residual structures of IDPs or how these structural properties depend on protein binding. As illustrated by studies of p53 summarized above, residual local and long-range structures can play central roles in IDPs’ ability to support specific functions in cellular signaling and regulation [18, 248, 261, 349].

Recognizing the importance of characterizing various local and global structural features of IDPs and the limitation of the low-resolution CG force field in describing IDP local structures [350], we have developed a hybrid resolution model, named HyRes, which includes atomistic backbones and intermediate resolution sidechains [170]. HyRes is able to semi-quantitatively capture the secondary structure details, while qualitatively describing dynamic long-range interactions of IDPs. The model has been applied to simulate the dynamic interactions of p53 TAD with cyclophilin D (CypD), successfully predicting key molecular features that agree well with NMR experiments [351]. Nonetheless, HyRes was not originally designed for stand-alone simulations of IDPs, but for driving more efficient atomistic sampling within the framework of multi-scaled enhanced sampling (MSES), where accurate descriptions of nonspecific interactions are not crucial. Indeed, further analysis of the structural properties of p53 TAD in unbound and CypD-bound states revealed some important limitations of HyRes for independent simulation of IDPs. In particular, the model over-estimates the protein-protein interactions and tends to generate overly compact structural ensembles. Further improvement is needed to better balance the protein-protein and (implicit) protein-water interactions in HyRes for accurate simulations of dynamic protein conformations and interactions.

Here, we describe the development of a greatly improved model, HyRes II, for more accurate description of both the local and long-range conformational properties of IDPs as well as their dynamic interactions. This is mainly achieved by introducing a solvent-accessible surface
area (SASA)-based implicit solvent, followed by careful re-parameterization of sidechain and backbone interactions. The new model was evaluated using a set of globular and disordered proteins and shown to be able to describe a wide range of conformational properties of IDPs, some of them at a level of accuracy almost comparable to atomistic explicit solvent protein force fields. Importantly, we showed that HyRes II could be effectively applied to simulate the dynamic interaction of p53 TAD with the folded DBD and provide a molecular-level description of the nonspecific TAD/DBD interaction that is highly consistent with available NMR data [88, 89]

4.2 Methods

4.2.1 Original HyRes Protein Model

As illustrated in Figure 4.1, the HyRes model includes a united-atom representation of the protein backbone, while the side chains are represented using five (TRP), three (HIS, PHE, TYR), two (LYS, ARG) or one bead (all others except Gly) (also see Table 4.1). Note that higher-resolution mapping is adopted for amino acids containing ring structures to achieve more accurate geometric representations.

![Figure 4.1 HyRes representation of amino acids.](image)

The effective energy function of HyRes includes eight bonded and nonbonded terms [170]:

\[ U_{\text{HyRes}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{improper}} + U_{\text{CMAP}} + U_{\text{LJ}} + U_{\text{Hbond}} + U_{\text{elec}} \]  

\[ U_{\text{bond}} = \sum_{\text{bonds}} k_b (b - b_0)^2 \]  

\[ U_{\text{angle}} = \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \]  

81
\[ U_{\text{dihedral}} = \sum_{\text{dihedrals}} k_\chi [1 + \cos(n\chi - \delta)] \quad (4.4) \]

\[ U_{\text{improper}} = \sum_{\text{impropers}} k_\psi (\psi - \psi_0)^2 \quad (4.5) \]

\[ U_{\text{CMAP}} = \sum_{\text{non-Gly, non-Pro residues}} U_{\text{CMAP}}(\varphi, \psi) \quad (4.6) \]

\[ U_{\text{Hbond}} = \sum_{\text{Hbonds}} \varepsilon_{\text{HB}} \left[ 5 \left( \frac{\sigma}{r} \right)^{12} - 6 \left( \frac{\sigma}{r} \right)^{10} \right] \cos^4 \theta_{\text{AHD}} \quad (4.7) \]

\[ U_{\text{LJ}} = \sum_{i,j} \varepsilon_{i,j} \left[ \left( \frac{r_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{ij}^{\text{min}}}{r_{ij}} \right)^{6} \right] \quad (4.8) \]

\[ \text{with } r_{ij}^{\text{min}} = \frac{r_{i}^{\text{min}} + r_{j}^{\text{min}}}{2}, \text{ and } \varepsilon_{i,j} = \sqrt{\varepsilon_i \varepsilon_j} \]

\[ U_{\text{elec}} = \sum_{\text{chrg.pairs}} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} e^{-\frac{r_{ij}}{b}} \quad (4.9) \]

Parameters for the bond and angle terms involving sidechains (Eqs. 4.2 and 4.3) were assigned to reproduce the averages and fluctuations of corresponding pseudo bonds and angles derived from atomistic simulations of dipeptides in the GBSW implicit solvent protein force field [352]. The parameters for the bonded terms involving the backbone were taken directly from the CHARMM19 united atom force field [353]. The hydrogen-bonding (Hbond) interaction (Eq. 4.7) was only applied to the backbone, which was tuned together with the dihedral and CMAP terms (Eqs. 4.4 and 4.6) to properly describe secondary structure propensities of model IDPs. The van de Waals (vdW) parameters (Eq. 4.8) of the backbone were also taken from CHARMM19, except that \( \varepsilon_i \) values were scaled down to 30% of original values to account for a lack of solvent screening of dispersion interactions. For the sidechains, the vdW radii were assigned to reproduce the total effective sidechain volumes from atomistic simulations, while \( \varepsilon_i \) were first assigned based on the Miyazawa and Jernigan’s (MJ) statistical potentials [354] and then scaled to reproduce the free energy profiles of sidechain pairwise interactions in explicit solvent from a previous work [355].
All backbone atoms and sidechain beads are charge neutral except for the terminal beads of charged sidechains. The electrostatic interactions (Eq. 4.9) were described by a Debye-Hückel-type potential, where the effective dielectric constant $\varepsilon_r$ were assigned to 20 to balance the short and long-range interactions using a set of (EK)$_{25}$ peptides [356]. The Debye screening length D is set to $\sim 8$ Å, corresponding to the physiological ionic strength of $\sim 0.15$ M.

### 4.2.2 HyRes II: design and optimization

A key limitation of HyRes is that none of the three non-bonded interaction terms is able to capture context-dependence of effective protein-protein interactions, which are crucial for accurate simulation of large-scale protein conformational transitions and protein-protein interactions [357, 358]. In particular, the over-compaction bias of HyRes may be directly attributed to inadequate description of the solvent screening effects in the treatment of non-bonded interactions. To address this limitation, we introduced a SASA-based implicit solvent term in HyRes II,

$$U_{\text{HyRes II}} = U_{\text{HyRes}} + U_{\text{SASA}}.$$  \hspace{1cm} (4.10)

SASA implicit solvent assumes that the solvation effects mainly come from the first solvation shell and are thus approximately proportional to surface area of the solute [358]. The simplicity and efficiency of SASA implicit solvent make it particularly suitable for CG modeling. Specifically, we use an efficient pair-wise de-screening approximation of SASA as implemented in CHARMM [359],

$$U_{\text{SASA}} = \sum_l \sigma_l A_l.$$  \hspace{1cm} (4.11)

The SASA term is only applied to the atomistic backbone in HyRes II. Key parameters include the effective surface tension coefficients ($\sigma_l$) for carbonyl groups (C=O) and amide groups (NH). The final optimized values are -0.010 kcal $\cdot$ mol$^{-1} \cdot$ Å$^{-2}$ (see Results and Discussion).

With the introduction of SASA implicit solvent, other nonbonded interactions in HyRes II need to be re-optimized to achieve an adequate balance of various type interactions (e.g., backbone
hydrogen bonding vs. electrostatic vs. vdW) for accurate simulation of protein conformational equilibrium and dynamic interactions. For this, we deployed a multi-stage iterative optimization strategy. The first stage focuses on balancing the competition between backbone conformational entropy, nonspecific interactions and SASA, using poly-glycine peptides (Gly_{10} and Gly_{20}) that do not contain any sidechain beads. As both Gly_{10} and Gly_{20} are random coils, the main objective at this stage is to assign the backbone SASA parameters (\sigma_i) to reproduce the chain dimensions derived from atomistic simulations using the ABSINTH implicit solvent force field [360]. In the second stage, we balance hydrogen bonding and other interactions (including CMAP) of the backbone to recapitulate the residue helical propensities of model peptides, mainly (AAQAA)_3 and KID (see Table 4.4 for sequences). The key parameter tuned in this stage is the hydrogen bond strength \epsilon_{HB}. Finally, we rebalanced the side chain vdW and electrostatic interactions to reproduce the chain expansion (e.g., R_g or end-to-end distance) of a set of disordered peptides, including the (EK)_{25} series, (AAQAA)_3 and KID (see Table 4.4). In the final stage, the main parameters to be optimized include the uniform scaling of vdW parameters \epsilon_i of side chain beads and the relative dielectric constant \epsilon_r for electrostatic interactions. These three stages are repeated iteratively if needed to achieve sufficient balance (and cancellation of errors) for accurate description of both local and long-range structural properties of all model peptides.

4.2.3 Folded and disordered proteins for training and testing

The sequences of all proteins used in the optimization and benchmark are provided in Table 4.4. (AAQAA)_3 is widely used in protein force field optimizations with residue helicity well characterized by NMR (~50% at 270 K) [361]. The (EK)_{25} peptide series includes 30 variants with different Glu and Lys mixing patterns as characterized by parameter \kappa, which ranges from 0 (well mixed) to 1 (fully segregated charges). The 30 variants were used to examine if HyRes II captures
the balance of vdW and electrostatic interactions appropriately compared to ABSINTH [356, 362].

KID, the 28-residue kinase inducible transactivation domain of transcription factor CREB, is a prototypical model IDP that has been extensively characterized by experiments and MD simulations [59, 61]. It contains two helical elements, with helicity estimated to be about 50-60% (residues 120-129) and 10-15% (residues 134-144) [279]. The 61-residue p53 TAD contains multiple residual helical regions and transient long-range intrachain contacts, and has been well studied by multiple experimental techniques including NMR, SAXS and smFRET [231]. The large amount of available experimental data makes p53 TAD a particularly useful system for assessing the quality of HyRes II in describing nontrivial local and long-range structural properties of IDPs.

Three folded proteins were used to assess the ability of HyRes II to distinguish the overall compactness of globular and disordered proteins (Table 4.4). The Villin headpiece is a 36-residue protein domain containing three stable helixes (PDB ID: 1VII) [363]; protein A is a 46-residue 3-helix bundle protein (PDB ID: 1BDC) [364]; and protein G B1 domain contains 56 residues with the mixed sheet and helix structures (PDB ID: 3GB1) [365]. Note that these folded proteins were chosen because their sequence lengths are comparable to the IDP test cases, allowing one to evaluate if HyRes II recapitulates the different degrees of compaction between folded proteins and IDPs. Last but not least, we critically evaluate HyRes II’s ability to describe dynamic interactions of p53 TAD and DBD (see below).

4.2.4 CG simulation protocols

All HyRes and HyRes II simulations were carried out using CHARMM [223, 366]. All CG simulations were simulated at constant temperature using a 2-fs integration step with Langevin thermostat. The temperature is 300 K unless otherwise noted. All bonds involving hydrogen atoms were constrained by SHAKE algorithm [367]. Nonbonded interactions were smoothly switched
off from 1.6 nm to 1.8 nm. All peptides/proteins were capped with an acetyl group at the N-terminus and N-methyl amide at the C-terminus. Unless otherwise noted, the parametrization and testing equilibrium simulations were performed with two independent replicas, initiated from distinct conformations (e.g., helical and unfolded) (see Figure 4.8). This allows rigorous assessment of the convergence of simulated ensembles. The unfolded conformations were generated using short simulations at 800 K. All simulations last 400 ns, which was found to provide well-converged ensembles for all IDPs evaluated. Two additional (AAQAA)$_3$ simulations were carried out at 270 K to for direct comparison with the NMR results acquired at the same temperature [361].

Dynamic interactions of p53 TAD/DBD was simulated using a monomeric construct of p53 that includes residues 1 through 291 using both HyRes and HyRes II. We first generated a set of initial structures with the disordered TAD and protein rich region (PRR, residue 62 to 94) in various extended and compact conformations, while the DBD structure was taken from PDB ID: 3KMD (Figure 4.8, residue 95 to 291). The fully extended TAD conformations (s5-s8) were generated using a high temperature simulation at 800 K. These initial structures were used to start eight independent 400-ns MD simulations at 300 K, where the C$_\alpha$ atoms of DBD were harmonically restrained with a force constant of 1.0 kcal·mol$^{-1}$·Å$^{-2}$.

4.2.5 Atomistic simulations

Monte Carlo (MC) simulations of free Gly$_{10}$ and Gly$_{20}$ were performed in the ABSINTH [368] implicit solvent as implemented in the CAMPARI [369] package. The ABSINTH model has been shown to be capable of accurately describing the global structural preferences of IDPs [259, 356, 360] and thus provides a good benchmark reference for simulation of poly-Glycine peptides. The simulation setup follows the ‘file abs3.2_opls.prm’ protocol in CAMPARI. The peptides were
placed in droplets of 80 Å and 160 Å radius for Gly$_{10}$ and Gly$_{20}$ respectively. Two simulations were carried out at 300 K under NVT conditions lasting 41,000,000 MC steps, with the first 1,000,000 steps excluded in the analysis.

Atomistic ensembles of (AAQAA)$_3$, KID, and p53 TAD were generated using replica exchange with solute tempering (REST2) simulations [159] using the a99SB-disp explicit solvent force field [123]. Our previous benchmark study showed that a99SB-disp was particularly suitable for IDP simulations, successfully recapitulating a wide range of nontrivial local and global structural properties of p53 TAD [231]. We used a similar REST2 protocol for simulations of (AAQAA)$_3$ and KID as in our previous works [231]. Briefly, simulations were carried out using GROMACS 5.1.4 patched with PLUMED 2.3.0 [283] under NVT conditions. Only the protein region was subjected to tempering, which was realized by scaling the solute-solute ($\lambda$) and solute-solvent ($\sqrt{\lambda}$) interactions correspondingly. 16 replicas were used to cover effective temperatures ranging from 298 to 500 K with $\lambda$ values set to 1.00, 0.97, 0.93, 0.90, 0.87, 0.84, 0.81, 0.79, 0.76, 0.73, 0.71, 0.68, 0.66, 0.64, 0.62, and 0.60. Replica exchange was attempted every 2 ps, with average acceptance rates of ~20%. The vdW interactions were cut off at 1.0 nm and the long-range electrostatic interactions were treated using the particle mesh Ewald (PME) method [370, 371]. The lengths of hydrogen-containing bonds were constrained using the LINCS algorithm [282]. The REST2 simulation of (AAQAA)$_3$ was initiated from an extended conformation and lasted 1.0 $\mu$s per replica. For KID, two independent REST2 runs were performed from extended and folded structures, respectively, for rigorous assessment of convergence. Each REST2 run lasted 1.0 $\mu$s per replica for KID. Atomistic ensembles of p53 TAD were taken from our previous REST2 simulations [231] but extended to ~3.0 $\mu$s per replica in both independent REST2 simulations initiated from extended and helical conformations, respectively.
4.2.6 Analysis

Unless otherwise noted, only the last 300 ns of all 400-ns coarse-grained simulation trajectories were included in subsequent analysis, which was performed using a combination of CHARMM, in-house scripts, and the python codes with MDTraj package [286]. The helical propensities were identified by the standard Dictionary of Secondary Structure of Proteins (DSSP) protocol [372]. Only $\alpha$-helices here were considered. $R_g$ values from HyRes/HyRes II simulations were uniformly shifted up by 4 Å to roughly account for larger side chain beads when comparing with atomistic sampling results except for poly-glycine peptides without sidechains, as done previously [170]. The error bars shown were estimated from differences between results calculated from two independent simulations initiated from distinct initial structures (Figure 4.8). For REST2 atomistic simulations in the a99SB-disp force field [123], structure ensembles were collected by including conformations sampled at effective temperature of 298 K ($\lambda = 1.00$). The first 200 ns were excluded in KID and (AAQAA)$_3$ trajectories (1.0 $\mu$s in total) and the first 300 ns were excluded in p53 TAD trajectories (3.0 $\mu$s in total). The error bars shown in all figures were estimated based on the difference between results from the two independent REST2 runs for KID and p53 TAD. For (AAQAA)$_3$, the 800-ns trajectory was divided into two 400-ns trajectories for estimating the error bars. All molecular visualizations were done using VMD [373].

Transient long-range structural properties of p53 TAD in the unbound state as well as its dynamic interaction with DBD are mainly evaluated using available paramagnetic relaxation enhancement (PRE) experimental data, which are back calculated from the simulated ensembles as,

$$\frac{I_{ox}}{I_{red}} = \frac{R_g \exp (-R_g^3 \mu t)}{R_g + R_g^3 \mu t}$$

(4.12)
where $R_{2}^{sp} = K < r^{-6} > (4\tau_{c} + \frac{3\tau_{c}}{1+\omega_{H}\tau_{c}})$, $r$ is the distance to the paramagnetic spin label, which was approximated by the Ca-Ca distance, and $<>$ denotes ensemble averaging. All other parameters were kept consistent with the experimental conditions ($K = 1.23 \times 10^{-32} cm^6 s^{-2}$, $\omega_{H} = 600 MHz$, $\tau_{c} = 3.3 ns$, $R_{2} = 16 s^{-1}$, $t = 9.8 ms$).

Principal component analysis (PCA) was performed using python SciKit-learn package [288], to evaluate the sampling convergence as well as to visualize the simulated ensembles. For this, snapshots were taken every 100 ps from the entire 400 ns trajectories to collect all conformations sampled. For unbound TAD, the ensembles generated from HyRes and HyRes II simulations (4 trajectories in total) were combined and then aligned using the backbone atoms before performing the PCA analysis. For p53 TAD/DBD simulations, all sixteen trajectories (eight each from HyRes and HyRes II runs) were aligned using the backbone atoms of the structured DBD core region (residues 95 to 291) before PCA analysis based on the coordinates of Cα atoms of residue 1-61. This analysis captures both internal conformations of TAD and how it interacts DBD. The free energy surfaces shown were derived directly from the 2D probability distributions along the first two principal components (PCs).

4.3 Results and Discussion

4.3.1 Balance of IDP chain dimensions and secondary structures

We applied a multi-stage optimization strategy to balance various interactions in the HyRes II force field for accurate description of conformational equilibrium of IDPs (see Methods). The balance of interactions was mainly examined based on the conformational distributions of a diverse set of model peptides. The convergence of all simulated ensembles was assessed by comparing the results generated from two independent replicas starting from distinct initial conformations (see Figure 4.8). All conformational properties examined, such as $R_g$, end-to-end distances, and helicity
profiles, are well converged within 400 ns, as indicated by the small error bars for all properties shown in this work. The high level of convergence greatly facilitates the tuning and assessment of the HyRes II protein force field.

The balance of backbone conformational entropy, nonspecific interactions within the backbone, and backbone-solvent interactions may be best captured in the conformational properties of Gly$_{10}$ and Gly$_{20}$, which do not have sidechains. In the original HyRes model, the strengths of backbone-backbone vdW interactions are scaled down to 30% of the CHARMM19 values to roughly account for the effects of solvent screening of dispersion interactions. However, accurate description of the dimension of poly-glycine peptides proved to be very difficult, and HyRes significantly underestimates both $R_g$ and end-to-end distances compared to results from ABSINTH (Figure 4.9 and Table 4.5). Including the SASA implicit solvent term helps to promote the population of extended conformations where the backbone has increased solvent exposure, especially for the longer Gly$_{20}$. However, we found that fully reproducing the polymer chain dimensions obtained from ABSINTH simulations would require very negative $\sigma_l$ values (e.g., < - 0.04 kcal \cdot mol^{-1} \cdot \text{Å}^{-2} ), which would abolish secondary structures of IDPs and require unphysically large effective backbone HB strength to restore correct helical propensities. This likely reflects an inherent limitation of the simplified representation in CG models, where the peptide conformational entropy is likely underestimated. As it will be shown later, the final choice of $\sigma_l = -0.010$ kcal \cdot mol^{-1} \cdot \text{Å}^{-2}$ represents a compromise that allows relatively accurate description of chain dimensions and local structural propensities for most IDPs (as well as folded proteins) examined in this work, particularly larger ones such as p53 TAD, even though poly-glycine peptides in HyRes II remain significantly more compact compared to atomistic ABSINTH ensembles (Fig. 4.9, Table 4.5).
In Figure 4.2, we examine the ability of HyRes II to recapitulate secondary structural propensities of (AAQAA)$_3$ and KID. Note that the secondary structural properties are impacted by several factors, including backbone dihedral potentials (including CMAP), vdW interactions, backbone hydrogen bonding and SASA terms. In HyRes II, the hydrogen bonding strength is increased slightly to -2.0 kcal/mol to compensate for the effect of SASA (which promotes the coil state). At the same time, we found it necessary to further scale down the sidechain vdW interaction strengths. The reason is that, in the original HyRes model, we assigned side chain vdW scaling to reproduce the pair-wise free energy profiles derived from the CHARMM22 explicit solvent force field, which is known to generate overly compact structures for IDPs [122]. In HyRes II, $\varepsilon_i$ in Eq. 8 was further uniformly reduced to 50% of the original values (see Table 4.1). With these optimizations, HyRes II retains the ability to semi-quantitatively predict the helicity of (AAQAA)$_3$ at 270 K (Figure 4.2A), modestly over-estimating the helicity (~75% vs 50%). Application to KID, HyRes II shows significant improvement compared to HyRes, predicting 30–45% helicity in $\alpha_1$ and 10–20 % in $\alpha_2$ at 300 K, in excellent agreement with the previous NMR study [374] as well as atomistic ensembles calculated using a99SB-disp (Figure 4.2B). Importantly, the overall chain dimensions of both (AAQAA)$_3$ and KID are also improved in HyRes II, which are more comparable to results from atomistic simulations in a99SB-disp (Table 4.2, Figure 4.10). We also note that all CG simulations are well-converged within 400 ns, while much larger errors persist atomistic ensembles of KID despite a very substantial total of 32 $\mu$s sampling time using REST2 (Figure 4.2B, red vs blue trace).
Figure 4.2 Residue helicity profiles of (AAQAA)$_3$ and KID calculated using HyRes and HyRes II models. (A) (AAQAA)$_3$ simulated at 270 K and (B) KID simulated at 300 K. The experimental or atomistic simulation results are shown as references. The experimental NMR result of (AAQAA)$_3$ was obtained from ref [361].

Table 4.1 HyRes II vdW interaction strength parameters for all side chain beads.

<table>
<thead>
<tr>
<th>Residue</th>
<th>CG bead</th>
<th>$\epsilon_i$ (kcal/mol)</th>
<th>$\epsilon_i$ (1-4) (kcal/mol)</th>
<th>Residue</th>
<th>CG bead</th>
<th>$\epsilon_i$ (kcal/mol)</th>
<th>$\epsilon_i$ (1-4) (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>CB</td>
<td>-0.1540</td>
<td>-0.3080</td>
<td>ARG</td>
<td>CC</td>
<td>-0.0675</td>
<td>-0.1350</td>
</tr>
<tr>
<td>VAL</td>
<td>CB</td>
<td>-0.3100</td>
<td>-0.6200</td>
<td>HIS</td>
<td>CB</td>
<td>-0.0540</td>
<td>-0.1080</td>
</tr>
<tr>
<td>LEU</td>
<td>CB</td>
<td>-0.4500</td>
<td>-0.9000</td>
<td>HIS</td>
<td>CC</td>
<td>-0.0405</td>
<td>-0.0810</td>
</tr>
<tr>
<td>ILE</td>
<td>CB</td>
<td>-0.3860</td>
<td>-0.7720</td>
<td>HIS</td>
<td>CD</td>
<td>-0.0405</td>
<td>-0.0810</td>
</tr>
<tr>
<td>MET</td>
<td>CB</td>
<td>-0.3180</td>
<td>-0.6360</td>
<td>PHE</td>
<td>CB</td>
<td>-0.1100</td>
<td>-0.2200</td>
</tr>
<tr>
<td>ASN</td>
<td>CB</td>
<td>-0.0900</td>
<td>-0.1800</td>
<td>PHE</td>
<td>CC</td>
<td>-0.1100</td>
<td>-0.2200</td>
</tr>
<tr>
<td>ASP</td>
<td>CB</td>
<td>-0.0740</td>
<td>-0.1480</td>
<td>PHE</td>
<td>CD</td>
<td>-0.1100</td>
<td>-0.2200</td>
</tr>
<tr>
<td>GLN</td>
<td>CB</td>
<td>-0.1020</td>
<td>-0.2040</td>
<td>TYR</td>
<td>CB</td>
<td>-0.0984</td>
<td>-0.1968</td>
</tr>
<tr>
<td>GLU</td>
<td>CB</td>
<td>-0.0700</td>
<td>-0.1400</td>
<td>TYR</td>
<td>CD</td>
<td>-0.0984</td>
<td>-0.1968</td>
</tr>
<tr>
<td>CYS</td>
<td>CB</td>
<td>-0.2660</td>
<td>-0.5320</td>
<td>TYR</td>
<td>CC</td>
<td>-0.0492</td>
<td>-0.0984</td>
</tr>
</tbody>
</table>
Table 4.2 Ensemble averaged properties of model peptides calculated from CG and AT simulations. The end-to-end distances ($d_{e2e}$) and $R_g$ of (AAQAA)$_3$ and KID are calculated from simulations using HyRes, HyRes II, and a99SB-disp. All values are in Å.

<table>
<thead>
<tr>
<th>Model</th>
<th>Peptide</th>
<th>$&lt;R_g&gt;$</th>
<th>$&lt;d_{e2e}&gt;$</th>
<th>Peptide</th>
<th>$&lt;R_g&gt;$</th>
<th>$&lt;d_{e2e}&gt;$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HyRes</td>
<td>(AAQAA)$_3$</td>
<td>11.1</td>
<td>15.3</td>
<td></td>
<td>13.2</td>
<td>18.0</td>
</tr>
<tr>
<td>HyRes II</td>
<td></td>
<td>11.8</td>
<td>19.1</td>
<td>KID</td>
<td>15.0</td>
<td>22.9</td>
</tr>
<tr>
<td>a99SB-disp</td>
<td></td>
<td>10.4</td>
<td>26.2</td>
<td></td>
<td>13.9</td>
<td>31.3</td>
</tr>
</tbody>
</table>

4.3.2 Balance between short-range and electrostatic interactions

Compared to globular proteins, IDPs are deficient in hydrophobic groups and enriched in polar and charged residues. Their conformations are highly sensitive to the balance between short-range vdW/HB and long-range electrostatic interactions. Herein, to better quantify if proper balance between short- and long-ranged nonspecific interactions has been achieved in HyRes II, we calculated the ensemble averaged $R_g$ values of the set of 30 (EK)$_{25}$ variants with a wide range of charge mixing patterns and conformational preference (Table 4.4). The key parameter to be tuned here is the effective dielectric constant $\varepsilon_r$, which may require further refinement to be...
compatible with the uniformly weakened vdW interactions of sidechain beads. This proved unnecessary in HyRes II. As summarized in Figure 4.3, HyRes II reproduces the trend of the ensemble averaged $R_g$, which generally decreases with increasing $\kappa$ (that is, increasing charge segregation). Even though HyRes II derived ensembles remain systematically more compact compared to atomistic ones calculated using ABSINTH [356], they are more extended compared to the original model and more consistent with atomistic results. These results demonstrate that HyRes II provides an appropriate balance between various short-range vdW/HB and (long-range) electrostatic interactions in governing the disordered protein ensembles.

![Figure 4.3](image_url)

**Figure 4.3** Averaged $R_g$ as a function of the charge patterning parameter $\kappa$ for 30 variants of the (EK)$_{25}$ peptide. ABSINTH results are taken from ref [170].
4.3.3 Ability of HyRes II in describing nontrivial local and long-range structures of p53 TAD

As discussed above, p53 TAD is a biologically important IDP with a rich set of nontrivial local and long-range structure features that have been extensively characterized by simulation and experiments [231]. These properties make p53 TAD an ideal and challenging system for rigorous evaluation of protein force fields for IDP simulations. Here we evaluate the ability of HyRes II to describe p53 TAD disordered ensemble, comparing to results from explicit solvent simulations using the a99SB-disp force field. As summarized in Figure 4.4, HyRes II provides substantial improvements compared to HyRes in the description of the overall chain dimensions. While HyRes yields conformational ensembles with severe over-compaction, HyRes II is able to sample much broader distributions that are more consistent with atomistic distributions using a99SB-disp, even though HyRes II clearly still under-estimates the level of overall peptide chain extension of p53 TAD compared to a99SB-disp (e.g., Figure 4.4B). As discussed above, the persisting bias towards modest over-compaction in HyRes II seems to derive from the inherent limitation of CG representation and necessary for retaining the ability to semi-quantitatively describing residual helical propensities of IDPs. Nonetheless, we note that a99SB-disp slightly over-estimates the chain expansion of p53 TAD compared to experimental results. For example, a slightly longer p53-
TAD peptide (residues 1–73) was found to have a Stokes radius of 23.8 Å [375], smaller than ~26 Å given by a99SB-disp for the shorter 61-residue TAD peptide. Instead, the average $R_g$ of ~23.6 Å predicted by HyRes II (Table 4.3) seems more consistent with the experimental measurement.

The 2D histogram of $R_g$ and $d_{e2e}$ shown in Figure 4.11 further illustrate that HyRes II dramatically reduces the systematic bias of HyRes towards over-compaction and properly captures large conformational fluctuations of p53-TAD. Importantly, HyRes II retains the ability to semi-quantitatively describe the residual helical propensities of p53 TAD, yielding a comparable distribution compared to those from the original HyRes and a99SB-disp (Figure 4.4C). The result also agrees well with NMR studies showing an approximate 10% helical among residues 17-29 ($\alpha 1$), with additional partial helices in the AD2 region (e.g., residue 40-44 and 48-52) [321, 376].

![Figure 4.5 PRE results of p53 TAD from experiments, AT simulations, and CG simulations. Calculated (lines) and experimental [375] (gray bars) PRE effects induced by paramagnetic spin](image)
labeling at residues (A) 28, (B) 39, (C) 7 and (D) 61. Error bars were calculated from two independent REST2 simulations (a99SB-disp) or MD runs (HyRes and HyRes II) (see Methods).

The ability of HyRes II to describe transient long-range structural features of p53 TAD is further evaluated by comparing the back calculated PRE effects with the experimental data [375]. PRE combined with site-specific spin labeling is a powerful NMR technique for detecting transient interactions between the spin label and the rest of the whole protein up to 35 Å [377, 378]. Specifically for the unbound p53 TAD, experimental PRE data are available for spin labelling at four sites, namely, D7, E28, A39, and D61 [379]. In Figure 4.5, we compare the calculated PRE effects from HyRes, HyRes II and a99SB-disp simulations with the experimental values. The results show that HyRes II significantly improves the ability to capture the overall PRE profiles compared to HyRes, which systematically over-estimates the PRE effects due to the strong bias towards over-compaction (Figure 4.4, green vs red traces). Compared to a99SB-disp, HyRes II appears to slightly over-estimate the PRE effects, especially for spin labels near either N- (residue 7) or C-terminus (residue 61) (Figure 4.5 C & D). This is consistent with the observation that HyRes II still slightly underestimates the p53 chain expansion (Figure 4.4 A and B). Nonetheless, HyRes II seems to be able to capture many of the fine features in the PRE profiles, to a level that is comparable to a99SB-disp. For example, experimental PRE data measured with spin labelling at residue 28 reveal transient interactions with N-terminal residues 12-15 as well as C-terminal residues 43-48 and 52-58, which are all well captured in the HyRes II ensembles (Figure 4.5A). Interestingly, atomistic ensembles generated using a99SB-disp largely fail to detect transient interactions between residue 28 with the C-terminal motifs (Figure 4.5A). The ability of HyRes II to recapitulate nontrivial fine features of PRE profiles is noteworthy, which has not been demonstrated by another CG protein model to the best of our knowledge. Furthermore, we note that generation of the atomistic ensembles here required a total of 48 μs sampling time (3 μs per
replica), which took about three months on 16 Nvidia 1080Ti GPUs. In contrast, the efficiency of HyRes II allows superior convergence to be achieved with standard MD simulations of 400 ns each, which takes less than 100 hours using 1 Xeon CPUs or a few hours on a single Nvidia 1080Ti GPU. This reflects a ~10,000-fold improvement in computational efficiency, while without substantial apparent reduction of accuracy in the simulated disordered ensembles. The ability of HyRes II to accurately capture both local and long-range nontrivial structural features of p53 TAD strongly suggests its suitability for general simulation of disordered protein states of IDPs and potentially their dynamic interactions with other proteins.

**4.3.4 Ability of HyRes II to distinguish compactness of structured and disordered proteins**

We further examined the ability of HyRes II to properly distinguish the level of chain compaction between folded and unstructured proteins, to ensure that including the SASA implicit solvent term and subsequent optimization does not lead to a strong and systematic bias towards extended conformations. As shown in Figure 4.12, even though all three proteins quickly lost most of their secondary and tertiary structures, reflected in the large RSMD values, all three small proteins largely retain their level of compactness during 400-ns HyRes and HyRes II simulations. Clearly, the coarse-grained treatment of sidechains is still far from adequate to describe the precise packing and interactions of the protein required for maintaining β-sheet and tertiary structures. Interestingly, snapshots taken from simulation trajectories (Figure 4.12D) suggest that some global topological features appear to be retained during HyRes and HyRes II simulations. Importantly, HyRes II does not over-destabilize the folded structures such that they would all behave like IDPs. As summarized in Table 4.3, HyRes II generates structural ensembles with average $R_g$ in good agreement with the PDB structures or experimental measurements for both folded and unfolded
proteins. In particular, $R_g$ for the 56-residue protein G B1 domain is only ~13 Å compared to that of ~24 Å for 61-residue p53 TAD despite the similar sequence lengths (Table 4.3). Therefore, HyRes II is able to distinguish the overall levels of compaction between structured and disordered proteins.

Table 4.3 Ensemble averaged radius of gyration ($R_g$) of ordered and disordered peptides simulated using HyRes and HyRes II (in Å). Reference values were calculated from the PDB structures (folded proteins) or SAXS (p53, KID).

<table>
<thead>
<tr>
<th>PDBID</th>
<th>Length</th>
<th>Ref</th>
<th>HyRes</th>
<th>HyRes II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1VII</td>
<td>36</td>
<td>9.4 [363]</td>
<td>9.5 ± 0.02</td>
<td>11.0 ± 0.04</td>
</tr>
<tr>
<td>1BDC</td>
<td>46</td>
<td>10.2 [364]</td>
<td>10.4 ± 0.02</td>
<td>12.8 ± 0.07</td>
</tr>
<tr>
<td>3GB1</td>
<td>56</td>
<td>10.8 [365]</td>
<td>11.0 ± 0.02</td>
<td>13.3 ± 0.03</td>
</tr>
<tr>
<td>KID</td>
<td>28</td>
<td>~ 11 [61]</td>
<td>9.2 ± 0.1</td>
<td>11.0 ± 0.1</td>
</tr>
<tr>
<td>TAD</td>
<td>61</td>
<td>~23.8 [231, 380]</td>
<td>16.4</td>
<td>23.6 ± 0.3</td>
</tr>
</tbody>
</table>

4.3.5 Dynamic interactions of p53 TAD/DBD

As discussed above, dynamic interactions of p53 TAD and DBD play a key role in promoting and regulating specific interactions of p53 with DNA [88, 89]. In particular, NMR experiments have identified that two short amphipathic motifs, AD1 and AD2, on TAD as binding hot spots. However, the molecular details of the transient and dynamic p53 TAD/DBD interactions are poorly understood. For example, it is not known if AD1 and AD2 bind alternatively or simultaneously to DBD or if they compete for similar or overlapping sites on DBD. Also, it is unclear that whether these two motifs would become more helical in the nonspecific DBD bound state, which was speculated to be the case in order to for p53 TAD to mimic the DNA double helix
conformation [88, 89]. The efficiency and improved accuracy of HyRes II may now allow us to address these important questions regarding dynamic interactions of p53 TAD/DBD.

Direct simulations of the structure, dynamics and interaction of IDPs tethered to a folded domain have proven to be much more challenging than those of unbound IDPs [250, 381]. In this work, we first generated a diverse set of eight initial TAD conformations and performed independent simulations to ensure adequate sampling of the relevant conformational space of disordered TAD (and PRR) in the presence of folded DBD (see Methods). Figure 4.6 summarizes all the conformational space sampled by p53 TAD tethered to DBD in the eight replicas as well as regions sampled by each individual run, projected on the same first two PCs derived from all conformations sampled. Note that the PCA analysis was applied to the disordered TAD segment aligned using DBD, such that the projections shown in Figure 4.6 captures both the (internal) structure and binding configuration of TAD. The results show that none of the eight independent simulations alone covers all relevant conformational space within 400 ns (Figure 4.6A vs B). However, spaces sampled by these eight simulations are broad and overlap; together they seem to provide a complete coverage of the accessible conformational space of p53 TAD in the presence of DBD. In particular, while some replicas mainly sample one of major states (namely, state I, II, III, IV and V in Figure 4.6A), all replicas sampled multiple major states within 400 ns simulations. Furthermore, distributions of TAD interacting residues on DBD from the eight replicas are also highly consistent (Figure 4.13), further supporting the convergence in the current HyRes II sampling of p53 TAD/DBD interactions.
Figure 4.6 PCA analysis of the conformational space sampled by p53 TAD in dynamic interactions with DBD. (A) From all HyRes II trajectories combined, and (B) from each independent HyRes II run. Snapshots sampled from all sixteen HyRes and HyRes II simulations were used to derive the first two PCs, which are then used to calculate all the projections. The heat map shows the free energy (in kcal/mol) derived from the 2D probability distributions. (C) Representative structures for the five major basins identified. The structures are drawn in cartoon, with DBD colored in cyan and TAD and PRR in grey except that AD1 and AD2 are highlighted in green and red, respectively. The DNA binding surface of DBD is highlighted in blue, with key TAD-interacting residues from NMR [89] T118 and V112 (on L1 loop), G245 and N247 (on L3 loop), and A276, G279, R280, R283 (on C-terminal helix) shown as yellow vdW beads. The paramagnetic spin labeling site S121 is colored in orange. S15 and P58 are also labelled in red and green, respectively.

In contrast, HyRes simulations using the same setup failed to achieve converged sampling of the structure and interaction of p53 TAD with DBD. The conformational spaces sampled by independent replicas are fragmented, each restricted to much smaller regions (Figure 4.14). This is likely a direct consequence of over-stabilized non-specific interactions in HyRes, such that the elevated ruggedness in the energy landscape significantly limits the conformational diffusion. Critically, despite being initiated from the same starting conformations, HyRes simulations largely fail to sample about half of the major states observed in HyRes II runs (left side of Figure 4.6A).
As will be shown later, these states mainly involve transient interactions of AD1 in addition to AD2. The lack of convergence in HyRes simulations is also reflected in varying distributions of TAD contacting residues from independent simulations (Figure 4.15). Comparison of Figures 4.13 and S8 illustrates that TAD can broadly sample on the DNA binding surface of DBD in HyRes II, but it is restricted to much smaller regions in the original HyRes model. These observations strongly support that HyRes II is superior not only in describing transient long-range structural features of IDPs but also in simulation of the dynamic interactions IDPs with other proteins.

Further inspection of the overall conformational space suggests two major modes of p53 TAD/DBD dynamic interactions. Mode 1 includes major states I, II, and V on the PC1-PC2 space (Figure 4.6A). It mainly involves dynamic interactions dominated by AD2, which binds directly on top of the known DNA-binding interface of DBD [382] (Figure 4.6C). AD2, on the other hand, will transiently bind to the flank or boundary of the target surface. Mode 2 includes major states III and IV on the lower left region of Figure 4.6A. Here, both AD1 and AD2 bind dynamically to the vicinity of the DNA binding surface (Figure 4.6C), and the N-terminal segment of TAD, including AD1 in some cases, can get close to residue S121 (site of paramagnetic spin labeling [89]; also see below). Mode 2 is less populated than Mode 1. States associated with Mode 2 are shorter-lived and often visited between TAD sampling of states I, II, and V during all eight simulations (e.g., see Figure 4.6B). The prediction of AD1 and AD2 as the DBD interaction hot spots is highly consistent with the NMR chemical shift analysis [89]. Moreover, TAD mostly binds to the DNA binding surface of DBD, apparently driven by electrostatic interactions between the highly negatively charged TAD (-14) and the positively charged DBD surface (Figure 4.16B). A more dominant role of AD2 in p53 TAD/DBD dynamic interactions may also be explained the fact that AD1 is neutral whereas AD2 has a – 5 net charge. Secondary structure analysis shows that
there is no significant increase in residual helicity of TAD in the nonspecific bound state (Figure 4.17), which is consistent with the highly dynamic and electrostatic dominant nature of its interaction with DBD. We note that NMR chemical shift analysis also revealed no significant helicity increase in p53 TAD/DBD interactions [89].

![PRE results of p53 TAD/DBD interactions from experiments and HyRes II model](image)

**Figure 4.7 PRE results of p53 TAD/DBD interactions from experiments and HyRes II model.** Calculated (lines) and experimental (gray bars) PRE effects induced by paramagnetic spin labeling at residues S121 (on p53 DBD). The PRE calculations are averaged over eight replicas excluding the first 100 ns.

We further calculated PRE effects on p53 TAD with spin labelling at S121 on DBD, and compared the results with experimental values [89]. As summarized in Figure 4.7, there is a very high level of agreement between prediction and experimental measurement, suggesting that the simulated ensemble provides a realistic molecular basis of dynamic p53 TAD/DBD interaction. We note that the current HyRes II simulations significantly overestimated the PRE effects in the N-terminal segment of TAD. A couple factors may contribute to this artifact. First, there are 4 negative charges in the N-terminus of TAD (E2, E3, D7 and E11) and it can thus compete with
AD1 and AD2 for interaction with the DNA-binding interface (e.g., in states IV and V, see Figure 4.6C). Given the $< r^{-6} >$ weighted averaging of PRE, even a relatively small population with the N-terminus binding to the vicinity of S121 could lead to signification over-estimation of PRE effects in this region. As shown in Figure 4.18, interaction between p53 TAD N-terminus and DBD is dynamic and transient, such that the average distances to S121 remain larger in the N-terminus compared to AD2. Second, the N-terminus mainly binds to regions of DBD that would become less accessible in the tetramer state. Therefore, over-estimation of PRE effects in the N-terminus of TAD may be attributed to the monomer system simulated here.

4.4 Conclusions

In this work, we developed a new hybrid-resolution coarse-grained protein model, called HyRes II, that may be used to reliably characterize the heterogeneous conformation ensembles and dynamic interactions of IDPs. The model was built upon a previous HyRes model but includes a SASA-based implicit solvent term to better account for the solvent screening effects and address the bias towards over-compaction in the original HyRes model. We deployed a multi-stage iterative optimization strategy to carefully balance the HyRes II model, by tuning SASA implicit solvent parameters for backbone atoms, strengths of vdW interactions for sidechains, and strengths of backbone hydrogen bonding interactions. The final HyRes II model is able to recapitulate key secondary and long-range conformational properties of a set of model IDPs, including residual helicity, peptide chain dimension and transient long-range interactions. Importantly, the ability of HyRes II to generate highly dynamic disordered ensembles without severe over compaction is achieved without over destabilization of structured proteins compared to HyRes. HyRes II is able to distinguish large differences in overall chain dimensions between structured proteins and IDPs. Application of HyRes II to study the disordered ensemble of a nontrivial IDP, the 61-residue TAD
domain of tumor suppressor p53, demonstrates that HyRes II is able to describe various local and long-range structural properties, including partially formed secondary structures, overall chain dimension and transient formation of long-range interactions. The achievable accuracy of HyRes II for these properties is almost comparable to that of a99SB-disp, one of the most accurate explicit solvent atomistic force fields, but with a ~10,000-fold reduction in the computational cost.

We further demonstrate that HyRes II provides an efficient and relatively reliable tool for studying the dynamic interactions of IDPs that are now recognized to be important in cellular signaling and regulation. In particular, we studied the dynamic interactions of p53 TAD with the DBD domain, a process that is important in promoting and regulating specific interactions of p53 with DNA. The calculated structural ensemble from HyRes II simulations is highly consistent NMR chemical shift and PRE data. It provides a reasonable molecular level description of how various segments of p53 TAD dynamically compete for interaction with DBD, which cannot be easily probed by experimental approaches or atomistic simulations. HyRes II can be trivially extended to include the description of post-translational modifications that often regulate IDP structure and interactions. Enabled by GPU acceleration, HyRes II may be further extended to study complex biological processes that occur at large length and long time-scales, such as liquid-liquid phase separations [22, 383-385].
### Supporting Material

Table 4.4 Sequences of training set, test set peptides used in the work. All peptides were capped with an acetyl group and N-methyl amide at N-terminus and C-terminus respectively. (The sequence of p53 DBD can be found from PDB ID: 3KMD [386]).

<table>
<thead>
<tr>
<th>Peptide</th>
<th>( \kappa )</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly(_{10})</td>
<td></td>
<td>GGGGG GGGGG</td>
</tr>
<tr>
<td>(AAQAA)(_3)</td>
<td></td>
<td>AAQAA AAQAA AAQAA</td>
</tr>
<tr>
<td>Gly(_{20})</td>
<td></td>
<td>GGGGG GGGGG GGGGG GGGGG</td>
</tr>
<tr>
<td>KID</td>
<td></td>
<td>TDSQK RREIL SRRPS YRKIL NDLSS DAP</td>
</tr>
<tr>
<td>1VII [363]</td>
<td></td>
<td>MLSDE DFKAV FGMTR SAFAN LPLWK QQNLK KEKGL F</td>
</tr>
<tr>
<td>1BDC [364]</td>
<td></td>
<td>QQNAF YEILH LPNLN EEQRN GFIQS LKDDP SQSANN LLAEA KKLND A</td>
</tr>
<tr>
<td>0.0009</td>
<td>KEKEK KEKEK EKEKE KEKEK EKEKE KEKEK EKEKE KEKEK</td>
<td></td>
</tr>
<tr>
<td>0.0025</td>
<td>EEEKK KEEEK KKEEE KKKEE EKKKK EEEKK EEEKK KEKEE KEKEE</td>
<td></td>
</tr>
<tr>
<td>0.0139</td>
<td>KEKKK KEKEE KEKEK KEKEK KEKEK KEKEK KEKEE KEKEE</td>
<td></td>
</tr>
<tr>
<td>0.0140</td>
<td>KEKEK KEKEE KEKEE KEKEK KEKEK KEKEK KEKEE KEKEE</td>
<td></td>
</tr>
<tr>
<td>0.0245</td>
<td>KEKEE KEKKK EEEKK EEEKK EKEKE EKEKE KKKKE EEEKE KEKEE</td>
<td></td>
</tr>
<tr>
<td>0.0273</td>
<td>EEEKK KEKEE KEKEK KEKEK KEKEE KEKEE KEKEE KEKEE</td>
<td></td>
</tr>
<tr>
<td>0.0450</td>
<td>KEKKK EEEKK KKEEE KKKEE KKKEE KKKEE KKKEE KKEEE</td>
<td></td>
</tr>
<tr>
<td>0.0450</td>
<td>KKKKE EEEKK KKEEE KKKEE KKKEE KKKEE KKKEE KKEEE</td>
<td></td>
</tr>
<tr>
<td>0.0624</td>
<td>EKKEK KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE</td>
<td></td>
</tr>
<tr>
<td>0.0834</td>
<td>EKKKK KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE</td>
<td></td>
</tr>
<tr>
<td>0.0841</td>
<td>EKKEK KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE</td>
<td></td>
</tr>
<tr>
<td>0.0864</td>
<td>EKKEK KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE KEKKE</td>
<td></td>
</tr>
<tr>
<td>0.0951</td>
<td>KEKKK EKEKK EKKEK EKEKK EKEKK EKEKK EKEKK EKEKK</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1311</td>
<td>EKKEK EEKEE EKKK KKEEK EKKEK KKKKE EEEEEE KEKEE KEKEE</td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>0.1354</td>
<td>KKEKK EKKKE KKEKK EEEKE KEKKE KKKKE KEKKE EEEEEE KEKEE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.1458</td>
<td>EKEKE EKKKE EKKKK EKEK EKKEK KEKKE KEEEKE EEEEEE KEKEE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.1643</td>
<td>EKEKK KKKKE KEEKK KEKKE KEEEK EKEK EKKEE EEEEEE KEKEE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.1677</td>
<td>KEKKE EEEEEE EEEEEE KKKKE KKEEE KKKKE EEEEEK KEKEE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.1941</td>
<td>EEEEE KKKKK EEEEEE KKKKK EEEEE KKKKK EEEEEE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.2721</td>
<td>EEEKE EEEKE EEEKE KEKKE EKEK EKKEK KKKKE KKEKE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.2737</td>
<td>EEEEE EEEEEE EKKKK KEKKE KKKKE KKKKE KEEEK EEEKE KEKKE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.3218</td>
<td>KEKKE EEEEEE EEEEEE EKKKK KEKKE KKKKE KEKKE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.3545</td>
<td>EEEEEE EEEEEE EEEEEE EEEEEE KKEKE KKKKE KKKKE KEKKE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.4456</td>
<td>EEEEEE EEEEEE EKEK KKKKE KKKKE KKKKE KKKKE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.5283</td>
<td>EEEEE EEEEEE EKEK EKEK KKKKE KKKKE KKKKE KEEEK KEKEE</td>
</tr>
<tr>
<td></td>
<td>0.6101</td>
<td>KEKKE EEEEEE EEEEEE EKKKK KKKKE KKKKE KKKKE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.6729</td>
<td>KKKKE KEKKE EEEEEE EEEEEE EEEEEE EKEK KKKKE KKKKE KKEKE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.7666</td>
<td>EKKKK KKKKK KKKKK KKKKE EEEEEE EEEEEE KKKKE EEEEEE KKKKE</td>
</tr>
<tr>
<td></td>
<td>0.8764</td>
<td>KEKKE EEEEEE EEEEEE EEEEEE EEEEEE EEEEEE EEEEEE KKKKE EEEEEE KKKKE</td>
</tr>
<tr>
<td></td>
<td>1.0000</td>
<td>EEEEEE EEEEEE EEEEEE EEEEEE EEEEEE KKKKE KKKKE KKKKE KKKKE</td>
</tr>
</tbody>
</table>

| 3GB1 [365] | MTYKL ILNGK TTKGE TTEA VDAAT AEKVF KQYAN DNGVD GEWTY DDATK TFTVT E |
| TAD         | MEEPQ SDPSV EPPLS QETFS DLWKL LPENN VLSPL PSQAM DDLML SPDDI EQWFT EDPGP D |
Table 4.5 Ensemble averaged conformational properties of poly-glycine simulated using HyRes, HyRes II, and ABSINTH. The end-to-end distance ($d_{e2e}$) and radius of gyration ($R_g$) are listed. All numbers are in Å.

<table>
<thead>
<tr>
<th>Model</th>
<th>Peptide</th>
<th>$&lt; R_g &gt;$</th>
<th>$&lt; d_{e2e} &gt;$</th>
<th>Peptide</th>
<th>$&lt; R_g &gt;$</th>
<th>$&lt; d_{e2e} &gt;$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HyRes</td>
<td>Gly</td>
<td>6.1</td>
<td>11.7</td>
<td>Gly</td>
<td>7.8</td>
<td>15.2</td>
</tr>
<tr>
<td>HyRes II</td>
<td>Gly$_{10}$</td>
<td>6.3</td>
<td>12.6</td>
<td>Gly$_{20}$</td>
<td>8.3</td>
<td>17.3</td>
</tr>
<tr>
<td>ABSINTH</td>
<td></td>
<td>6.9</td>
<td>14.8</td>
<td></td>
<td>10.1</td>
<td>20.1</td>
</tr>
</tbody>
</table>

Figure 4.8 Initial conformations of peptides/proteins used in all simulations. For p53, a total of 8 replicas were simulated. In each initial conformation (s1-s8), DBD region (residue 95-291) is colored cyan, and TAD (residue 1-61) and PRR (residue 62-94) are colored in grey. The AD1 (residue 15-26) and AD2 (residue 40-56) segments are highlighted in green and red, respectively, with S15 and P58 Cα atom shown as vdW beads and colored green and red similarly. Hot spots on the DNA binding surface of DBD identified by NMR work [89]: T118, V112 (on L1 loop), and G245, N247 (on L3 loop), and A276, G279, R280, R283 (on C-terminal helix) are shown in vdW beads and colored yellow, with S121 colored in orange. Random coils were used as initial structures of (EK)$_{25}$ peptide simulations.
Figure 4.9 Probability distribution of the $R_g$ and end-to-end distance of Gly$_{10}$ and Gly$_{20}$. (A) $R_g$ for Gly$_{10}$, (B) end-to-end distance for Gly$_{10}$, (C) $R_g$ for Gly$_{20}$, and (D) end-to-end distance for Gly$_{20}$. Conformational properties were obtained from HyRes, HyRes II and ABSINTH simulations at 300 K.
Figure 4.10 Probability distribution of the $R_g$ and end-to-end distance of (AAQAA)$_3$ and KID. (A) $R_g$ for (AAQAA)$_3$, (B) end-to-end distance for (AAQAA)$_3$, (C) $R_g$ for KID, and (D) end-to-end distance for KID. Conformational properties were obtained from HyRes, HyRes II and ABSINTH simulations at 300 K.
Figure 4.11 2D histogram of p53 TAD geometric properties based on the $R_g$ and end-to-end distance simulated in HyRes and HyRes II. (A) and (B) show results from HyRes and HyRes II, respectively. The heat map shows the probability densities of the geometry propensities derived from simulation statistics. The ensembles shown above are derived from the combined trajectories of two independent replicas (totally 800 ns) every 5 ns. The ensembles are shown in Ribbons style, colored by RWB style (from red at the N-terminus to blue at the C-terminus).
Figure 4.12 Stability and dimension of folded proteins in HyRes and HyRes II. (A) Averaged secondary structure propensities of Villin Headpiece, Protein A of *S. aureus*, and protein G B1 domain in comparison to the PDB structures and (B) $R_g$ as function of time and (C) Backbone RMSD from the initial structure during HyRes and HyRes II simulations at 300 K. (D) Representative snapshots taken every 10 ns from 400 ns trajectories (transparent cartoons) in comparison to the PDB structures (solid cartoon).
Figure 4.13 Major states of dynamic interactions of p53 TAD/DBD sampled in each of the eight independent HyRes II simulations. See Figure 4.6 for the states. DBD is shown in cartoon with the DNA binding surface colored in blue. TAD is shown in grey Tube with AD1 and AD2 colored in green and red, respectively. The ensembles shown contain TAD conformations sampled every 1 ns from the 400-ns trajectories.
Figure 4.14 PCA analysis of conformational space sampled by HyRes simulations of the disordered p53 TAD with respect to DBD. Snapshots sampled from all sixteen simulations (HyRes and HyRes II) were used to derive the first two PCs, which are then used to calculate all the projections shown, (A) from all eight HyRes trajectories combined, and (B) from each independent HyRes run (s1-s8). The heat map shows the free energy (in kcal/mol) derived from the 2D probability distributions. (C) Representative structures are shown for the five major basins identified, which are drawn in cartoon style and similar coloring style as shown in Figure 4.8 and Figure 4.6.
Figure 4.15 Major states of dynamic interactions of p53 TAD/DBD sampled in each of the eight independent HyRes simulations. See Figure 4.14 for the states. DBD is shown in cartoon with the DNA binding surface colored in blue. TAD is shown in grey Tube with AD1 and AD2 colored in green and red, respectively. The ensembles shown contain TAD conformations sampled every 1 ns from the 400-ns trajectories.
Figure 4.16 TAD interaction hot spots on DBD. A) TAD-contacting probabilities of DBD residues from HyRes and HyRes II simulations. The L1, L2, L3, and C-terminal helix regions, which constitute the DNA-binding interface of DBD are marked with blue boxes. B) The binding interface of DBD domain (residue 95 to 291). The charged residues are shown in blue (positive) and red (negative). Hot spots on DNA binding surface identified by NMR including S121 (red), N247 and T118 (green), V122, G245, A276, G279, R280, and R283 (orange) are shown in sticks.
CHAPTER 5

TOWARDS ACCURATE SIMULATION OF COUPLING BETWEEN PROTEIN SECONDARY STRUCTURE AND PHASE SEPARATION

Intrinsically disordered proteins (IDPs) frequently mediate phase separation that underlies the formation of biomolecular condensates. Together with theory and experiment, efficient coarse-grained (CG) simulations have been instrumental in understanding sequence-specific phase separation of IDPs. However, the widely-used Cα-only models are limited in capturing the peptide nature of IDPs, particularly backbone-mediated interactions and effects of secondary structures, in phase separation. Here, we describe a hybrid resolution (HyRes) protein model toward a more accurate description of the backbone and transient secondary structures in phase separation. With an atomistic backbone and coarse-grained side chains, HyRes can semi-quantitatively capture the residue helical propensity and overall chain dimension of monomeric IDPs. Using GY-23 as a model system, we show that HyRes is efficient enough for direct simulation of spontaneous phase separation, and at the same time appears accurate enough to resolve the effects of single His to Lys mutations. HyRes simulations also successfully predict increased β-structure formations in the condensate, consistent with available experimental CD data. We further utilize HyRes to study the phase separation of TPD-43, where several disease-related mutants in the conserved region (CR) have been shown to affect residual helicities and modulate phase separation propensity as measured by the saturation concentration. The simulations successfully recapitulate the effect of these mutants on the helicity and phase separation propensity of TDP-43 CR. Analyses reveal that the balance between backbone and sidechain-mediated interactions, but not helicity itself, actually

---

determines phase separation propensity. These results support that HyRes represents an effective protein model for molecular simulation of IDP phase separation and will help elucidate the coupling between transient secondary structures and phase separation.

5.1 Introduction

Biomolecular condensates are now recognized to be an important mechanism of subcellular compartmentalization and contribute to myriad biological functions such as RNA storage and processing, stress response, cellular trafficking, metabolism, and cellular signaling [22, 383, 384, 387]. They have also been linked to many diseases including cancer, neurodegeneration, and infectious diseases [388-393].Extensive studies in recent years have established that biomolecular condensates form and dissolve through spontaneous phase separations driven by dynamic and multivalent biomacromolecules such as RNAs and proteins [394-397]. It is further recognized that condensation of macromolecules likely involves coupling of both associative and segregative phase transitions [397, 398]. One of the major drivers of the phase separation is intrinsically disordered proteins or regions (IDPs/IDRs) [95, 399-401]. IDPs are enriched with charged and polar residues, have lower sequence complexity, lack stable tertiary structures, and are key components of cellular signaling and regulatory networks [1, 2, 4, 261, 309, 402]. The dynamic and multivalent nature allows IDPs to frequently serve as scaffolds of biomolecular condensates [308, 383]. Maturation frequently follows phase separation of IDPs, where the droplets become increasingly viscoelastic and turn into fibrils, gels, glasses, or other solid-like materials [387].

Intensive efforts have been dedicated to understanding how the balance of various solvent and co-solvent mediated backbone and side chain interactions together give rise to a rich spectrum of sequence-specific phase transitions of IDPs, by themselves or together with RNAs and other molecules [403-405].
Classical concepts from polymer physics have been central to understanding the general physical principles governing the biomolecular condensates of IDPs [384, 397, 398, 406-408]. For example, the Flory-Huggins mean-field formalism for homopolymers captures the balance of polymer-solvent and polymer-polymer interactions with a single Flory-Huggins $\chi$ parameter [409, 410], and can be effectively used to fit and understand the phase diagram of IDPs [411-413]. Random phase approximations (RPAs) can account for electrostatic effects on the phase transitions of polyampholytes with any charge pattern, allowing the prediction of sequence and pH/salt dependence of IDP phase separation [414, 415]. Field-theoretic simulations (FTS) have been further developed to numerically sample the fully fluctuating field-theoretic Hamiltonian and alleviate the restriction of analytical approximations in RPAs [416-418]. FTS approaches enable more realistic descriptions of the roles of charge-charge interactions as well as sequence-dependent short- and long-range interactions in driving phase separations of IDPs. The stickers-and-spacers models for associative polymers provide a particularly effective framework for coarse-grained (CG) lattice simulations of the phase transitions of multivalent proteins to decipher the impacts of various sequence features [404, 419-421]. These models are applicable to multi-component systems and can describe additional structural features such as percolation and gelation in dilute and condensed phases [422].

Compared to theoretical approaches, molecular dynamics (MD) simulations using transferable energy functions can be particularly attractive for deriving system-specific molecular details of phase separation [423, 424]. In particular, CG protein models are often necessary to access the large length ($> \sim 100$ nm) and time scales ($> \mu$s) required for direct simulation of IDP phase separation. Among them, $\alpha$-only models have been the most successful in simulating phase separation of IDPs. Here, each residue is represented by a single “$\alpha$” CG bead, interacting
through effective potentials derived from structural statistics, atomistic simulations, and/or experimental observables. Similar Cα-only models, such as Gō-like models, have been quite successful in studies of protein folding [425-428], transient protein-protein interactions [429, 430], and coupled binding and folding of IDPs [74, 167, 213, 215]. The original Kim-Hummer (KH) model was first adopted by Mittal, Best, and coworkers for simulating IDP phase separations [431], where an alternative version with the original Miyazawa-Jerningan (MJ) statistical contact potential replaced by a hydrophobicity scale (HPS)-based one was also explored. Both KH and HPS models were shown to be capable of providing qualitative prediction of changes in the phase diagram of low complexity domain (LCD) of protein FUS due to mutations or the presence of folded protein LAF-1. Many variants of Cα-only models have since been developed with contact pair potentials assigned and optimized using different strategies including machine learning and sometimes with attention to specific interaction types such as π-π contacts [175, 178, 182, 432-438]. Together, these models have demonstrated increasingly robust abilities to recapitulate experimental observables such as monomer radius of gyration (Rg) and phase transition boundaries. Multi-bead CG protein models have also been explored for simulation of IDP phase separation [168, 169, 342]. For example, the popular MARTINI force field can be rebalanced to simulate the droplet formation of FUS LCD. [342] A range of liquid properties including densities, surface tensions, and sheer viscosities could be reproduced with moderate downscaling (~0.65) of the original MARTINI protein-protein interactions.

A critical limitation of Cα-only or MARTINI models, however, is their inability to accurately describe the interactions and secondary structures of the peptide backbone, due to the single bead representation of the entire residue (Cα-only models) or backbone (MARTINI). Yet, it is well recognized that anisotropic backbone interactions are important in driving IDP phase
separations [384, 423, 439-441]. Critically, IDPs are not simple polymers, and transient secondary and sometimes long-range structures are prevalent [248, 260]. Recent studies have highlighted the strong interplay between transient secondary structures and phase separation behaviors [108, 109, 442, 443]. For instance, IDPs and IDR often adopt more expanded conformations within condensates and often with an increased propensity for $\beta$-structures [444-446]. It has also suggested phase separation can induce the folding of poly-(PR) peptides into helical conformations [108, 447], which is a characteristic feature of aging-related phase transition pathologies. Furthermore, recent studies have revealed that a conserved region on C-terminal domain of TAR DNA-binding protein 43 (TDP-43 CR) undergoes transient conformational changes during phase transitions, forming helices during phase separation and subsequently aging into amyloid structures [448]. Intriguingly, TDP-43 mutants with enhanced helicity have lower measured saturation concentrations ($C_{\text{sat}}$), suggesting a driving role of residual helices in TDP-43 phase separation [449, 450].

Accurate description of the peptide backbone and transient secondary structures arguably requires near atomistic representation beyond the single bead C$\alpha$-only resolution [451]. This is because backbone interactions are highly anisotropic in nature, and the anisotropic nature is responsible for the conformational dependence of these interactions. For example, the backbone within a helix segment is no longer available to form hydrogen bonding with water molecules or the rest of the protein. This conformational dependence of backbone interactions cannot be captured using C$\alpha$-only models, even if empirical potentials are introduced to mimic sequence-dependent helical structures [452]. At present, atomistic simulations largely remain the only option for examining details of peptide backbone interactions and potentially the effects of transient structures in phase separation of IDPs [441, 453-457]. While these courageous attempts have
provided useful insights, atomistic simulations of phase separation are severely limited by the achievable simulation timescales, and deriving statistically meaningful predictions of biomolecular condensates is largely out of reach. There is a critical need for simplified protein models that accurately describe the peptide backbone and at the same time are computationally trackable for simulation of IDP phase separation.

Numerous CG models have been introduced to provide higher-resolution presentations of the protein backbone (and side chains) and/or more accurate description of backbone-mediated interactions [354, 458-460], including more recent ones designed specifically for IDPs [169, 342, 461]. In particular, early approaches have utilized virtual potentials of important interactions (e.g., electrostatics or hydrogen bonding) while maintaining a single-bead backbone representation [460, 462]. We recently developed a hybrid resolution (HyRes) protein model that was specifically designed and parameterized for accurate descriptions of local and long-range structures of IDPs [170, 463]. As illustrated in Figure 5.1A, the HyRes model contains explicit atomistic backbones, and the side chains are represented by up to five CG beads (for Trp). The energy function of HyRes contains bond, angle, dihedral/improper, backbone CMAP dihedral cross-term, Lennard-Jones (LJ) van der Waals (vdW), Debye-Hückel electrostatics, and backbone hydrogen bonding and solvent accessible surface area (SASA)-based implicit solvent. The bonded terms as well as the LJ radii were parameterized to reproduce the side chain geometry, volume, and flexibility from atomistic simulations. The relative strengths of LJ potentials ($\varepsilon_i$) for all side chains CG beads were first assigned based on the MJ statistical contact potentials [464], and then uniformly rescaled to balance electrostatics and hydrogen-bonding terms based on a set of small model peptides. The latest version, HyRes II, has been shown to provide a realistic description of a range of nontrivial local and long-range structure features of IDPs, including residual helicity and transient long-range
contacts, at a level comparable to atomistic simulations for the N-terminal transactivation domain (TAD) of tumor suppressor p53 [463]. It was also able to describe the dynamic interaction between p53-TAD and protein cyclophilin D, predicting structure features that are highly consistent with NMR titration and binding experiments [351]. Therefore, HyRes may provide an appropriate CG protein model for direct simulation of peptide backbone and secondary structures in phase separation of IDPs.

Here, we develop a GPU-accelerated HyRes model, and critically assess its suitability for simulating spontaneous phase separation of IDPs and studying the roles of backbone-mediated interactions and residual structures in phase separation. We show that the HyRes-GPU model is able to directly simulation phase separation of GY-23 peptides derived from histidine-rich beak proteins (HBPs) and recapitulate the effects of single His-to-Lys mutations on ability to phase separate. The simulations also correctly predict increased $\beta$-structures in the condensed phase, consistent with previous circular dichroism (CD) measurements. We further examine the molecular basis of how residual helicity of the conserved region (CR) of TDP-43 modulates its phase separation propensity. HyRes simulations correctly recapitulate the effects of monomeric peptide helicity on the condensate stability as measured by $C_{\text{sat}}$. However, analyses based on HyRes simulations reveal that residual helices do not play a direct role by increasing helix-helix interactions in the condensed phase as previously proposed. Instead, it is the balance among solvation, side chain hydrophobicity, and availability of backbone that actually governs the condensate stability of TDP-43 CR. These results suggest that HyRes provides an efficient and viable model for simulating IDP phase separation and interrogating the role of peptide backbone and residual structures.
5.2 Methods

5.2.1 HyRes simulations of IDP monomers

A total of 26 IDPs ranging from ~20 to ~300 residues were simulated in this work. The sequences of these IDPs are provided in Table 5.2, with experimental radii of gyration and solution conditions summarized in Table 5.3. The N- and C-termini of all proteins are capped with acetyl and N-methyl amide groups, respectively. All simulations were carried out using CHARMM [465] and OpenMM [466] on GPUs. Examples of the HyRes-GPU setup and simulation scripts can be found on GitHub at https://github.com/mdlab-um/HyRes_GPU. The initial extended structures were generated using CHARMM and then energy was minimized. Langevin dynamics was performed with a collision frequency of 0.2 ps\(^{-1}\) and an MD timestep of 4 fs. The salt concentration is set at 150 mM, consistent with most experimental conditions (Table 5.3). All bonds involving hydrogen atoms were constrained by the SHAKE algorithm. Nonbonded interactions were smoothly switched off from 1.6 nm to 1.8 nm. All monomers were simulated for 1 μs at 300 K in HyRes unless otherwise noted. Additional simulations of GY-23 mutants were simulated for 400 ns per replica, which proved to be long enough for achieving sufficient convergence (e.g., see error bars in Figure 5.15. Four additional 400-ns production simulations were also performed for WT GY-23 monomer at 260 K, 270 K, 280 K, and 290 K to examine the temperature dependence of conformational properties in comparison to results from condensate simulations.

The first 200 ns trajectories were excluded in subsequent structural analysis, which was performed using a combination of CHARMM and in-house scripts. The secondary structures were identified using the standard Dictionary of Secondary Structure of Proteins (DSSP) protocol within the MDTraj package [467]. \(R_g\) values from HyRes-GPU simulations are uniformly shifted up by 4 Å to roughly account for larger side chain beads as done previously. Viscosity of GY-23 peptides
in dilute and condensed phases was calculated from MSD correlation functions of the Cα atom of the central residue (residue 12). The error bars for all properties were estimated from differences between results calculated from two independent simulation replicas. All molecular visualizations were done using VMD [468].

5.2.2 Simulation of phase separation and equilibrium

To prepare the initial configuration for phase separation simulation of GY-23, we first generated a minimized extended conformation and then packed 200 copies together using the Packmol Software [469]. The initial cubic box was used for packing 30*30*30 nm$^3$ with a distance tolerance of 10 Å to avoid atom overlap (Figure 5.10A). The initial packed state was subject to 1-ns constant pressure and temperature (NPT) equilibration at 300 K, which would result in a high-density compact configuration of ~12 nm in diameter. The cubic box size was subsequently increased to 45 nm for production simulation of phase equilibrium. Multiple copies of constant volume and temperature (NVT) simulations were then performed at every 5 K from 260 K to 320 K for 2 µs, which have been shown to be sufficient to reach phase equilibrium (see main text). A separate NVT simulation was performed at 600 K to generate fully dispersed initial configurations (Figure 5.10A), which were then used to initiate three independent 2-µs simulations of spontaneous phase separation of WT GY-23 at 300 K. The convergence of the phase equilibrium will be examined by comparing simulations initiated from compact and dispersed initial configurations (see main text). Additional simulations were performed for WT GY-23 to examine the finite-size efforts on the phase equilibrium, using two additional constructs with 100 copies in a 35.2-nm cubic box and 300 copies in a 51.7-nm cubic box (Figure 5.12). Similar protocols were used to generate the initial compact configurations and the simulations were performed for 2 µs at 300 K. A similar protocol was used to generate the compact initial configuration of TDP-43 CR
(Figure 5.10B), except for the larger packing box used (45*45*45 nm³). The phase equilibrium simulations of TDP-43 CR were performed at 300 K in a 60-nm cubic box and lasted 2 μs each. For each sequence (WT and 8 mutants), three independent simulations were run to examine the convergence and estimate the errors. For G335A and G338A with the lowest \( C_{\text{sat}} \) values, we performed additional triplicates of simulations in a larger 120-nm cubic box for 6 μs each to further examine the convergence. The larger box size allows 8 times more copies of peptides in the dilute phase, which should provide better statistics for calculating \( C_{\text{sat}} \). The WT peptide was also re-ran in triplicates in the 120-nm box as a control. The results are summarized in Table 5.5 and Figure 5.19.

### 5.2.3 Phase diagram and properties of the condensates

A single dominant droplet was observed in all phase equilibrium simulations. They were first identified using the density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm within the Python Scikit package [470]. Density profiles show that the density of the condensate is uniform within the droplet (e.g., Figure 5.3D, 5.11, 5.12B and D). We chose a cutoff radius of 3 nm (√\( n \)) to determine the concentration of the condensates of GY-23, averaged over the last 500 ns were used, where the phase equilibrium is fully established (e.g., see Figure 5.12 B and D). Residue-residue contact maps were determined based on the minimum distance of two residues (only heavy atoms), where the contact was identified if the minimum distance was no longer than 0.5 nm.

### 5.3 Results and Discussion

#### 5.3.1 GPU acceleration of the HyRes protein model

We first implemented the HyRes model on GPU within the OpenMM package [146, 225, 233] (see Methods) to dramatically expand the accessible length- and time-scales required for
studying IDP phase separation. We note that the SASA solvation energy term, applied only to the backbone, is removed in the GPU implementation and compensated by recalibrating the backbone LJ parameters for computational efficiency (Table 5.1). The recalibration was guided by examining the conformational properties of two model IDPs, (AAQAA)$_3$ and KID (see Table 5.2 for sequences). The new parameters for backbone atoms, summarized in Table 5.1, allow HyRes-GPU to essentially reproduce both residual helical structures and overall compaction of both peptides (Figure 5.8). The final re-calibrated model as implemented on GPU is referred as HyRes-GPU hereafter, which is available through GitHub at https://github.com/mdlab-um/HyRes_GPU. The GPU implementation can provide several hundred-fold speed-ups compared to the CPU calculations for large systems (Figure 5.1B). The production rate is near 200 ns/day for a 100 K atoms system on a single Nvidia GTX 4070Ti card. We further validated the balance of HyRes-GPU using p53-TAD (residue 1-61), a longer IDP with nontrivial structural properties (Figure 5.1 C-D). The result shows that HyRes-GPU largely retains the ability of HyRes II to capture residual helical propensities. In particular, HyRes-GPU provides further improvement in capturing the level of overall compaction as reflected in the distribution of radius of gyration ($R_g$), compared well to the result from a99SB-disp, one of the best explicit solvent atomistic protein force fields [123, 231].
Figure 5.1 Representation, efficiency, and accuracy of HyRes-GPU. (A) Physical representation of the HyRes model. The backbone atoms are drawn in CPK style with carbon in cyan, nitrogen in blue, oxygen in red, and hydrogen in white. The side-chain beads are illustrated by vdW spheres. (B) Computational efficiencies of HyRes-GPU in comparison to the same model run on CPU (that is, no SASA term) for systems of different sizes. The timing simulations were performed using either a single Xeon E5-2620 CPU core or a single Nvidia RTX 4070Ti GPU card. (C-D) Probability distributions of $R_g$ and residue helicity profiles of p53-TAD (1-61) simulated using HyRes-GPU, in comparison to results from HyRes II (grey traces) and atomistic force field a99SB-disp (blue traces). The vertical dashed lines mark the average $R_g$ values of the simulated ensembles. The atomistic results are taken from a previous work. [184]

5.3.2 Ability of HyRes-GPU to describe properties of monomeric IDPs

It has been shown that single-molecule properties of IDPs are intimately connected to their phase behavior [471, 472]. For example, the coil-to-globule transition temperature can be used as a proxy for predicting the upper critical solution temperature [471]. In fact, single-chain properties, particularly $R_g$, have been frequently used in the optimization of CG protein models for the simulation of phase separation [182, 437, 438]. Models better reproducing experimental single-
chain $R_g$ have been shown to more accurately predict homotypic phase separation propensities. Therefore, we first examine the ability of HyRes-GPU to reproduce the experimental $R_g$ of a set of IDPs ranging from 24 to 236 residues in length (Table 5.2). Some of these IDPs have been shown to drive phase separation, while others have not. The test set also exhibits significant variation in charge distribution (from -44 to +16), which can help to evaluate balance among local contacts, long-range nonpolar and electrostatic interactions in the model. As summarized in Figure 5.2A, $R_g$ values computed from HyRes-GPU simulations agree well with the experimental results, with a Pearson’s correlation score of $r = 0.66$. This compares well to early Cα-only models such as KH ($r \sim 0.54$) [182] and HPS or HPS-Urry ($r \sim 0.3 – 0.6$) [437]. This is notable because HyRes-GPU has not been directly optimized to reproduce experimental $R_g$ of a large number of IDPs. Instead, HyRes has only been parameterized based on small model peptides including poly-Gly, (AAQAA)$_3$, KID and the (EK)$_{25}$ series, to achieve qualitative long-range non-specific interactions and semi-quantitative residual structure descriptions [170, 463]. Along this line, it is unsurprising that HyRes-GPU yields a lower Pearson’s correlation score compared to $R_g$-optimized models, such as Mpipi [182], HPS-cation-π [436] and others [437, 438]. The implication is also that HyRes-GPU could be further improved following similar optimization strategies [473, 474].

We further analyzed the residual structures of several IDPs known to exhibit dynamic local conformations with available experimental data (Figure 5.2B). The results support that HyRes-GPU can semi-quantitatively predict sequence-specific helical propensities. Although there are some mismatched local structures that are over- or under-estimated, residual helices within the structural ensembles generated by HyRes-GPU are generally consistent with experiments. Overall, HyRes-GPU is able to reproduce the experimental helicity profiles with RMSD $\sim 0.1$ (or 10%; Figure 5.2B). For example, the model accurately predicts $\sim 20\%$ helicity between 25-40 residues.
of ACTR, ~15% helicity between 55-65 residues of Ash1, and the helicity between residues 105-118 of FUS. Particularly, hNHE1cdt is not overly helical even though its chain dimension is slightly underestimated by HyRes-GPU (as indicated by its \( R_g \) values in Figure 5.2A). The correlation coefficients between simulation and NMR-derived residue helicities are \( r > 0.5 \) in 5 out of 10 cases (Figure 5.9), but can be as low as \(< 0.1\) in other cases such as FUS, hNHE1cdt and \( \alpha \)-synuclein. These are often cases where the residual helicity levels are low. Even though HyRes-GPU is generally able to recapitulate the low helical propensities (as reflected in small RMSD values, Figure 5.2B), not all transient helices in simulated ensembles are well aligned with the NMR results.
Figure 5.2 Monomeric conformational properties of IDPs in HyRes-GPU. (A) Correlation between $R_g$ from simulation and experiment for a set of IDPs. The sequences are provided in Table 5.2. (B) Residue helicity profiles from HyRes-GPU in comparison to experimental results calculated from NMR secondary chemical shifts using SSP [475]. The references and experimental conditions of NMR studies are listed in Table 5.4. The RMSDs between the experimental and simulation profiles are also listed. The correlations between experimental and simulated residue helical propensities are shown in Figure 5.9.

5.3.3 HyRes-GPU simulation of spontaneous phase separation of GY-23

To examine the ability of HyRes-GPU to simulate phase separation, we first focus on a short 23-residue low complexity peptide GY-23 (see Table 5.2 for sequence), the central segment of squid-beak derived protein HBP-1 that regulates its pH-dependent phase separation [441]. GY-23 alone can undergo homotypic phase separation at room temperature and neutral pH of 7.0.
Importantly, its phase behavior is highly sensitive to even single-point mutations [476]. Therefore, GY-23 provides a good model system for evaluating the suitability of HyRes-GPU for direct simulation of phase separation and further examining if the model is accurate enough to recapitulate the effects of single-point mutations. For this, we performed multiple simulations of 200 copies of wild-type (WT) GY-23 in a 45-nm cubic box for a total effective concentration of ~3.6 mM, starting from either preformed high-density compact or fully dispersed initial states at 300 K (see Methods and Figure 5.10 for the generation of initial configurations). As illustrated in Figure 5.3, the system reached similar phase-separated states within ~1 µs regardless of the initial configurations. Representative movies of the simulation trajectories are provided in the Supplementary Materials (Movies S1 and S2). A single dominant droplet was observed in the second half of all six independent simulations. The droplets are of comparable sizes (Figure 5.3C) and have largely identical density profiles (Figures 5.11 and Figure 5.3D). In simulations initiated from the dispersed initial configuration, small clusters were observed to form and dissolve during the initial stages of phase separation (Figure 5.3C). Once exceeding certain critical size, the droplet would grow rapidly until reaching the final equilibrium. We further examine the finite size effects on the droplet by simulating the phase equilibrium with 100, 200, and 300 copies in three different box sizes at 300 K. The results, summarized in Figure 5.12, show that the droplets have similar density profiles and reach the same high concentration in the droplet interior in all three cases. As such, all additional simulations of WT GY-23 and mutants at various temperatures were performed using 200 copies in the 45-nm cubic box.
Figure 5.3 Spontaneous phase separation of GY-23 in HyRes-GPU. (A, B) Snapshots from representative simulations initiated from either compact or dispersed initial states, showing the system reaching a similar phase separated (PS) final state during the course of 2 µs simulation timescale. (C) The size of the largest cluster during three replicas of simulations initiated from either the compact or dispersed initial states. The panels illustrate that multiple small nuclei (red circles) often form and dissolve during the initial stage of simulations from the dispersed state. (D) The concentration profiles of the final condensed phase were derived from all 6 simulations. $R_{\text{in}}$ is the radius around the center of the droplet. The error bars were estimated from standard deviation of five block averages within the last 500 ns trajectories.

We note that the peptides undergo rapid exchange between the condensed and dilute phases throughout HyRes-GPU simulations. Figure 5.13A plots the numbers of GY-23 peptides that never leave the droplet as a function of time from representative trajectories of simulations initiated from the compact configuration at different temperatures. It shows that the number of peptides that had yet exchanged with the dilute phase decreased rapidly to zero, within ~250 ns at 300 K, even though the equilibrium droplet size is ~125 (Figure 5.3C). The rapid dynamics of peptide exchange between two phases in HyRes-GPU is a highly desirable property; it allows efficient simulation of
the dynamic phase equilibrium to derive reliable thermodynamic properties and potentially resolve the effects of mutations and other perturbations. We further characterize the peptide diffusion coefficients of dilute and condensed phases by analyzing the mean-squared displacement (MSD) functions (Figure 5.13B). The results show that the diffusion coefficient of GY-23 is about 10-fold smaller in the droplet compared to the dilute phase. This is smaller than the ~100-fold experimental estimate. [477] Multiple factors may contribute to the apparent underestimation of viscosity increase in the condensate, including smoother effective potentials used in CG models such as HyRes-GPU and the small droplet size simulated here (~10 nm in diameter, Figure 5.3D).

5.3.4 **HyRes-GPU recapitulates mutational effects on GY-23 phase equilibrium**

It has been shown experimentally that the ability of GY-23 to phase separate is sensitive to Histidine-to-Lysine (H-to-K) mutations. [476] Either single H-to-K mutations, H2K, H12K, H15K, and H20K, or all H-to-K mutations, H/K, can significantly shift the phase co-existence, such that phase separation is completely abolished for H2K, H20K, and H/K mutants at all tested conditions and can only be achieved at higher pH (> 8) and/or salt concentrations (0.5 or 1 M NaCl) for H12K and H15K mutants. A set of simulations at different temperatures spanning 260 to up to 310 K were performed for each GY-23 construct (including WT) to probe the impact of mutations on phase equilibrium. All simulations were initiated from a preformed high-density compact configuration (see Method and Figure 5.12). The results show that, with a total concentration of ~3.6 mM, the preformed droplet was quickly dissolved for all H-to-K mutants at 300 K (Figure 5.14) and there is no phase separation at the end of the 2-µs simulations unlike the case of WT GY-23 (Figure 5.4A). This is apparently consistent with the experimental observations [476]. Interestingly, whereas the H/K mutant shows no phase separation ability even at 260 K, the single H-to-K mutants can undergo phase separation at lower temperatures. To further quantify the
impacts of H-to-K mutations on phase separation of GY-23, we constructed the phase co-existence curves by fitting them to the Flory-Huggins mean field model with three-body interactions using FIREBALL [413] (Figure 5.4B). The H/K mutant was not included in the diagram due to the absence of phase separation at all temperatures simulated. The apparent critical concentration for phase separation of the WT peptide at 300 K is ~1.5 mM, consistent with the experimental result showing that phase separation occurs at 0.5 mM or higher [476]. For the single H-to-K mutants, the critical concentration of phase separation is >6 mM at 300 K, which is consistent with a lack of phase separation experimentally examined between 0.1 to 1.0 mM. The ability of HyRes-GPU to recapitulate the impact of single H-to-K mutations on the phase behavior of GY-23 is encouraging. It suggests that the balance between various interactions, particularly nonpolar vs. electrostatic, is reasonably well captured in HyRes-GPU, which is crucial for accurate simulation of how sequence and other structural properties of IDPs may modulate phase separation.

Figure 5.4 Phase separation of GY-23 peptide variants. (A) Representative final snapshots of the 2-µs simulations of WT GY-23 and four single H-to-K mutants in HyRes-GPU at 300 K. (B)
Phase diagram of GY-23 WT and single H-to-K mutants. The error bars shown were estimated from the standard deviation of three block averages within the last 500 ns trajectories. The total peptide concentration is \( \sim 3.6 \text{ mM} \) in all simulations, marked by the red dashed vertical line. Dashed smooth binodal curves were generated by fitting to the Flory-Huggins model with three-body interactions using FIREBALL. [413] (C) Probabilities of intermolecular residue-residue contacts in the condensed phase of WT GY-23 at 300 K.

### 5.3.5 Molecular basis of GY-23 phase separation: importance of aromatic residues and coupling with \( \beta \)-structure formation

To further elucidate the molecular basis of GY-23 phase separation, we first analyze the probabilities of intermolecular contacts in the condensate and compare the peptide conformational properties in the dilute and condensed phases. The results reveal that there is only a slight increase in the peptide chain expansion upon forming the condensate, as reflected in the distributions of \( R_g \) and end-to-end distance (\( R_{ee} \)) (Figure 5.5 A and B). This is not surprising because GY-23 is highly disordered in both phases. Monomeric GY-23 peptides display minimal secondary structures at 300 K (~1.5%, Figure 5.15), which is consistent with previous CD measurements [477]. The propensity for greater extension in the condensed phase has been observed for IDPs in numerous other cases [171, 478-480].

A notable difference in the GY-23 conformations in dilute and condensed phases is a significant increase in the propensity for \( \beta \)-structures within the condensed phase (Figure 5.5C). In contrast, the peptide contains similarly low helical content in both phases (Figure 5.5D). Although the overall probability of forming \( \beta \)-strands and sheets remains low (~3% per residue), the presence of \( \beta \)-structures (including \( \beta \)-bridges) within the condensate likely plays a significant role in stabilization. In particular, within the droplet at 300 K, even though only ~5% of the peptides participate in intermolecular parallel or anti-parallel \( \beta \)-sheets, almost 35% of the peptides are involved in \( \beta \)-bridges and over half of these are intermolecular ones (Figure 5.5 E and F). It is possible that intermolecular \( \beta \)-bridges may enable networking of peptides within the condensate.
and contribute to percolation [398]. A representative snapshot of the equilibrium droplet at 300 K is shown in Figure 5.5G, where the presence of several intra- and intermolecular β-sheets are highlighted. We note that the level of intermolecular β-structures increases dramatically and faster than intramolecular ones at lower temperatures (Figure 5.16). At 260 K, over 50% of peptides are involved in β-structures, of which over 70% are intermolecular ones, apparently contributing to more stable phase separation. The importance of β-structures, and particularly an apparent coupling between β-structures and condensate formation, has also been suggested in a previous experimental study of a longer 161-residue HBP-1 construct [477]. CD spectra reveal a negative peak at 215 - 220 nm upon condensation formation at high HBP-1 concentration, consistent with a shift towards higher β-structure. These results strongly support a coupling between protein secondary structures and phase separation and highlight the importance of accurate description of protein backbone and transient secondary structures.

Figure 5.5 Conformational properties of WT GY-23 at phase coexistence at 300 K. (A-D) Distributions of $R_g$, $R_{ee}$, residue β-structure propensity and helicity in dilute (black) and condensed (red) phases. (E) Probability of each residue forming intra- vs inter-peptide β-sheet
structures in the condensed phase. (F) Probability distributions of the fraction of peptides that contain any \(\beta\)-structures (blue), participate in \(\beta\)-bridges (green), or \(\beta\)-sheets (cyan) in the condensed phase. (G) A representative snapshot of the droplet with \(\beta\)-sheet and helical structures colored in blue and red, respectively. Note that \(\sim1/3\) of peptides are involved in \(\beta\)-bridges and they are not highlighted.

We further examine the molecular basis of how H-to-K mutations affect the ability of GY-23 to phase separate. Analysis of the monomeric conformational properties, summarized in Figure 5.15, shows that single H-to-K mutations have minimal impacts on compaction (< 1%) or secondary structural propensities. This is not surprising given the highly unstructured nature of GY-23. Only replacing all four histidines with lysines in the H/K mutant can lead to a modest increase (~ 5%) in the overall expansion of the peptide, apparently due to the charge-charge repulsion. Therefore, the conformational properties of the peptide itself are unlikely a significant contributor to the reduced phase separation propensity. Instead, analysis of the residue-residue contact probability map suggests that aromatic residues likely play major roles in stabilizing the condensate. As shown in Figure 5.4C for WT GY-23, the most dominant intermolecular contacts in the condensate exclusively involve histidine (H), phenylalanine (F), and tyrosine (Y) residues. H-to-K mutations do not only remove favorable \(\pi\)-\(\pi\) stacking interactions but also introduce extra charge-charge repulsion forces. It is conceivable that replacing any of these aromatic residues could significantly reduce condensate stability. Examination of the contact probability maps of condensates of the single H-to-K mutants at 280 K, where phase separation can still occur, reveals a significant reduction in the probabilities of intermolecular contacts involving the mutation site (Figure 5.17). Intriguingly, it appears that the location of the H-to-K mutation may have differential impacts on the condensate. Mutations at the middle region (H12K and H15K) appear to be less disruptive compared to those near the termini (H2K and H20K). Indeed, H12K and H15K can be more readily driven to phase separation, by increasing salt concentration and/or increasing pH, compared to H2K and H20K [476]. The difference in sensitivity H-to-K mutation at central vs
terminal positions is likely an entropic one, that disrupting the terminal intermolecular contacts may more significantly increase the peptide flexibility and hinder the formation of intermolecular contacts. Taken together, the apparent ability of HyRes-GPU to capture these nontrivial effects of GY-23 His-to-Lys mutations further supports that the model appears reasonably balanced among structural propensities and various polar and nonpolar interactions.

5.3.6 Phase separation of TDP-43: correlation between helicity and condensate stability

RNA-binding protein TDP-43 is a key component of ribonucleoprotein granules and has been linked to diseases such as amyotrophic lateral sclerosis (ALS), frontotemporal dementia, and Alzheimer’s [481-483]. Particularly, a conserved region (CR, residues 311-360) within the disordered C-terminal domain (CTD) of TDP-43 is crucial for the function of TDP-43 and harbors several ALS-associated mutations [484]. TDP-43 CR is partially helical and can independently drive phase separation [485, 486]. Importantly, recent NMR studies further revealed that the level of residual helicity of various CR mutations, including ALS-associated ones, is positively correlated with the stability of the condensate as measured by $C_{sat}$ [109, 449]. It has been suggested that transient helices may play a direct role in mediating intermolecular interactions during phase separation [487-489]. However, existing CG simulations using Ca-only models can not directly describe backbone-mediated interactions and residual helices during phase separation, and atomistic simulations cannot adequately sample the dynamic phase equilibrium to derive reliable intermolecular interactions within the condensate or predict the condensate stability [109]. As such, the molecular details that underlie the apparent correlation between residual helicity and condensate stability remain poorly understood.
Figure 5.6 Residue helicity profiles and saturation concentrations of TDP-43 CR variants. (A) Residue helicity profiles of WT TDP-43 CR and mutants G335A, G338A, M337P, and A326P (the last 9 residues 352-360 were not shown due to the low helicity). The error bars were calculated as the standard deviation of block averages. The comparison between residue helicity from simulation and experiments can be found in Figure S11. (B) Comparison between experimental and simulated $C_{sat}$ values. The experimental values were taken from Ref. [109] for the entire CTD domain of TDP (residues 267-414). Quantitative experimental value was not reported for mutant A326P but was stated to be similar to that of M337P. The simulated values were derived from the last 500 ns of HyRes-GPU trajectories in the 60-nm box at 300 K, and the error bars are the standard deviation calculated from three replicas. The Pearson’s correlation coefficient is $r = 0.77$ (0.85 without including G338A). (C) Representative snapshots of phase co-existence of WT, G335A, G338A, and A326P TDP-43 CR at 300 K.

We first examine the ability of HyRes-GPU to recapitulate the effects of mutations on the residue helicity of four selected mutations. The results show that the residue helicity profiles predicted by HyRes-GPU are highly consistent with secondary NMR chemical shifts for the WT...
peptide and all four variants (Figure 5.6A and Figure 5.18). In particular, for WT TDP-43 CR, the simulation correctly predicts a major partial helix in residues 321-334, with residue helicity max ~40%, and a second minor partial helix in residues 338-346 with a lower helicity of ~20%. This is in quantitative agreement with the NMR secondary chemical shift analysis [109]. Mutations G335A and G338A extend the major and minor partial helices, respectively, which is again in excellent agreement with NMR secondary chemical shifts (Figure 5.18). The correlation coefficients between NMR secondary chemical shifts and simulated residue helicities are above 0.8 for both alanine mutants. Mutation A326P reduces the helicity of the major partial helix, while M337P, located in a coiled region, has minimal impact on the residue helicity profile. Both predictions are in excellent agreement with NMR, even though the correlation between NMR and simulation is only 0.19 for A326P due to the low residual helicity (Figure 5.18D). We note that no ad hoc tuning is required in the above simulations. HyRes-GPU is a transferable CG protein force field that can semi-quantitatively describe the sequence-dependent secondary structure propensities, as well as the overall peptide chain dimension (e.g., Figure 5.2).

We then directly simulated the phase equilibrium of WT TDP-43 CR and several single mutations using HyRes-GPU, starting from preformed high-density configurations (see Methods and Figure 5.9). Movies of representative simulation trajectories of WT, G335A, and A326P systems are provided in the Supplementary Materials (Movies S3, S4, and S5). Despite the substantially larger size, HyRes-GPU allows relatively fast exchange of peptides between the dilute and condensed phases. The rapid exchange is critical for accurate quantification of $C_{\text{sat}}$ because of the high stability of TDP-43 CR condensates. For example, experimental $C_{\text{sat}}$ is ~15 $\mu$M for the WT peptide, which translates to only ~2 peptides in the dilute phase in the 60 x 60 x 60 nm$^3$ simulation box. To further examine the convergence, additional triplicates of simulations
were performed for G335A and G338A, which have the smallest $C_{\text{sat}}$, in a larger 120-nm box and for a much longer length of 6 $\mu$s. The WT sequence was also simulated using the larger box as a control. The results, summarized in Table 5.5 and Figure 5.19, show that the results using two box sizes are largely consistent, even though $C_{\text{sat}}$ values for G335A and G338A are now more similar as expected based on the experimental results.

Figure 5.6B compares $C_{\text{sat}}$ values from HyRes-GPU simulations (with the 60-nm box) for WT TDP-43 CR and 8 mutants to the experimental results on corresponding TDP-43 CTD constructs, with numerical values provided in Table 5.4. Representative snapshots of the phase coexistence are given in Figure 5.6C. Overall, there is a strong correlation between the simulation and experiment ($r = 0.77$ for all and 0.84 if using the values derived from the 120-nm box simulations for G335A and G338A), even though HyRes-GPU predicts $C_{\text{sat}}$ about 2-3-fold of the experimental values. The modest systematic over-estimation of $C_{\text{sat}}$ could reflect limitations of HyRes-GPU, but may also be attributed to the longer protein construct characterized in the experiment study (the whole CTD instead of CR alone) [109]. Importantly, the simulations successfully recapitulate the apparent positive correlation between residual helicity and condensate stability of TDP-43. Specifically, the simulations correctly show that disruption of partial helices through A326P mutation (Figure 5.6A) significantly increases $C_{\text{sat}}$ (Figure 5.6B and Table 5.4). In contrast, enhancing partial helices via G335A or G338A (Figure 5.6A) also leads to a significant reduction in $C_{\text{sat}}$. This is noteworthy given the transferable and CG nature of HyRes-GPU.

5.3.7 How residual helicity modulates TDP-43 CR saturation concentration

The ability of HyRes-GPU to correctly reproduce the apparent correlation between residual helicity and $C_{\text{sat}}$ as discussed above allows us to further analyze the underlying molecular basis. Examination of the probability map of residue-residue contacts in the condensate of WT TDP-43
CR, shown in Figure 5.7A, reveals that three aromatic residues, F313, F316, and W334, play dominant roles in mediating intermolecular interactions. Previous in vitro and in vivo experiments also suggest that aromatic residues play critical roles in phase separation of TDP-43 [490-493]. It is noteworthy that the helical regions do not appear to be a major contributor to intermolecular contacts in the condensed phase. Inspection of the distribution of partial helices within the condensates reveals that helical regions spread around uniformly in the condensed phase and there is no apparent tendency for preferential helix-helix contacts (e.g., see Figure 5.20A). To quantify the fraction of residues involved in intermolecular contact when both belong to partial helices, we plot the fraction of helix-helix contacts for each residue of WT, G335A, G338A, A326P, and M337P in Figure 5.20B. Overall, the helix-helix contacts made only small contributions to the phase separation, up to 6% for residues in the major partial helical region for G335A. Importantly, the fractions of helix-helix contacts are strictly proportional to the residue helicity (Figure 5.20C). There is no apparent cooperativity in helix-helix contacts. That is, increasing helicity does not directly enhance intermolecular contact formation to help drive phase transition. This notion is further supported by the observation that M337P, which does not change helicity, has a similar effect on phase separation compared to A326P.
Figure 5.7 Contact map and the relationship between helicity and phase separation behavior of TDP-43 CR. (A) Probabilities of intermolecular residue-residue contacts in the condensed phase of WT TDP-43 CR at 300 K, derived from the last 500 ns of the trajectories. The major partial helix is shown on top. (B) Single-chain helicity profiles of WT, WT+ΔHel, G335A, and G335A-ΔHel. The last 9 residues 352-360 were not shown due to the low helicity and intermolecular contact probabilities. (C) $C_{Sat}$ of WT, WT+ΔHel, G335A, and G335A-ΔHel from the last 500 ns of the co-existence simulations at 300 K. All error bars were calculated as the standard deviations of three replicas.

Besides affecting residual structures, mutations have direct impacts on intermolecular interactions by introducing different side chain groups. For example, the G-to-A mutation adds a hydrophobic methyl group. In addition, changing the helicity has a direct impact on the availability
of backbone for intermolecular interactions. These two factors can compete with one another to influence the phase separation in a nontrivial way. To evaluate how side chain hydrophobicity and backbone availability contribute to TDP-43 CR phase separation, we designed two artificial mutants named as WT$^{+\Delta Hel}$ and G335A$^{-\Delta Hel}$ (Figure 5.7B). Here, additional dihedral restraints were imposed at residue 335 to artificially increase the helicity of WT to the same level of G335A (WT$^{+\Delta Hel}$), or to reduce the helicity of G335A to the level of WT (G335A$^{-\Delta Hel}$). The simulations show that WT$^{+\Delta Hel}$ actually reduces the condensate stability despite the increased helicity, resulting in an increased $C_{sat}$ (Figure 5.7C). This is opposite of what one may expect if helicity has a direct and positive contribution to condensate stability. Instead, artificially increasing helicity in the WT sequence reduces the availability of backbone for intermolecular interactions, destabilizing the condensate. This is supported by analysis of residue-residue contact probabilities, showing backbone-backbone contacts are the most affected in WT$^{+\Delta Hel}$ (Figure 5.21). For G335A$^{-\Delta Hel}$, the calculated $C_{sat}$ is similar to that of G335A despite reduced helicity near residue 335 (Figure 5.7C). The implication is that the introduction of the hydrophobic Ala side chain is the dominant contributor behind increased condensate stability. This effect becomes evident when comparing contact probabilities between WT and G335A$^{-\Delta Hel}$ (Figure 5.22A), as well as between WT$^{+\Delta Hel}$ and G335A (Figure 5.22B). Remarkably, the hydrophobic Ala side chain substantially augments contacts of residues near position 335, even with different levels of residual helicities.

For A326P and M337P, the proline mutation leads to a global reduction in the contact probabilities (e.g., see Figure 5.23), presumably because prolines reduce the ability of peptides to pack closely in the condensed phase. As such, M337P can lead to a similar increase in $C_{sat}$, even though its helicity is very similar to that of WT. Taken together, these analyses strongly suggest that mutations will perturb the complex balance among multiple factors, including peptide
conformation, backbone availability, and sidechain hydrophobicity, to modulate the IDP phase equilibrium. The nontrivial net effects of mutations on the balance are likely responsible for the contrasting observations regarding the influence of helicity on the propensity for phase separation in various experimental studies [109, 494]. HyRes-GPU should provide a powerful tool for studying these nontrivial effects with the ability to accurately describe residual structures and backbone-mediated interactions.

5.4 Conclusions

In summary, we have developed an efficient GPU-accelerated HyRes CG protein model that provides atomistic description of the peptide backbone to enable more accurate simulations of phase separation of IDPs. The HyRes-GPU model is physics-motivated and transferable. It provides a semi-quantitative description of both local residual structures and overall level of compaction of the monomeric form of a diverse set of IDPs, even though the model could be further parameterized to better reproduce experimental $R_g$. Application to two phase-separating IDPs, GY-23 and TDP-43 CR, demonstrates that HyRes-GPU is sufficiently balanced to capture rapid peptide exchange between the dilute and condensed phases and could recapitulate the effects of various single mutations on the dynamic phase equilibrium. Our in-depth analysis of GY-23 and TDP-43 CR phase separation also provides new insights on the role of backbone-mediated interactions and coupling of residual structures with phase separation. In the case of highly disordered GY-23, it is found that phase separation can promote the formation of $\beta$-structures within the condensed phase, which likely augment intermolecular interactions and stabilize the condensates. For TDP-43 with substantial residual helicity, HyRes-GPU simulations reveal that contacts in the condensate predominantly involve aromatic residues and there are minimal preferential helix-helix interactions. Instead, the apparent positive correlation between helicity and
condensate stability as measured by $C_{sat}$, observed in both experiment and simulation, is a result of nontrivial effects of perturbing peptide conformations, backbone availability, and sidechain hydrophobicity. In particular, increasing helicity itself can actually reduce the availability of backbone for intermolecular interactions and destabilize the condensate. Taken together, HyRes-GPU expands the capability of CG simulations of IDP phase separation and provides a viable tool for studying how detailed structural and sequence features may affect the phase equilibrium at the molecular level.
## 5.5 Supporting Material

<table>
<thead>
<tr>
<th>Atom Type</th>
<th>$\varepsilon_i$ (kcal/mol)</th>
<th>$\varepsilon_i$ (1-4) (kcal/mol)</th>
<th>$R_{min}$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>-0.0149</td>
<td>-0.0498</td>
<td>0.225</td>
</tr>
<tr>
<td>C</td>
<td>-0.0180</td>
<td>-0.1000</td>
<td>2.100</td>
</tr>
<tr>
<td>Cα</td>
<td>-0.0073</td>
<td>-0.1000</td>
<td>2.365</td>
</tr>
<tr>
<td>Cα (Gly)</td>
<td>-0.0172</td>
<td>-0.1000</td>
<td>2.235</td>
</tr>
<tr>
<td>N</td>
<td>-0.0356</td>
<td>-0.2384</td>
<td>1.600</td>
</tr>
<tr>
<td>O</td>
<td>-0.2390</td>
<td>-0.1159</td>
<td>1.600</td>
</tr>
</tbody>
</table>
Table 5.2 Sequences of all IDPs studied in the work. Peptides that have already been shown to drive phase separation are colored in blue.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(AAQAA)$_3$</td>
<td>AAQAA AAQAA AAQAA</td>
</tr>
<tr>
<td>KID</td>
<td>TDSQK RREIL SRRPS YRKIL NDLSS DAP</td>
</tr>
<tr>
<td>α-synuclein</td>
<td>MDVFM KGLSK AKEGV VAAAE KTKQG VAAEA GKTKE GVLVYV GSKTK EGVVH GVATV AEKTQ EQVTN VGAVG VTGTV AVAQK TVEGA GSIAA ATGCV KKDQL GKNHE GAPQE GILED MPVD DNEAY EMPSE EGYQD YEPEA</td>
</tr>
<tr>
<td>ACTR</td>
<td>GTQNR PLLRN SLDDL VGPPS NLEGQ SDERA LLDDL HTLLS NTDAT GLEEI DRALG IPELV NQGQA LEPKQ D</td>
</tr>
<tr>
<td>Ash1</td>
<td>GASAS SSPSP STPTK SGKMR SRSSS PVRPK AYTPS PRSPN YHRFA LDSPP QSPRR SSNSS ITKGG SRRSS GSSPT RHTTR VCV</td>
</tr>
<tr>
<td>hNHE1cdt</td>
<td>MVPAH KLDSP TMSRA RIGSD PLAYE PKEDL PVITI DPASP QSPES VDLVN EELKG KVLGL SRDPA KVAAE DEDDD GGIMM RSKET SSPGT DDVFT PAPSD SPSSQ RIQRC LSPDG PHPEP GELEG FPFGQ Q</td>
</tr>
<tr>
<td>IBB</td>
<td>GCTNE NANTP AARLH RFKMK GKDST EMRRR RIEVN VELRK AKKDD QMLKR RNVSS FPDDA TSPLQ ENRNN QGTVN WSVVL IVKGI NSSNV ENQLQ AT</td>
</tr>
<tr>
<td>K18</td>
<td>MQTAP VPMDP LKNVK SKIGS TENLK HQPGG GKVQI INKKL DLSNV QSKCG SKDNI KHVPG GGSVQ IVYKP VDLK VTSKQ GSLGN IHHKP GGGQV EVKSE KLDFK DLDFQ QISDG QISDN HVTG QISMQ GNQCN</td>
</tr>
<tr>
<td>N49</td>
<td>GCQTS RLFGF NNVTN NINNS SSGMN NASAG LFSGK P</td>
</tr>
<tr>
<td>N98</td>
<td>GCFNK SFGTP FGGGT GGFGT TSTFG QNTGF GTTSG GAFGT SAFGS SNNTG GLGFN SQTKP GGLFG TSSFS QPATS TSTGF GFGT GFTS TGTAN TLFGT ASTGT SLLFG QUHAF AQNKP TGFGN FSTGT SSGGL FGTGN TTSNP GSTG GFSF G</td>
</tr>
<tr>
<td>NLS</td>
<td>ACETN KRKRE QISTD NEAKM QIQEE KSPKK KRKKR SSSAN KPNPE</td>
</tr>
<tr>
<td>NSP</td>
<td>GCNFN TPQPN KTPFS FGTAN NNSNT TNQNS STAGA AFGTQ QSTFG FNNSA PNNTN NANSS ITPAF GNNTF GNTAF GNSNP TNVSF GNNSN TTNTF GSNSA GTSLF GSASQ QKTQS NQTAG NQTFG SSSLF NNSTN SNTTK PAFGQ LNFAG GNNTT PSSTG NANTG NNLFN ATANA N</td>
</tr>
<tr>
<td>NUL</td>
<td>GCGFK GFDTK SSSSN SAASS SFKFG VSSSS SGGQI TLSTG GNKFQ GDQGG FKIGV SSDES SINPM SEGFK FSFKF GDFKF GVSSQ SKPPQ VKKDS KNDFN SGLGF SVGLN P</td>
</tr>
<tr>
<td>NUS</td>
<td>GCPSA SPAFG ANQTP TFGQS QGASQ PNPPQ FGSIS SSTAL FPTGS QPAFQ TFQGV SSSQ PPVFG QPQPS RAFGS GTTPN</td>
</tr>
<tr>
<td>p53 (1-93)</td>
<td>MEEPQ SDPSE EPPLS QETFS DLWKL LPENN VLSPL PQSAM DDLML SPDDT EQWFT EDPGP DEAPR MPEAA PPVAP APAAP TPAAP APAPS WPL</td>
</tr>
<tr>
<td></td>
<td>ProTa</td>
</tr>
<tr>
<td>----</td>
<td>-------------------------------</td>
</tr>
<tr>
<td></td>
<td>MSDAA VDTSS EITTK DLKEK KEVVE EAENG RDAPA NGNAE NEENG EQEAD NEVD EESEE GREEE GEEEEE DGGEE DGDED EEAES ATGKR AAEDD EDDDV DTKKQ KTDDE D</td>
</tr>
</tbody>
</table>
Table 5.3 Experimental $R_g$ and the corresponding measurement conditions. Conditions including temperature ($T$), salt concentration ($C_{\text{salt}}$), and pH. The lengths of the proteins (\(N_{\text{res}}\)) are also listed.

<table>
<thead>
<tr>
<th>Protein</th>
<th>(N_{\text{res}})</th>
<th>(R_g) (nm)</th>
<th>(T) (K)</th>
<th>(C_{\text{salt}}) (M)</th>
<th>pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha)-synuclein [495]</td>
<td>140</td>
<td>3.55</td>
<td>293</td>
<td>0.2</td>
<td>7.4</td>
</tr>
<tr>
<td>ACTR [496]</td>
<td>71</td>
<td>2.6</td>
<td>278</td>
<td>0.2</td>
<td>7.4</td>
</tr>
<tr>
<td>Ash1 [497, 498]</td>
<td>83</td>
<td>2.9</td>
<td>293</td>
<td>0.15</td>
<td>7.5</td>
</tr>
<tr>
<td>hNHE1cdt [496]</td>
<td>131</td>
<td>3.63</td>
<td>278</td>
<td>0.2</td>
<td>7.4</td>
</tr>
<tr>
<td>IBB [499]</td>
<td>97</td>
<td>3.20</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>K18 [499]</td>
<td>130</td>
<td>3.80</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>N49 [499]</td>
<td>36</td>
<td>1.59</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>N98 [499]</td>
<td>151</td>
<td>2.86</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>NLS [499]</td>
<td>44</td>
<td>2.40</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>NSP [499]</td>
<td>176</td>
<td>4.10</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>NUL [499]</td>
<td>112</td>
<td>3.00</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>NUS [499]</td>
<td>115</td>
<td>2.49</td>
<td>277</td>
<td>0.16</td>
<td>7.4</td>
</tr>
<tr>
<td>p53 (1-93) [276]</td>
<td>93</td>
<td>2.87</td>
<td>293</td>
<td>0.15</td>
<td>6.8</td>
</tr>
<tr>
<td>ProTa [168, 500]</td>
<td>111</td>
<td>3.79</td>
<td>--</td>
<td>0.15</td>
<td>7.0</td>
</tr>
<tr>
<td>SH4-UD [501]</td>
<td>85</td>
<td>2.90</td>
<td>277</td>
<td>0.22</td>
<td>7.5</td>
</tr>
<tr>
<td>Sic1 [502]</td>
<td>92</td>
<td>3.21</td>
<td>298</td>
<td>0.16</td>
<td>7.5</td>
</tr>
<tr>
<td>Hst5 [503]</td>
<td>24</td>
<td>1.38</td>
<td>293</td>
<td>0.15</td>
<td>7.5</td>
</tr>
<tr>
<td>(Hst5)$_2$ [503]</td>
<td>48</td>
<td>1.87</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
</tr>
<tr>
<td>FUS [504]</td>
<td>163</td>
<td>3.32</td>
<td>297</td>
<td>0.15</td>
<td>7.4</td>
</tr>
<tr>
<td>LAF-1 RGG (WT) [92]</td>
<td>176</td>
<td>3.08</td>
<td>293</td>
<td>0.15</td>
<td>7.4</td>
</tr>
<tr>
<td>LAF-1 RGG (shuffled) [92]</td>
<td>176</td>
<td>3.0</td>
<td>293</td>
<td>0.15</td>
<td>7.4</td>
</tr>
<tr>
<td>A1 LCD [505]</td>
<td>137</td>
<td>2.76</td>
<td>298</td>
<td>0.15</td>
<td>7.0</td>
</tr>
<tr>
<td>TDP-43 CTD [450]</td>
<td>147</td>
<td>2.8</td>
<td>298</td>
<td>0.2</td>
<td>6.1</td>
</tr>
<tr>
<td>Ddx4 LCD [506]</td>
<td>236</td>
<td>3.61</td>
<td>297</td>
<td>0.15</td>
<td>6.5</td>
</tr>
</tbody>
</table>
Table 5.4 IDPs with available experimental data on residue helicity profiles and the corresponding measurement conditions. Conditions including temperature ($T$), salt concentration ($C_{salt}$), and pH. The methods for secondary structure analysis are also listed.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Methods</th>
<th>$T$ (K)</th>
<th>$C_{salt}$ (M)</th>
<th>pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-synuclein [507, 508]</td>
<td>NMR, SSP</td>
<td>288</td>
<td>0.05</td>
<td>7.4</td>
</tr>
<tr>
<td>ACTR [507, 509]</td>
<td>NMR, SSP</td>
<td>304</td>
<td>0.05</td>
<td>6.7</td>
</tr>
<tr>
<td>Ash1 [510]</td>
<td>NMR, SSP</td>
<td>298</td>
<td>0.137</td>
<td>6.95</td>
</tr>
<tr>
<td>hNHE1cdt [511]</td>
<td>NMR, SSP</td>
<td>278</td>
<td>0.15</td>
<td>7.2</td>
</tr>
<tr>
<td>K18 [507, 512]</td>
<td>NMR, SSP</td>
<td>283</td>
<td>0.1</td>
<td>7.4</td>
</tr>
<tr>
<td>p53 (1-93) [507, 513]</td>
<td>NMR, SSP</td>
<td>288</td>
<td>0.15</td>
<td>7.4</td>
</tr>
<tr>
<td>Sic1 [302, 507]</td>
<td>NMR, SSP</td>
<td>278</td>
<td>0.14</td>
<td>7.0</td>
</tr>
<tr>
<td>FUS [507, 514]</td>
<td>NMR, SSP</td>
<td>298</td>
<td>0.15</td>
<td>5.5</td>
</tr>
<tr>
<td>A1 LCD [515]</td>
<td>NMR, SSP</td>
<td>298</td>
<td>0.05</td>
<td>7.0</td>
</tr>
<tr>
<td>TDP-43 CTD [486, 507]</td>
<td>NMR, SSP</td>
<td>283</td>
<td>0.15</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Table 5.5 Experimental and simulated $C_{sat}$ (in μM) for TDP-43 CR variants. Simulated values were calculated for all constructs using a 60-nm cubic box and additionally in a 120-nm cubic box for WT, G335A and G338A for validating convergence.

<table>
<thead>
<tr>
<th>TDP-43 CR</th>
<th>Experimental $C_{sat}$ [487]</th>
<th>Simulated $C_{sat}$ (60 nm)</th>
<th>Simulated $C_{sat}$ (120 nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>13.7 ± 1.5</td>
<td>23.1 ± 3.4</td>
<td>22.2 ± 3.4</td>
</tr>
<tr>
<td>WT$^{ΔHel}$</td>
<td>--</td>
<td>27.4 ± 3.2</td>
<td>--</td>
</tr>
<tr>
<td>G335A</td>
<td>5.4 ± 0.7</td>
<td>14.2 ± 2.0</td>
<td>16.8 ± 3.3</td>
</tr>
<tr>
<td>G335A$^{ΔHel}$</td>
<td>--</td>
<td>14.7 ± 0.8</td>
<td>--</td>
</tr>
<tr>
<td>G338A</td>
<td>4.6 ± 0.6</td>
<td>21.1 ± 3.1</td>
<td>18.7 ± 3.5</td>
</tr>
<tr>
<td>A326P</td>
<td>--</td>
<td>44.5 ± 4.2</td>
<td>--</td>
</tr>
<tr>
<td>M337P</td>
<td>15.9 ± 0.9</td>
<td>44.7 ± 4.5</td>
<td>--</td>
</tr>
<tr>
<td>G335S</td>
<td>10.9 ± 0.3</td>
<td>16.8 ± 4.9</td>
<td>--</td>
</tr>
<tr>
<td>Q331K</td>
<td>15.5 ± 0.3</td>
<td>34.4 ± 3.5</td>
<td>--</td>
</tr>
<tr>
<td>G335D</td>
<td>10.5 ± 0.4</td>
<td>22.5 ± 2.5</td>
<td>--</td>
</tr>
<tr>
<td>G335N</td>
<td>9.0 ± 0.4</td>
<td>18.1 ± 5.3</td>
<td>--</td>
</tr>
</tbody>
</table>
Figure 5.8 The probability distributions of $R_{gt}$, $R_{ee}$, and residue helicity of model peptides.  
(A) (AAQAA)$_3$ and (B) KID sampled from HyRes-GPU (red) and HyRes II (CPU, with SASA term, in grey) at 300 K.

Figure 5.9 Correlation of residue helicities between HyRes-GPU and NMR SSP results. The available NMR reference and experimental conditions are listed in Table 5.4.
Figure 5.10 Preparation of the initial conformations for phase separation simulations. (A) and (B) illustrate the preparation for GY-23 and TDP-43 CR systems, respectively. Similar procedures were used for both WT peptides and their mutants.

Figure 5.11 The density of the largest droplets during six simulations of WT GY-23 at 300 K. The Compact-to-PS and Dispersed-to-PS initiated from the compact and dispersed initial conformation, respectively.
Figure 5.12 Finite-size effects on the simulation of GY-23 condensates. (A) The droplet was first identified using DBSCAN with a radius of 2.5 nm. The density profile was then calculated as a function of radius around the center of the droplet. (B) Density profiles of the final droplet of WT GY-23 at equilibrium at four selected temperatures. The box size is 45 nm. (C) The initial compact configurations WT GY-23 with three different box sizes but at the same total concentration of 3.6 mM. (D) Density profiles were calculated for droplets at phase equilibrium from simulations of WT GY-23 with three different box sizes as shown in (C).

Figure 5.13 GY-23 phase separation properties. (A) Numbers of WT GY-23 peptides that remain in the preformed droplet as a function of simulation time at different simulation temperatures. (B) MSD of GY-23 in the dilute (black) and condensed (blue) phases at 300 K as a function of time.
Figure 5.14 Phase separation of WT and mutant GY-23 peptides. The left column plots the number of peptides in the largest cluster as a function of simulation time at temperatures ranging from 260 up to 310 K. Traces at 260 K, 280 K and 300 K are highlighted using solid lines, with representative final snapshots shown.
Figure 5.15 Conformational properties of monomeric GY-23 peptides. Probability distributions of (A) $R_g$, (B) $R_{ee}$, (C) average residue helicity, and (D) $\beta$-structure propensity of WT GY-23 and five mutants at 300 K. The vertical lines in panels A and B mark the average values.
Figure 5.16 β-structure properties in the condensed phase of WT GY-23 at 280 K and 260 K. (A) The propensities of intra- and inter-β structures. (B) Distributions of the fractions of peptides involved in all β-structures, β-bridges, and β-sheets in the droplet.
Figure 5.17 Residue contacts of GY-23 variants. Probability maps of inter-molecular residue-residue contacts in the condensed phases of (A) WT (at 300 K), and (B-E) H2K, H12K, H15K, and H20K at 280 K. (F) Average numbers of intermolecular contacts per frame for residues H2, H12, H15, and H20 in the condensate of WT GY-23 at 300 K.
Figure 5.18 Helicity of TDP-43 CTD variants. (A-E) Comparison between residue helicities from simulation (red line with right y axis) and experimental secondary chemical shifts\textsuperscript{27, 28} (black line with left y axis) of WT, G335A, G338A, A326P, and M337P TDP-43 CTD peptides. The Pearson’s correlation coefficients are shown to the top. (F) Correlation between simulated residue helicities and experimental secondary chemical shifts.
Figure 5.19 $C_{\text{sat}}$ of WT, G335A, and G338A TDP-43 CTD. (A) Comparison between experimental and simulated $C_{\text{sat}}$ values with a cubic box size of 60 nm (●) or 120 nm (▲). The simulated values were derived from the last 500 ns of 6-$\mu$s HyRes-GPU trajectories at 300 K, and the error bars are the standard deviation calculated from three replicas. (B-D) Representative snapshots of phase co-existence of WT, G335A, and G338A TDP-43 CR at 300 K in the 120-nm cubic box.
Figure 5.20 Helix-helix interactions in condensed phase. (A) A representative snapshot of the condensate of WT highlighting the presence of partial helices. Note that these partial helices are dispersed and do not form preferential interactions. (B) Fraction of helix-helix contacts ($f_{\text{helix-helix}}$) for each residue of WT, G335A, G338A, A326P, and M337P in blue, orange, green, red and purple, respectively. (C) Correlation between the average residue helicity and fraction of involvement in helix-helix contacts for WT, G335A, G338A, A326P, and M337P in blue, orange, green, red and purple, respectively. Dashed lines are the linear fit with $R^2$ shown in the key.
Figure 5.21 Probability differences of intermolecular backbone-backbone contacts between WT and WT^{+\Delta H} TDP-43 CR condensates. The difference was calculated through subtracting the contact probabilities in the condensed phase of WT by those of WT^{+\Delta H}. Positive values (red) represent higher contact probabilities in the WT case, while negative values (blue) represent lower contact probabilities compared to the WT^{+\Delta H} case.
Figure 5.22 The effect of Ala on intermolecular residue-residue contact probabilities. (A-B) Probability differences of inter-molecular residue-residue contacts between WT and G335AΔHel and between WTΔHel and G335A. The difference was calculated through subtracting the contact probabilities in the condensed phase of former variant by those of the latter one. Positive values (red) represent higher contact probability in the former variant, while negative values (blue) show lower contacts compared to the latter one.
Figure 5.23 Probability differences of inter-molecular residue-residue contacts between WT and M337P. The difference was calculated through subtracting the contact probabilities in the condensed phase of WT by that of M337P. Positive values (red) represent higher contact probability in the WT case, while negative values (blue) show lower contact probabilities compared to the M337P case.
CHAPTER 6
RESOLVING DIFFERENT CONFORMATIONAL STATES OF WNV NS2B/NS3 PROTEASES BY ENGINEERING THE CYTOPLYSIN NANOPORE TWEEZER: INSIGHTS FROM COARSE-GRAINED SIMULATIONS

West Nile Proteases NS2B/NS3 play a crucial role in viral replication, rendering them an attractive target for antiviral drug discovery. However, the highly charged and relatively flat active pocket of the proteases poses a challenge in effective drug developments. The dynamics between the ligand-bound/unbound (closed/open) states of the proteases are particularly challenging to measure. To address this, we utilized the ClyA nanopore tweezers, a single-molecular and label free technique, to investigate the dynamics of WNV NS2B/NS3 in its two conformational states. Coarse-grained simulations were applied to explore potential interactions between ClyA and WNV proteases, which helped to identify a preferred binding region on ClyA (mid-region). By specifically engineering this mid-region especially the residue 57, which were found to play a significant role in stabilizing protein-pore interactions, ClyA was able to capture and hold the proteases in its lumen and generate dynamic-related signals. Further analyses revealed distinct blockage and Ires% signals at the binding region for NS2B/NS3 open and closed states, respectively. This simulation work provides crucial insights into the interactions between proteases and the pore, particularly in resolving signals between the two conformational states by engineering the nanopore, which can be helpful for drug discovery in the future.

6.1 Introduction

The West Nile virus (WNV), a member of the Flaviviridae family within the flavivirus genus, poses a rising global threats as a mosquito-borne human pathogen and causes hundreds to thousands yearly infections as reported in U.S along [516-518]. Despite advancements in medical
research and drug discovery, the lack of effective therapeutics against WNV reflects the complexity of combatting this viral pathogen [519]. During the viral life circle, the NS2B/NS3 (NS stands for non-structural) proteases is a key player responsible for cleaving the long non-structural polyprotein during the viral replication process. Therefore, targeting the NS2B/NS3 proteases has emerged as a promising way for antiviral drug development [520-522]. However, the NS2B/NS3 (Figure 1.2B), as the classical serine protease (Asp-His-Ser), presents unique challenge for drug targeting due to its highly charged and shallow groove shaped catalytic triad [523]. The C-terminal of NS2B (NS2B CTD), which would fold into a hairpin and bind to the active pocket on NS3 in the ligand bound state (closed state), can remain disordered or fluctuate away from the active site in the free state (open state) in the solvent [38, 83]. Although the NMR and X-ray studies have shown that the closed of WNV NS2B/NS3 is pre-dominant in the solvent [84, 85], the conformational dynamics of NS2B and NS3 remain unclear. Therefore, clearly elucidating the dynamic behavior of the NS2B/NS3 protease becomes crucial to facilitate the identification of the drug bound states, which can unveil potential vulnerabilities and aid in the identification of novel drug targets.

As a label-free technique, the biological nanopores have evolved into a robust tool for characterizing single-molecular dynamics [524-528]. The protein nanopores can form a single hole with the diameter range from ~ 1 nm (e.g., OmpF pore) to 5.5 nm (e.g., PlyAB pore) in an insulating membrane [529], creating two separate electrolyte chambers, where the ionic current will sensitively respond to the analyte’s entering, conformational changes, and interactions in high temporal resolution (μs-ms) [530, 531]. Therefore, the Cytolysin A (ClyA) nanopore, featuring a cis- entrance of ~5.5 nm and a trans-entrance of ~ 3.3 nm long (constriction region) [532-536], serves as an optimal tool for probing the structural dynamics of NS2B/NS3 proteases (~ 4.4 nm
for the longest dimension, Figures 1.2B and 6.4d) due to its potential capacity to trap the proteases within its lumen. However, in order to effectively monitor the dynamics of NS2B/NS3 in the ClyA pore, it is essential to have a prior understanding towards specific proteases/pore interactions that can yield robust, reproducible, and dynamics-related distinguishable current signals. Subsequently, it will enable the further investigation of the conformational dynamics of proteases from the produced signals such as the blockage and residual current (Ires%).

Molecular dynamics simulations provide in-depth molecular-level insights and, when synergistically combined with the ClyA nanopore tweezer, offer a powerful approach to link the current signal changes in line with proteases conformational change and dynamic interactions [528]. In this work, a hybrid-resolution model, HyRes [90, 170], was applied to study the proteases/pore interactions. HyRes is a hybrid resolution model, which can semi-quantitatively describe the secondary structure and long-range non-specific interactions of disordered proteins, proved robustness and effective in efficiently simulating IDP-related interactions [53, 185]. By doing the steered MD, we have identified a favorable binding region in ClyA, where the 57-site residue play a critical role in stabilizing protein/pore interactions. Engineering the residue at position 57, specifically towards a non-negative residue like E57A/K, resulted in cleaner binding signals. Further analyses demonstrated that the specifically engineered ClyA nanopore can effectively distinguish the structural dynamics between NS2B/NS3 proteases (here after referred as proteases).

6.2 Methods

6.2.1 Steered MD simulation setup

Due to the system complexity, the simulations were performed using the HyRes CG force field to allow for sufficient sampling of potential dynamic and non-specific interactions between
NS2B/NS3 and ClyA. The HyRes model has atomistic backbone and coarse-grained sidechain beads, with energy potentials being optimized for more quantitative description of IDP conformation and non-specific interactions [90]. For the simulation of ClyA/proteases, the atomistic structures of WNV proteases and ClyA were first mapped to the HyRes model and then minimized with Cα atoms harmonically restrained. The GsSGs linker exits to connect NS2B/NS3 proteases.

Two enterobacterial species of ClyA, *Escherichia coli* (*E*.coli) and *Salmonella typhi* [533, 537] (*S*.typhi) have been applied to confine the proteases, which share high conformational similarity (12 units of protomer) and sequence identity (as shown in Figure 6.4a). The proteases were initially placed 45 Å away from the nanopore in 10 different rotations, with the center of mass facing to the center of the ClyA pore lumen. Since the cis entrance of ClyA is larger than the longest dimension of proteases, more rotations of proteases were used to allow for diverse orientation of proteases when entering the pore. The pulling force was added to the Cα atom of residue H122, which is close to the center of mass, with 60 pN directing to the nanopore center (cis-trans direction, Figure 6.4b).

Except for NS2B CTD, all Cα atoms on the structured region of NS2B/NS3 have been harmonically restrained to avoid melting with a force of 0.5 kcal/mol. Two sets of the simulations have been performed with two ClyA species, with the Cα atoms on NS2B CTD (residue 48 to 59) being harmonically restrained (0.5 kcal/mol) for the closed state and restraints-free for open states simulations, respectively. All Cα atoms on ClyA have been harmonically positional restrained with a force of 1.0 kcal/mol.

All simulations were carried out using CHARMM [223, 366]. All simulations were simulated at 300 K with constant volume and temperature (NVT), using a 2-fs integration step.
with Langevin thermostat. All bonds involving hydrogen atoms were constrained by SHAKE algorithm [367]. Nonbonded interactions were smoothly switched off from 1.6 nm to 1.8 nm. All simulations last for 50 ns.

### 6.2.2 Standard MD simulation setup

Following the same mapping scheme, the standard MD simulations have been performed for *E.coli*, *S.typhi* WT, *S.typhi* E57A, and *S.typhi* E57K variants, respectively. Only residues 36-86, 123-176 and 208-256 on ClyA have been kept, which consist of the “mid-region” of ClyA lumen. Proteases were placed in the center of ClyA with the center of mass in line with the pore center. Six different orientations were used as initial conformations, where the NS2B CTD faces up, down, right, left, front, and back to the ClyA pore, to enable more sufficient sampling (Figure 6.5). Except for NS2B CTD, all Cα atoms on the structured region of NS2B/NS3 have been harmonically restrained to avoid melting with a force of 0.5 kcal/mol. Two sets of the simulations have been performed for all four ClyA variants, with the Cα atoms on NS2B CTD (resid 48 to 59) being harmonically restrained (0.5 kcal/mol) for the closed state and restraints-free for open states simulations, respectively. Two replicas have been performed for each orientation of proteases in ClyA pores. All Cα atoms on ClyA have been harmonically restrained with a force of 1.0 kcal/mol.

All simulations were performed using similar setup as the steered MD except being running at 400 K. Higher temperature was used to accelerate the simulation and prevent the trapping of unfavorable proteases/pore interactions. All standard MD simulations last for 200 ns.

### 6.2.3 Analysis

In the pulling simulations, the z-depth measured the center of mass of NS2B/NS3 proteases, where the origin (z = 0) was set at the center of the membrane (Figure 6.1d). The z-direction aligns with the trans-cis direction. VMD is used for protein visualization [468].
In the standard MD simulations, the first 100 ns was excluded for all 200-ns coarse-grained simulation trajectories in subsequent analysis, which was performed using a combination of CHARMM, in-house scripts, and the python codes with MDTraj package [286]. Principal component analysis (PCA) was performed using python SciKit-learn package [288], to evaluate the sampling convergence as well as to visualize the simulated ensembles. For this, snapshots were taken every 100 ps from the entire 100 ns trajectories to collect all sampled proteases conformations. Particularly, the symmetric conformation of proteases along z direction have been aligned to one single orientation to avoid over-counting. The ensembles generated from all simulations (6 orientations of proteases in four ClyA pore with two replica each orientation, 24 trajectories in total) were combined together and then using the coordinates of Cα atoms except for the NS2B CTD for the PCA analysis. This analysis captured both orientation of proteases and how it interacts DBD. The free energy surfaces shown were derived directly from the 2D probability distributions along the first two principal components (PCs).

To estimate the solvated ion cross-section and Ires%, the in-house CHARMM scripts were used to count the whole area of ClyA lumen, and the rest area being occupied by proteases. The 3-dimensional pore lumen was first divided into multiple layers with each layer’s height up to 0.1 Å. The counting of the surface area in each layer were done by subtracting the area occupied by coarse-grained beads from a square. For the non-constriction region in the pore lumen, the width of the square is 50 Å (the longest lumen diameter is ~ 55 Å). For the constriction region, the width of the square is 40 Å (the shortest lumen diameter is ~ 35 Å). The water probe is set to1.4 Å. As for the Ires%, the residual current is inversely proportional to the resistance, while the resistance can be calculated from the cross-section by integrating the cross-section through the whole pore,
\[
I_{res}^\% = \frac{\int_{S_{\text{ClyA + Proteases}}^{\text{cross-section}}(Z)} dz}{\int_{S_{\text{ClyA}}^{\text{cross-section}}(Z)} dz}
\]  
(Eq. 6.1)

where \(S_{\text{ClyA + Proteases}}^{\text{cross-section}}(Z)\) stands for the cross section area with proteases in the pore and \(S_{\text{ClyA}}^{\text{cross-section}}(Z)\) stands for the cross section of ClyA whole lumen without proteases.

### 6.3 Results and Discussion

#### 6.3.1 Steered MD revealing a preferred mid-pore binding region on ClyA

We first performed steered MD simulations by pulling the proteases through the ClyA pore, which can help us to explore if there are “favorable” binding sites in the lumen region. The pulling force was set to be small enough, so that the proteases would be pulled towards the pore but also could “stop by” the inner pore sites when favorable interactions occurred with particular orientations. As shown in Figure 6.1 a and c, several resident cases can be found in both \(S.typhi\) (such as R1\_open, R3\_close, R4\_close and R10 open cases) and \(E.coli\) (such as R2\_close, R3, R6, and R7 cases) pulling simulations. This is not very surprising as ClyA lumen is highly charged and conserved between two species types (Figure 6.4a), which could potentially interact with the charge rich NS2B/NS3 proteases as well. However, a noteworthy observation is that the majority of these mid-binding events occurred in the mid-pore region (except for the R1\_open pulling case in \(S.typhi\)) and stayed relatively longer in \(E.coli\) pore during the pulling simulations, where the center of the mass of proteases is near z-depth of approximately 5 nm (Figure 6.1 b and d). For example, four resident events (R2\_close, R3\_close, R6\_close, and R7\_open) have lasted for the whole 50 ns pulling simulations in \(E.coli\). On the contrary, all mid-binding occurred in \(S.typhi\) pore last much shorter (within 20 ns) and would quickly drop to the constriction region due to the external forces. In order to figure out the potential reasons contributing to the distinct mid-binding behavior in two pores, we first took a closer look at the interaction surface between proteases and
ClyA. As shown in Figure 6.2A, the bindings are mainly driven by electrostatic interactions, where the charged residues on proteases properly line with the charged residues on pore lumen site. Notable, a key difference between *S. typhi* and *E. coli* is the residue 57 on the pore middle-lumen, (Glu for *S. typhi* and Ala for *E. coli*). This specific variation in residue composition might serve as a pivotal factor contributing to proteases different binding behaviors.

### Figure 6.1 The pulling traces of proteases in *S. typhi* pore and *E. coli* pore, and representative snapshots for the mid-binding events.

- **a)** Pulling traces in *S. typhi*. c) Pulling traces in *E. coli*. The Y-axis is the relative distances of proteases to the pore membrane center (Z=0, illustrated in d). The traces of closed and open configurations of proteases are drawn in blue and red, respectively. Proteases pulled from different initial rotation were labeled on the top of each sub-trace figures, from R1-R10. b) and d) showed the representative snapshots of the middle trapping states. Key residues on ClyA pore that would interact with proteases were highlight out in CPK drawn style, colored by charge properties.

#### 6.3.2 Mid-binding simulations reveal major binding configurations

To elucidate the different binding behavior of proteases in two pores and investigate if the 57 site are crucial for protein/pore interactions, we executed the standard simulations specifically at the pore middle region (see Methods). To quantify the binding states of proteases in two types
of ClyA pore, we did the PCA analyses based on the dynamic proteases/pore interaction configurations. As shown in Figure 6.2b, more heterogeneous protease binding states can be found in *S.typhi* pore, where three major states co-exist in the NS2B/NS3 dynamic simulations (S1-S3). On the other hand, the dynamics and interactions of proteases in the *E.coli* pore are cleaner, as only S3 and S4 two states populated in all trajectories that initiated from six different orientations. We note that the open conformation of proteases is relatively more dynamic than the closed state in the same pore, which is plausible since NS2B CTD is highly charged and will unfold and fluctuate freely in the coarse-grained simulations at 400 K. The dynamic of NS2B CTD could thus interfere or change the bound states for the rest body. However, such effects mainly influence the morphology of the edges of the local minima in the free energy surface, rather than inducing a shit in binding towards alternative states.

We further mapped the configuration of four binding states by tracing back the snapshots used for PCA analysis. As illustrated in Figure 6.2c (only the charge residues on proteases surface have been highlighted), it is evident that the prevalence of the four binding states is likely driven by the electrostatic interactions. However, the stability of these states varies significantly. In the case of S1, the proteases are oriented in a way that only the middle and bottom two surface charge “rings” can effectively interact with the K147/E57 and D53/R49 belts on ClyA. Conversely, the S2 configuration interacts with more charged belts on ClyA. However, the binding site of S2 is slightly upper, and the charged belts on the pore are highly mixed, making its stability susceptible to even a slight positional shift. Particularly, the S3 binding state is a common configuration observed in all proteases/pore dynamics, indicating its potential as one of the most stable binding configurations. The R49 belt stabilizes 6 negative charged residues on proteases, and the D53 belt stabilizes one of the longest positive charge rings on proteases. Given that these two longest charge
“rings” are stabilized by the purely anti-charged residue belt on ClyA, the frequent occurrence of S3 in all trajectories regardless of proteases structural states and nanopore species is not surprising. The last state S4, is featured by the stabilization between the bottom positive charge ring and ClyA D53 belt, particularly prominent when proteases are in the open conformation in the E.coli pore. Notably, S3 and S4 are the two major binding states for proteases in E.coli pore, possibly benefiting from the neutral A57 site compared to the negatively charged E57 in S.typhi. As shown in Figure 6.2c bottom panel, A57 prevents the destabilization of two long negative charged rings (up to five residues on the proteases surface). This observation may explain the comparatively less stable nature of S3 and S4 binding states in the S.typhi pore and why S.typhi exhibits much shorter mid-binding events during the pulling simulations.

Figure 6.2 Mid-binding configurations and properties of proteases in E.coli and S.typhi pores. a). Representative pulling traces for proteases in S. typhi (top) and E.coli pore (down). The closed and open configurations of proteases are drawn in blue and red, respectively. The representative mid-trapping state is shown in the right panel, with residues on ClyA colored by binding probability, which follows the BGW coloring scheme with hot spots colored deeper (black for probability equals 1) and low contact residues colored lighter (white for probability equals 0). Proteases are drawn by surface and colored by residue type: blue for positively charged, red for negatively charged, white for nonpolar, and pink for polar residues. The zoom-in
figure shows the contacts between ClyA and proteases, with hot spots on ClyA drawn by sticker and colored using the same color scheme as protease. Residues on proteases that within 8 Å of ClyA are also highlighted in sticker form. b). The PCA clustering of proteases/ClyA binding states from standard MD simulations at 400 K. The representative snapshots for the four high populated states are shown in c) (top), with the illustrative charge patterning match figures shown in the bottom. Charge residues in the lumen of ClyA and on proteases surface are colored blue and red, for positive and negative ones, respectively. d). The blockage of closed (blue) and open (red) proteases in S.typhi and E. coli pore. d) The Ires% of closed (blue) and open (red) in S.typhi and E. coli pore.

We further analyzed the cross section and Ires% signals (Figure 6.2 d and e) to examine if two pores can distinguish the different structural state of proteases. In line with the binding states identified from the free energy surface, the average cross section profiles of the proteases in two pores are very identical, owing to the dominance of very similar binding states. Although the average blockage of two conformations in the S.typhi pore are slightly different (mostly due to the morphology), the Ires% signals between proteases two conformations are indistinguishable. Overall, the simulations at the specific pore middle region suggest a more stable binding of proteases two conformations in the E.coli pore, as indicated by the number of binding states derived from the free energy surface. This enhanced stability may be contributed to the neutral A57 site, which avoids destabilizing the popular binding configurations, especially S3. However, due to the high consistent binding patterns adopted by the proteases open and closed structures, subtle differences have been observed in the Ires% signals between two conformational states.

6.3.3 Computational designed S.typhi pore resolves different proteases states

To further validate our observations and hypothesis regarding the significance of site 57 on ClyA in stabilizing protein/pore interactions, we designed two mutants, S.typhi E57A pore and S.typhi E57K pore. The aim was to examine whether proteases would have more stable bindings in the engineered pores. Specifically, we aimed to investigate if this stabilization is achieved by a higher probability of S3 binding configuration occurrence, facilitated by the negative-to-neutral, more dramatically, by the negative-to-positive mutants. As shown in Figure 6.3, much cleaner
binding states have been observed in both engineered pores. Consistent with our proposal, the S3 state is dominant in all trajectories, and the *S.typhi* E57K pore leads to less heterogeneous binding dynamics. This observation aligns with the charge interaction patterns we showed in Figure 6.2c. Obviously, the E57K can contribute to the stabilization of the third charge “ring” for five negative residues on proteases surface.

As for the dynamic of proteases in the open conformation in *S.typhi* E57A pore, although we cannot fully elucidate why the S2 state is more favored by the open conformation (also observed in *S.typhi* WT pore), the preference for different binding configurations makes this E-to-A engineered pore even more powerful. It enables the resolution of the two dynamic states of proteases. This observation was further quantified by comparing the Ires% signal between two conformations in the same pore (Figure 6.3 b and c). Clearly, the S2 bound configuration for the open state has different cross section values and lead to smaller Ires% compared to the closed state. By specifically engineering the 57 site, the *S.typhi* E57A pore lowered the probability of unfavorable proteases/pore interactions, resulting in more robust signals corresponding to two distinct binding configurations representing NS2B/NS3 two different conformational states, which could be instrumental in helping the antiviral-drug development.
Figure 6.2 Mid-binding states and properties of proteases in *S.typhi* E57A and *S.typhi* E57K pores. a). The PCA clustering of proteases/ClyA (*S.typhi* E57A and *S.typhi* E57K) binding states from standard MD simulations at 400 K. b). The blockage of closed (blue) and open (red) proteases in *S.typhi* E57A and *S.typhi* E57K pore. c) The Ires% of closed (blue) and open (red) proteases in *S.typhi* E57A and *S.typhi* E57K pore.

6.4 Conclusions

Targeting the conformational dynamics of WNV NS2B/NS3 is significant for effective antiviral drug development. In this work, we utilize the coarse-grained simulation to gain molecular insights on how NS2B/NS3 will dynamically interact with the ClyA nanopore. From the pulling simulation, we found the mid-region of the pore, where several charged residues locate, is prone to interact with the proteases. However, although both *E.coli* and *S.typhi* pores have such mid-binding events, the resident time in *E.coli* pore is much longer than its in *S.typhi* pore. Further simulations at the pore middle region indicate significant binding configurations, where the 57 site on ClyA could play a significant role by stabilizing/destabilizing the protease/pore charge interactions. Therefore, we made two computational designs to mutate the E57 site on *S.typhi* pore to E57A and E57K, respectively. Much stable binding was observed in the engineered pores with the S3 binding configuration most dominant in closed structured NS2B/NS3 bindings, which again suggest the significant roles of 57 site on WNV proteases/ClyA interactions. Particularly, the open and closed conformations of proteases prefer distinct binding configuration in the *S.typhi* E57A pore, resulting in different Ires% signal. This finding is promising to be further applied to develop effective antiviral drugs.
Figure 6.4 Illustration of steered MD simulations pulling WNV NS2B/NS3 proteases through *S. typhi* (green) and *E. coli* (grey) ClyA. a) The inner pore profiles of *S. typhi* (green) and *E. coli* (grey) ClyA. Pore residues are drawn by residue type, which are blue for positive charged, red for negative charged, white for nonpolar and polar residues. b) Illustration of steered MD, where the proteases were initially putting 4.5 nm away from the top boundary of ClyA with different orientations c). The proteases would be pulled down by a constant force of ~60 pN at the center of mass (purple dot). c) The 10 different orientations of proteases before pulling simulations. The proteases are drawn in cartoon style, with NS3 highlighted cyan, NS2B highlighted orange, and the NS2B-CTD sheet at active site colored red.
Figure 6.5 Illustration of standard MD simulations to explore protease dynamics in the middle of the ClyA (highlighted region) in ClyA. The transparent region of ClyA is cleaved out to save computational time. The initial conformation of the six different rotations of proteases is shown in the right panel (r1 to r6). The proteases follow the same drawn and coloring style as Figure 6.4.
CHAPTER 7
SUMMARY AND FUTURE DIRECTIONS

7.1 Summary

It has been well recognized that IDP conformations, dynamics, and interactions are significant in cellular activities. However, how their structural plasticity is linked with their functions remain unclear. In this dissertation, the multi-scale molecular dynamic simulations have been performed to study the structures, dynamics, interactions, and functions of IDPs. In chapter 2, we performed exhaustive atomistic simulations to investigate how SPIN execute its inhibition capacity via coupled binding and folding. We first performed the dissociation simulations at high temperature to correspondingly interpret the folding-upon-binding process. The free-energy landscapes indicate different mechanism for two SPIN species (SPIN-\textit{aureus} and SPIN-\textit{delphini}). We then carried out simulations for SPIN-NTD along at room-temperature to compare the stability between two species. Our results indicate that SPIN-\textit{delphini} NTD has stronger stability in the unbound state, which is prone to adopt the conformational selection-like binding mechanism. We propose that this stronger stability of SPIN-\textit{delphini} NTD helps to significantly improve its inhibitory capacity. In chapter 3, we developed a new enhanced sampling method REST3 by calibrating the protein-protein and protein-water balance. We found that REST3 achieves more efficient temperature random walk compared to REST2 due to the avoid of the conformational segregation, which mainly comes from the conformational collapse at high temperature conditions in REST2. Although no big differences have been observed between the conformational ensembles derived from REST2 and REST3 two protocols for model peptides, REST3 allows for using smaller number of replicas to cover similar effective temperature range.
Our group have developed a hybrid resolution coarse-grained protein model (HyRes), which has atomistic backbone and coarse-grained sidechains. The HyRes model can qualitatively describe the long-range non-specific interactions and semi-quantitatively re-capitulate the secondary structures of IDPs. However, the original HyRes model lacks an accurate description of solvation effects and would generate overly compact protein ensembles. Therefore, in chapter 4, we optimized the original HyRes model and developed the new HyRes II model, which achieves more accurate descriptions of disordered protein chain dimensions and secondary structures. We showed that the optimization on HyRes II will not lead to the mistakenly extension of folded protein chain dimensions but more accurate sampling of IDP dynamics and interactions. We have further encoded HyRes II to allow for GPU acceleration as discussed in chapter 5, which help us to access the long time-scale events of biological condensation. Simulations of the phase separation of GY-23 prove that HyRes-GPU is able to simulate the spontaneous condensation efficiently with significant properties well converged within 1 \( \mu s \) and accurate enough to re-capitulate the non-trivial mutational effects consistent with experiments. Particularly, we have observed an increased formation of \( \beta \)-structures in GY-23 condensates, which is consistent with CD data. The further simulation of TDP-43 CR phase separation showed a high correlation between the simulation predicted \( C_{\text{sat}} \) value and experimental characterizations. By designing specific mutation and controlling the secondary structure, our simulations indicate that the helical properties of TDP-43 CR play an indirect role on regulating its phase separation. On the other hand, the balance between backbone availability and residual structure together controls the condensation capacity of TDP-43 CR. We also used the HyRes model to study the dynamics of WNV proteases in the nanopore. We first identified a favorable binding region on ClyA mid-lumen, and further found that the E57 site play a crucial role in stabilizing the proteases/pore
interactions. By specifically engineering the pore, we showed that the open and closed two structures of proteases prefers different binding states in the \textit{S.typhi} E57A ClyA pore. The different Ires% signal can help to resolve the two structural states of WNV NS2B/NS3 proteases.

7.2 Future Direction

A substantial portion of this dissertation delves into the exploration of the structural dynamics and interactions inherent to intrinsically disordered proteins (IDPs). Employing multiscale molecular dynamics (MD) simulations at both atomistic and coarse-grained levels, the primary objective is to unravel the intricate functioning of IDPs in biological activities, encompassing crucial processes like signaling and regulation. A particular focus of this investigation extends to deciphering how IDP secondary structure will be connected to or influence their biological condensation ability. Despite the strides discussed in this dissertation, it is well-acknowledged that accurately simulating IDP-related activities remains a formidable challenge. The recognized limitations are multifaceted, including the significant gap between the functional time-scale of proteins and the accessible simulation time, lacking of accuracy and transferability in methodologies, resulting in insufficient sampling of the IDP conformational ensemble and limitations in applicability across diverse biological systems. These challenges necessitate a concerted effort in advancing both force fields and enhanced sampling methods. In one hand, addressing the limitations in sampling requires innovative methods to capture the diverse conformational states of IDPs more effectively by overcoming the both the enthalpic and entropic barrier. On the other hand, the more transferable force fields are crucial to gain insights derived from IDP studies extended to various biological contexts, which enable simulations that mirror the complexity of biological systems.
To overcome the inherent limitations related to entropy and enthalpy in IDP simulations, our group has proposed a novel enhanced sampling technique known as the Multiscale Enhanced Sampling (MSES) method. This innovative approach aims to drive more efficient global or local folding of peptides exhibiting $\alpha$-$\beta$-structures. MSES uniquely combines coarse-grained (CG) modeling with atomistic (AT) simulations through a coupling potential. This integration accelerates atomistic-level conformational transitions by leveraging the fast dynamics at the coarse-grained level, where entropic barriers are lower. Simultaneously, MSES employs T-REX to effectively overcome the enthalpic barriers. However, the effectiveness of MSES relies on the ability to achieve an effective coupling between AT and CG models, necessitating highly correlated ensembles generated from CG simulations with the AT counterparts. Thus, CG models capable of consistently simulating structural fluctuations akin to AT models become imperative for enhancing the efficiency of MSES. The coupling between HyRes and AT has already demonstrated an increased conformational transition rate, exemplified by the $\alpha$-structured $(\text{AAQAA})_3$, compared to traditional Go-like models. Consequently, there is a clear trajectory for advancing CG models to better describe IDP conformations, particularly the residual structures, to propel more efficient MSES simulations.

As for force field optimization, one promising direction is to incorporate the parametrization process with the machine learning (ML) method. Leveraging advanced algorithms like neural networks, optimized CG models’ capacity to re-capitulate the chain dimensions of a diverse range of IDPs can be dramatically improved, with correlations up to 0.9 when compared to experimental data. The ML-parametrization approach serves as an optimal strategy to rapidly enhance model accuracy and extend applications to a broader system of interests. Taking the HyRes model, as discussed in this dissertation, as an example, ML-based optimization holds the
potential to elevate its accuracy on describing IDP global conformations, enabling its application to a more diverse range of systems for studying biological condensation. This could involve a deeper exploration of the intricate phase separation mechanisms observed in elastin-like peptides (ELPs). Notably, the phase separation of several ELPs, including GY-23, has been characterized by an increase in $\beta$-structures experimentally, which suggests a potential role for secondary structure in stabilizing their condensation. Furthermore, extending the application of HyRes to include biological molecules like RNA is of fundamental importance. This expansion offers the opportunity for a more comprehensive exploration of the heterotypic phase separation that IDP involved, unlocking new dimensions in our understanding of the multivalent network between disordered proteins and other biological molecules.


