

Analytical Framework to Uncover Structural Signatures from All-Atom Molecular Dynamics Simulations

by

Wai Shing Tang

B.Sc., The Chinese University of Hong Kong, Hong Kong, 2017

M.Sc., Brown University, Providence, Rhode Island, 2019

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Physics at Brown University

Providence, Rhode Island

May 2022

© Copyright 2022 by Wai Shing Tang

This dissertation by Wai Shing Tang is accepted in its present form by
the Department of Physics as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____

Professor Lorin Crawford, Advisor

Recommended to the Graduate Council

Date _____

Professor Brenda M. Rubenstein, Reader

Date _____

Professor Jay X. Tang, Reader

Approved by the Graduate Council

Date _____

Professor Andrew. G. Campbell
Dean of the Graduate School

Wai Shing, Tang

Curriculum Vitae

Education

- **Brown University, Providence, Rhode Island, USA, 2017-2022**
 - Ph.D. in Physics, Department of Physics (May 2022)
 - M.Sc. in Physics, Department of Physics (May 2019)
- **The Chinese University of Hong Kong, Shatin, Hong Kong, 2013-2017**
 - B.Sc. in Physics, Department of Physics (May 2017)

List of Publications

1. **W.S. Tang**, G.M. da Silva, H. Kirveslahti, E. Skeens, B. Feng, T. Sudijono, K.K. Yang, S. Mukherjee, B. Rubenstein, and L. Crawford (2022). A Topological Data Analytic Approach for Discovering Biophysical Signatures in Protein Dynamics. *PLOS Computational Biology*. In press.
2. A.C. Murthy, **W.S. Tang**, N. Jovic, A. Janke, D.H. Seo, T. Perdikari, J. Mittal, and N.L. Fawzi (2021). Molecular interactions contributing to FUS SYGQ LC/RGG phase separation and co-partitioning with RNA polymerase II heptads. *Nature Structural & Molecular Biology*. 28, 923–935.

3. A.S. Bock, A.C. Murthy, **W.S. Tang**, N. Jovic, F. Shewmaker, J. Mittal, and N.L. Fawzi (2021). N-terminal acetylation modestly enhances phase separation and reduces aggregation of the low-complexity domain of RNA-binding protein FUS. *Protein Science*.
4. **W.S. Tang**, N.L. Fawzi, and J. Mittal (2020). Refining All-Atom Protein Force Fields for Polar-Rich, Prion-like, Low-Complexity Intrinsically Disordered Proteins. *Journal of Physical Chemistry B*. 124(43): 9505-9512.
5. B.S. Schuster, G.L. Dignon, **W.S. Tang**, F.M. Kelley, A.K. Ranganath, C.N. Jahnke, A.G. Simpkins, R.M. Regy, D.A. Hammer, M.C. Good, and J. Mittal (2020). Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proceedings of the National Academy of Sciences*. 117(21): 11421-11431.
6. A. Yang, **W.S. Tang**, T. Si, and J.X. Tang (2017). Influence of Physical Effects on the Swarming Motility of *Pseudomonas Aeruginosa*. *Biophysical Journal*. 112(7): 1462-1471.

Abstract of “Analytical Framework to Uncover Structural Signatures from All-Atom Molecular Dynamics Simulations”, by Wai Shing Tang, Ph.D., Brown University, May 2022.

Structural features of proteins often serve as signatures of their biological function and molecular binding activity. Elucidating these structural and dynamical features is essential for understanding underlying biophysical mechanisms. Significant efforts in the past decade have resulted in highly accurate all-atom protein force fields for molecular dynamics (MD) simulations of proteins. These simulations provide new insights into the underlying physics of both folded proteins, which are easy to characterize by one or a few native states, and intrinsically disordered proteins (IDPs), which are structurally heterogeneous and challenging to characterize by experiments alone. However, state-of-the-art protein force fields still display significant discrepancies in MD simulation compared to experiments. Moreover, given accurate structural ensembles obtained from MD simulation, existing methods do not have the statistical capability to robustly identify geometric and dynamic variations in these simulated protein ensembles or describe the molecular interactions underlying the biophysical behavior of the protein systems.

In this thesis, I present the analytical approaches for optimizing the accuracy of MD simulations and extracting structural signatures from the simulated ensembles. In the following chapters, I present (1) a refinement to the all-atom molecular dynamic simulation force field to improve the representation of helicity for intrinsically disordered proteins, (2) a statistical approach to quantify and visualize the molecular interactions underlying the liquid-liquid phase separation of intrinsically disordered protein systems, and (3) a novel topological data analytic approach for discovering biophysical signatures in protein dynamics.

Acknowledgments

We live in an unusual era. COVID has changed things we took for granted and posed unprecedented challenges to us. Without the help of all of the people around me, I would not have been able to overcome these challenges. They made me who I am today.

First and foremost, I would like to express my sincere gratitude to my thesis advisor, Professor Lorin Crawford. Lorin is one of the most considerate advisors I have met. He always gave me his best support to overcome the challenges and gave me a lot of freedom in research. Whenever I was stuck, He advised me to steer my research in a productive direction. During the COVID work-from-home period, it was challenging for advisors and students to sustain a healthy relationship since it is challenging to communicate virtually. Lorin showed me his patience whenever I strumbled and prioritized my personal well-being. A year after I joined his lab, we finally had a chance to meet in person at the office. Now that I am leaving, I wish I had more opportunities to know more about Lorin in person and to learn more from Lorin academically. Without Lorin, I would have given up on my Ph.D. study, and I would not be writing this thesis.

I would like to thank Jay, Professor Jay X. Tang, my mentor, who knew me before I started my Ph.D. program at Brown. When I was still at the Chinese University of Hong Kong, I spent a summer at Brown, 2015, as a summer research intern, working with Jay on experimental work for bacterial motility. Two years later, I came back to Brown for my Ph.D. in Physics. Jay gave his best academic and life advice during my summer internship and my Ph.D. program using his decades of faculty experience.

I would like to thank Professor Brenda M. Rubenstein and her Ph.D. student, Gabriel. Brenda and Gabriel are the closest collaborators on my thesis project. Brenda is one of the most talented scientists I have ever worked with. Her span of expertise ranges from quantum chemistry to molecular biophysics. Gabriel is one of the most hard-working and responsible collaborators I have ever worked

with. Together we always come up with creative ideas and execute them swiftly.

I would like to thank Professor Nicholas L. Fawzi and Professor Jeetain Mittal. They advised me for 2 years during my Ph.D. program, allowed me to participate in high-impact research on intrinsically disordered proteins, and built my knowledge foundation in computational protein research.

I would also like to thank my colleagues at the Crawford Lab. We did not get to know each other very well during COVID. I still enjoy the small talk every time we met virtually and in person when we bumped into each other in the office. I would like to thank my colleagues at Fawzi Lab, especially Anastasia Murthy and Myrto Perdikari. I enjoyed working with them a lot when I was at the Fawzi Lab, and they accompanied me and helped me navigate through the most challenging time in my personal life.

I would like to thank everyone at Brown Taekwondo Club and my beloved coach, Master Sung S. Park. Brown Taekwondo Club is not just about Taekwondo. It is also about the family and supporting each other. Without their support, I would not have made it through all the challenges I faced during the past 5 years. Master Park is my life mentor who is always there to listen and coach. He knows me much better than I know myself. He tries his very best to teach me everything he knows about Taekwondo and life. I was honored to be invited to become an instructor. Being a leader and mentor in the club, I learned how to teach and take care of myself and the people around me.

Last but not least, I would like to thank my family for their love and support. They raised me with love and care. Most importantly, they allowed me the freedom of choice to pursue the career of being a scientist. Being 8000 miles away from my family, traveling back to Hong Kong was impractical due to COVID. I got very little chance to spend time with my family, yet they are still, and will forever be, the most indispensable part of my life.

Contents

List of Tables	xi
List of Figures	xii
1 Introduction	1
2 Residue-level refinement of all-atom protein force field for intrinsically disordered proteins	5
2.1 Introduction	5
2.2 Method	8
2.3 Result	9
2.4 Conclusion	13
2.5 Tables and Figures	14
3 Quantifying Molecular interactions contributing to liquid-liquid phase separation of intrinsically disordered protein systems	18
3.1 Introduction	19
3.2 Method for Molecular Dynamic Simulation and Analysis	21
3.3 Results	23
3.4 Discussion	27
3.5 Tables and Figures	29
4 SINATRA Pro: A Topological Data Analytic Approach for Discovering Biophysical Signatures in Protein Dynamics	36
4.1 Introduction	36

4.2	Material and Methods	48
4.3	Results	54
4.4	Discussion	64
4.5	Tables and Figures	66
5	Future directions and Preliminary results	76
5.1	A Topological Data Analytic Approach for Structural alignment for protein structural ensembles and distant homologs	76
5.2	Topological Data Analytic Framework for Evaluating and Predicting Protein Conformational Transition	81
5.3	Gaussian Process Regression Framework for identifying biophysically relevant structures	82
5.4	Tables and Figures	83
6	Discussion and Conclusion	88
A	Supporting Material for “Residue-level refinement of all-atom protein force field for intrinsically disordered proteins”	90
B	Supplementary material for “SINATRA Pro: A Topological Data Analytic Approach for Discovering Biophysical Signatures in Protein Dynamics”	103

List of Tables

2.1	Root-mean-squared deviation (RMSD) of simulated $\Delta\delta C_\alpha$, $\Delta\delta C_\beta$, $\Delta\delta C_\alpha - \Delta\delta C_\beta$ (in ppm) from NMR ^{13}C chemical shifts with various force fields	14
4.1	General guidelines for choosing values for the free parameters in the SINATRA Pro pipeline software	66
4.2	Detailed overview of the different protein systems analyzed in this study	67
4.3	Null hypothesis experiment to evaluate SINATRA Pro's ability to find regions of interest (ROI) in each of the proteins analyzed in this study	68
5.1	Root-mean-squared deviation (RMSD) with respect to manual alignment	83
A.1	Protein sequences used in simulations	91
A.2	Summary of simulations performed using amber ff03ws, where N is number of residues in given protein or peptide	91
A.3	Summary of simulations performed using modified force fields	91
A.4	Root-mean-squared deviation (RMSD) of simulated $\Delta\delta C_\alpha$, $\Delta\delta C_\beta$, $\Delta\delta C_\alpha - \Delta\delta C_\beta$ (in ppm) from NMR ^{13}C chemical shifts with various force fields.	92
B.1	Empirical runtimes for running the SINATRA algorithm as a function of its free parameters and inputs	130

List of Figures

2.1	Chemical shifts and α -helix propensities of TDP-43 _{310–350} and FUS _{0–43} in simulations and experiments	15
2.2	Deviation of chemical shifts $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta)$ (in ppm) in simulations from NMR ¹³ C chemical shifts	16
2.3	Secondary structure propensities of FUS _{0–43} , TDP- 43 _{310–350} , and RNA Pol II _{1927–1970} in simulations and experiments	17
3.1	Composition of fragments used for all-atom simulations	30
3.2	Contacts between FUS SYGQ LC and RGG domains	31
3.3	Diverse interaction modes contribute to FUS SYGQ LC and RGG3 interactions within the condensed phase	32
3.4	Lysine residues within RNA Pol II CTD contribute to interactions with FUS within the condensed phase	33
3.5	Contacts within a three-component phase containing FUS SYGQ LC and RGG3 and RNAP2 CTD	34
3.6	Multiple interaction modes contribute to multi-component condensed phases of FET proteins and RNA polymerase II	35
4.1	Schematic overview of SINATRA Pro: a novel framework for discovering biophysical signatures that differentiate classes of proteins	69
4.2	Receiver operating characteristic (ROC) curves comparing the power and robustness of SINATRA Pro to competing 3D mapping approaches in controlled molecular dynamic (MD) simulations	70

4.3	Real data analyses aimed at detecting structural changes in the Ω -loop of β -lactamase (TEM) induced by an R164S mutation	71
4.4	Real data analyses recover structural changes in the flap region of HIV-1 protease driven by a Ile50Val mutation	72
4.5	Real data analyses aimed at detecting structural changes in Domain 2 of the elongation factor EF-Tu upon guanosine triphosphate (GTP) hydrolysis	73
4.6	Real data analyses identify enrichment in the N-terminal pocket of the Abl1 Tyrosine protein kinase due to a M290A mutation in the α C helix	74
4.7	Real data analyses aimed at detecting uncoiling of the superhelix in Importin- β upon release of an IBB peptide	75
5.1	Protein structural alignment using directional DEC curves to align directions with similar topological profile	84
5.2	Coarse Alignment scheme to remove structural outliers and align topologically relevant structures	85
5.3	EC align recreate manual alignment in manually-curated protein distant homologs dataset MALISAM	86
5.4	SINATRA Pro results on Abl1 Transition trajectory	87
A.1	Comparison of α -helix propensities and NMR ^{13}C chemical shifts between experiments and simulations	94
A.2	Comparison of NMR ^{13}C chemical shifts between experiments and simulations	95
A.3	Secondary structure propensities of FUS _{0–43} , TDP-43 _{310–350} , and RNA Pol II _{1927–1970} with ff99SBws, ff99SBws-STQ, ff99SB-disp and experiments	96
A.4	Intrachain residue-to-residue contact propensities in FUS _{0–43} and TDP-43 _{310–350} with ff99SBws and ff99SBws-STQ	97
A.5	Change in Intrachain residue-to-residue contact propensities in FUS _{0–43} and TDP-43 _{310–350} from ff99SBws to ff99SBws-STQ	98
A.6	Average radius of gyration (Rg) in FUS _{0–43} , TDP-43 _{310–350} and RNA Pol II _{1927–1970} with ff99SBws and ff99SBws-STQ	99
A.7	DSSP secondary structure as a function of simulation time for each demultiplexed replica moving through the temperature space of FUS _{0–43} with ff99SBws-STQ	100

A.8	DSSP secondary structure as a function of simulation time for each demultiplexed replica moving through the temperature space of TDP-43 _{310–350} with ff99SBws-STQ	101
A.9	Fraction Helix of (AAQAA) ₃ with ff99SBws and ff99SBws-STQ (this study) compared to NMR experimental data	102
B.1	Power and sensitivity analysis assessing the robustness of SINATRA Pro to different free parameter settings in controlled molecular dynamic (MD) simulations	104
B.2	Effect of realignment after displacement in the controlled experiments aimed at assessing the ability of SINATRA Pro to detect artificial changes in the Ω -loop of β -lactamase (blue region)	105
B.3	Receiver operating characteristic (ROC) assessing the differentiating power of PCA under different model parameter configurations in controlled molecular dynamic (MD) simulations	106
B.4	Receiver operating characteristic (ROC) assessing the differentiating power of the Neural Network under different model parameter and architecture configurations in controlled molecular dynamic (MD) simulations	107
B.5	Atomic-level results for detecting biophysical signatures in (top row) TEM β -lactamase, (middle row) HIV-1 protease bound to Amprenavir, and (bottom row) GTP-bound EF-Tu	108
B.6	Atomic-level results for detecting biophysical signatures in (top row) Abl1 and (bottom row) IBB-bound Importin- β	109
B.7	Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in the Ω -loop of TEM β -lactamase induced by the Arg164Ser mutation using SINATRA Pro	110
B.8	Real data analyses aimed at detecting structural changes in the Ω -loop of TEM β -lactamase induced by the Arg164Ser mutation using atomic-level regularization with Elastic Net classification	111
B.9	SINATRA Pro analysis on the dimeric form of HIV-1 protease to detect structural change in the flap region driven by a Ile50Val mutation	112

B.10 Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in the flap region of HIV-1 protease driven by the Ile50Val mutation using SINATRA Pro	113
B.11 Real data analyses aimed at detecting structural changes in the flap region of HIV-1 protease driven by the Ile50Val mutation using atomic-level regularization with Elastic Net classification	114
B.12 Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in Domain 2 of the elongation factor EF-Tu upon guanosine triphosphate (GTP) hydrolysis using SINATRA Pro	115
B.13 Real data analyses aimed at detecting structural changes in Domain 2 of the elongation factor EF-Tu upon guanosine triphosphate (GTP) hydrolysis using atomic-level regularization with Elastic Net classification	116
B.14 Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in the N-terminal pocket of the Abl1 Tyrosine protein kinase due to the Met290Val mutation in the α C helix using SINATRA Pro	117
B.15 Real data analyses aimed at detecting structural changes in the N-terminal pocket of the Abl1 Tyrosine protein kinase due to the Met290Val mutation in the α C helix using atomic-level regularization with Elastic Net classification	118
B.16 Sensitivity analyses on different lengths of MD simulations aimed at detecting uncoiling of the superhelix in Importin- β upon release of an IBB peptide using SINATRA Pro	119
B.17 Real data analyses at detecting uncoiling of the superhelix in Importin- β upon release of IBB using atomic-level regularization with Elastic Net classification	120
B.18 Residue-level results from running root mean square fluctuation (RMSF) analysis on TEM β -lactamase, HIV-1 protease, and GTP-bound EF-Tu	121
B.19 Residue-level results from running root mean square fluctuation (RMSF) analysis on Abl1 and IBB-bound Importin- β	122
B.20 Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cut-offs r values used to construct the simplicial complexes for TEM β -lactamase	123
B.21 Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cut-offs r values used to construct the simplicial complexes for HIV-1 protease	124

B.22 Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cut-	
offs r values used to construct the simplicial complexes for EF-Tu	125
B.23 Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cut-	
offs r values used to construct the simplicial complexes for Abl1	126
B.24 Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cut-	
offs r values used to construct the simplicial complexes for Importin- β	127
B.25 Real data analysis results demonstrating the idea of running SINATRA Pro with	
sequence-independent (or correspondence free) structural alignment based on topo-	
logical summary statistics	128
B.26 Real data analysis results demonstrating the idea of running SINATRA Pro with	
sequence-independent (or correspondence free) structural alignment based on topo-	
logical summary statistics	129

Chapter 1

Introduction

1.0.1 All-Atom Protein Force Fields for IDPs

Significant efforts in the past decade have given us highly accurate all-atom protein force fields for molecular dynamics (MD) simulations of folded and disordered proteins. These simulations, complemented with experimental data, provide new insights into molecular interactions that underlie the physical properties of proteins, especially for intrinsically disordered proteins (IDPs), for which defining the heterogeneous structural ensemble is hugely challenging by experiments alone. Consequently, the accuracy of these protein force fields is of utmost importance to ensure reliable simulated conformational data.

In Chapter 2, I present my work on refining the all-atom protein force fields for polar-rich, prion-like, low complexity intrinsically disordered proteins [1]. The accuracy of current state-of-the-art force fields for IDPs (ff99SBws and ff03ws) was assessed by MD simulation of disordered proteins of low amino acid sequence complexity that can undergo liquid-liquid phase separation. A detailed comparison of NMR chemical shifts between simulation and experiment on several IDPs shows that regions surrounding specific polar residues contribute to simulated ensembles with exaggerated helicity compared to the experiments. To resolve this discrepancy, I introduced residue-specific modifications to the backbone torsion potential of three residues (Ser, Thr, and Gln) in the ff99SBws force field. The modified force field, ff99SBws-STQ, provides a more accurate representation of helical structure propensity in these LC domains without compromising the faithful representation of helicity in a region with distinct sequence composition. Our refinement strategy also suggests a

path forward for integrating experimental data to assess residue-specific deficiencies in the current physics-based force fields and improve them further for their broader applicability. Supplementary materials for this chapter are presented in Appendix A.

1.0.2 Quantifying Molecular interactions contributing to liquid-liquid phase separation of intrinsically disordered protein systems

Biomolecular condensates are formed by liquid-liquid phase separation (LLPS) of a mixture of proteins and nucleic acids with distinct sequence compositions. The RNA-binding protein FUS in Sarcoma (FUS) mediates LLPS and plays a role in transcription by clustering with RNA polymerase II (RNA Pol II). In Chapter 3, adapted from this paper [2], I present my computational effort for providing a biophysical molecular model for Liquid-Liquid Phase Separation (LLPS) of the RNA binding proteins FUS and RNA Pol II. This work revealed the liquid-liquid phase separation among the prion-like N-terminal low complexity domain (SYGQ LC), and RGG domains of FUS, and C-terminal heptad repeats of RNA Pol II are contributed by weak and distributed molecular interactions among various amino acid types.

Results from Nuclear magnetic resonance (NMR) spectroscopy and molecular dynamics (MD) simulations demonstrated that interactions involving many residue types, not only among arginine and tyrosine, form contacts with several additive interaction modes beyond previously appreciated sp^2/π and cation- π interactions, contributing to LLPS of FUS liquid condensates. In addition, lysine residues within the heptad repeats of RNA Pol II mediate cation- π and hydrogen-bonding interactions with FUS SYGQ LC. Our data point to a model where weak, distributed interactions between the intrinsically disordered SYGQ LC and RGG intrinsically disordered regions stabilize LLPS of FUS and its co-phase separation with RNA polymerase II.

1.0.3 Discovering Biophysical Signatures in Protein Dynamics using Topological Data Analytic Approach

Structural features of proteins often serve as signatures of their biological function and molecular binding activity. Nonetheless, identifying structural differences among proteins can be a non-trivial task. Elucidating these structural features is essential for a complete understanding of biophysical mechanisms underlying the physiological behavior of proteins. When contrasting ensembles of

protein structures obtained from molecular dynamics simulations, spurious fluctuations can easily overshadow biologically-relevant features. While there are existing methods to identify structural differences between protein variants, such methods cannot jointly infer both geometric and dynamic changes simultaneously.

In Chapter 4 of this thesis, I present SINATRA Pro, a computational framework for extracting crucial structural features between two sets of proteins [3]. SINATRA Pro robustly outperforms standard techniques in pinpointing the physical locations of both static and dynamic signatures across various types of protein ensembles with improved resolution. Algorithmically, SINATRA Pro works by first taking in the 3D atomic coordinates for each protein snapshot and summarizing them according to their underlying topology. Statistically significant topological features are projected back onto a user-selected representative protein structure, thus facilitating the visual identification of biophysical signatures of different protein ensembles. The ability of SINATRA Pro to detect minute conformational changes in five independent protein systems of varying complexities were assessed. In all test cases, SINATRA Pro identifies known structural features validated by previous experimental and computational studies and novel features that are also likely to be biologically relevant according to the literature. These results highlight SINATRA Pro as a promising method for facilitating the non-trivial task of pattern recognition in trajectories resulting from molecular dynamics simulations with substantially increased resolution. Supplementary materials for this chapter are presented in Appendix B.

1.0.4 Extended Application of Topological Data Analytic approach on Protein Analysis

Protein structure alignment is a fundamental problem in computational structural biology. It is especially challenging to compare protein structures with low sequence similarity. Structure alignment methods align any given protein structures using structural information such as secondary structures and distance maps. While prior biological knowledge such as evolutionary distances can improve the alignment accuracy of these algorithms, current algorithms still produce faulty alignment, hindering the quality of downstream analyses for heterogeneous sets of protein systems.

In Chapter 5, I extended the topological data analytic approach on protein structural alignment and develop a topology-based alignment method, EC align. I utilized an Euler-characteristic-based

topological summary statistic to summarize the local and global topology of the protein backbone. I then used this information to match relevant components between protein structures for more precise alignment. Results show that EC align can robustly and efficiently align protein homologs with low sequence identity. I also present two extended applications for this topological data analysis approach: (1) Topological data analytic framework for evaluating and predicting protein conformational transition and (2) Gaussian process regression framework for identifying biophysically relevant structures.

Chapter 2

Residue-level refinement of all-atom protein force field for intrinsically disordered proteins

This chapter is adapted from this paper [1] on refining all-atom protein force fields for polar-rich, prion-like, low complexity intrinsically disordered proteins. I would like to acknowledge my co-advisors, Jeetain Mittal, and Nicolas L. Fawzi, for their contributions to the work presented in this chapter.

2.1 Introduction

Intrinsically disordered proteins (IDPs) and regions (IDRs) are biologically functional without adopting a single well-defined folded structure.[4] They are present in a significant fraction in the proteome of all organisms [4–6] and participate in essential physiological and pathological functions including stress response [7] and signaling [8]. They are also involved in cell regulation and various neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) through the formation of protein aggregates [9]. IDPs are also involved in cellular liquid-liquid phase separation (LLPS), which underlies the formation of some membraneless organelles [10, 11] and may serve as precursors for pathogenic aggregates [12, 13]. Many IDPs can contain partial secondary

structures, including helices [14, 15] and β -sheet structures [16, 17], that can contribute to their intermolecular interactions and self-assembly processes. Therefore, accurately probing these transient secondary structures in experiments and simulations is essential for understanding the underlying molecular mechanisms of protein assembly in LLPS.

Experimental techniques such as X-ray crystallography and cryogenic electron microscopy (cryo-EM) can resolve folded protein structures down to the angstrom scale but are unable to provide detailed spatiotemporal information about the equilibrium structural ensembles of disordered proteins because these techniques and sample preparations quench and/or cannot observe the heterogeneous nature and rapid interconversion between a large number of microstates [18]. Nuclear magnetic resonance (NMR) in solution has provided a wealth of information about the local structural and dynamical properties of IDPs to generate important insights into their biological function. It is often helpful to combine NMR data with molecular modeling techniques to generate conformational ensembles consistent with the experimental observables such as chemical shifts and J-coupling to obtain atomic details obscured by averaging [19–23]. It is also highly desirable to obtain the necessary information about IDP properties directly from physics-based all-atom transferable force fields without requiring any experimental input on the system of interest. Such models can provide unprecedented details on the structural and dynamical properties as well as atomic interactions responsible for the observed behavior [24, 25]. This can be critical in establishing IDP sequence-structure/motions-function relationships and provide hypotheses for experimental work [26]. With the rapid development of computer hardware and advanced sampling techniques to conduct molecular dynamics (MD) simulations based on atomistic protein force fields over the past decade, it became apparent that improvements in their accuracy and transferability were necessary to realize the potential for molecular simulations to serve as a “computational microscope” [27, 28]. Progressive efforts in the past 10-15 years have resulted in significant improvements in the accuracy of all-atom protein force fields beginning with a substantial step forward on achieving secondary structure balance and thereby demonstrating true transferability in terms of their application to proteins of different structural classes [29–32]. The empirical approach by Best and Hummer [29] used helical propensity data on a short peptide, Ac-(AAQAA)₃-NH₂, to achieve such secondary structure balance, and demonstrated that the potential energy functions only needed fine-tuning as opposed to a significant reparametrization as in previous force field development efforts. This strategy to modify only backbone torsion potentials yielded many improved force field variants, marked by an

asterisk next to their name, such as ff03*, ff99SB*, and charmm22* [29, 33]. In these force fields, modifications on backbone torsion potentials were applied uniformly to all amino acids (except proline and glycine) and were not tuned to capture any residue-specific biases in the force field. Since then, several studies have taken additional steps to make residue-specific corrections to side-chain [34] and backbone [35] torsion potentials as well as more elaborate bottoms-up parametrization of energy function parameters, [36–39] resulting in highly accurate protein force fields.

The use of these optimized force fields has been especially fruitful for studying IDPs by overcoming challenges of using the experimental techniques alone [40–43]. At the same time, the flat conformational energy landscapes of IDPs can make conformational ensembles more sensitive to systematic errors in the model. Therefore, simulations of IDPs have been used successfully for evaluating the remaining deficiencies of these force fields and providing roadmaps for their continued improvement [44]. For example, the unfolded states of proteins remained too collapsed in the force fields tuned for secondary structure balance alone (such as ff03*), which was also corroborated by available experimental data from single-molecule FRET and SAXS [45]. Some improvements in preventing unfolded state collapse were visible by changing the water model (TIP4P/2005) [46] in protein simulations while trying to improve the temperature cooperativity of the helix-coil transition and the associated solvation effects. The new protein force fields, indicated by the letter “w” for an improved water model TIP4P/2005 (ff03w and ff99SBw), provided significant improvement in protein-protein interactions but only modest success in capturing IDP dimensions [47]. Based on the accumulated evidence from other studies, protein-water interactions were scaled (strengthened) [48] to model the size of solvated disordered proteins accurately with force fields suffixed by “ws” such as ff03ws and ff99SBws [49]. Previous work based on these optimization strategies has leveraged the force field improvements to probe the secondary structure, single-chain configurations, and contact formation of IDPs and demonstrated their applicability to reproduce the structure and dynamics of unfolded proteins accurately [50–53]. At this point, it is instructive to ask if there are areas of concern about the applicability of these refined force fields for IDPs. The recent surge of interest in biomolecular LLPS and the underlying atomic interactions responsible for stabilizing the protein-rich condensed phase has motivated us to look at low-complexity (LC) IDPs [54], in which a few amino acid types dominate their composition, and at prion-like domains, named for their polar-rich residue sequence composition that resembles the composition of yeast prion proteins [55]. These LC IDPs present new challenges in conducting accurate simulations as any minor deficiencies in particular

amino acid types that dominate the sequence composition can propagate additively. Therefore, LC protein sequences are useful benchmarks to assess the accuracy of current state-of-the-art all-atom force fields.

2.2 Method

All-atom MD simulations were conducted on 44-residue fragments from the disordered regions of FUS LC (FUS LC_{0–43}, FUS LC_{37–97}, FUS LC_{41–84}, FUS LC_{77–120}, and FUS LC_{120–163}), TDP-43 CTD (TDP-43_{310–350}), RNA Pol II CTD heptads (RNA Pol II_{1927–1970}), and hnRNPA2 LC (hnRNPA2_{190–233} and hnRNPA2_{265–308}). The length of the peptides was chosen such that sampling the conformational space of the protein is computationally tractable. The GROMACS 4.6.7 MD engine [56] was used for simulations with PLUMED 2.2.4 plugin [57]. Simulations were conducted by using different variants of Amber force fields as specified in the text (see Tables S2-S4), using the original and modified versions of ff03ws and ff99SBws [49] which includes TIP4P/2005 explicit water molecules [46]. Production simulations were conducted by using replica exchange molecular dynamics (REMD) [58] in a well-tempered ensemble (WTE) [59, 60], a type of metadynamics method [61]. This combination is known as parallel tempering in the well-tempered ensemble (PT-WTE) and is useful for reducing the number of replicas required to explore the conformational space efficiently. The temperature is held constant for each replica by using Langevin dynamics as a thermostat with a friction constant of 1 ps⁻¹ and a time step of 2 fs. Electrostatic interactions are calculated by using the particle-mesh Ewald method [62] up to fourth order with a real-space cutoff distance of 1.0 nm. A 1.0 nm cutoff distance is used for the van der Waals interactions. For each protein, a random protein structure is generated and solvated in a truncated octahedron box with a face-to-face distance of 6.5 nm. The system is initially equilibrated for 100 ps at each temperature used for parallel tempering. In the PT-WTE simulations, 16 replicas at temperatures ranging from 300 to 500 K are constructed. Differences in temperatures between the replicas are determined such that the exchange acceptance probability is about 30%. The positions and velocities of the thermally equilibrated system are then used as the initial configuration of the PT-WTE simulations. WTE was applied by using Gaussian functions of width $\sigma_i = 750$ kJ/mol added at a $\tau_G = 20$ ps interval, with a bias factor of $\gamma = (T + \Delta T)/\Delta T = 36$ and an initial hill height of 2.5 kJ/mol. Chemical shifts and helical propensities in these simulations are calculated by using the SPARTA+ algorithm [63],

and the secondary structure is computed by using DSSP [64]. Experimental secondary structure propensities were also derived from NMR secondary chemical shifts by using the $\delta 2d$ software [65]. The RMSD (root-mean-squared deviation) from experimental chemical shifts (see text in the Supporting Information) was calculated as $(x_{\text{sim}} - x_{\text{expt}})^2/N$, where N is the number of residues sampled, x_{sim} is the simulated chemical shifts from SPARTA+, and x_{expt} is the NMR chemical shifts. For each peptide, two residues from each terminus and residues missing from experimental data are not included in the analysis. The RMSD is calculated for each residue type based on the chemical shift differences between simulations and experiments.

2.3 Result

I conducted parallel tempering (PT) MD simulations in a well-tempered ensemble (WTE) [59] to enhance conformational sampling (see the Methods section) with two suitable force fields for IDPs (ff03ws and ff99SBws) on a set of prion-like domains including domains from TDP-43, Fused in Sarcoma (FUS), and hnRNPA2 as well as the C-terminal heptad repeat domain of RNA Polymerase II (RNA Pol II) [50, 51, 53, 66]. I used 44-residue fragments of these proteins to reduce the computational cost associated with simulations of full-length proteins using a multi-replica approach necessary to obtain converged equilibrium properties. I used the lowest temperature (300 K) replica for all the analyses presented in this paper.

First, I focused on the results of two protein fragments, TDP-43_{310–350} and FUS LC_{0–43}, as an example of IDPs where NMR data suggest partial helical structure and complete disorder, respectively. I calculated the chemical shifts from the simulated ensembles using the SPARTA+ algorithm [63] and subtract the sequence-based random-coil values from the Poulsen Web server [67] to compute secondary chemical shifts, which can be used to infer secondary structure populations. Figure 2.1(a) shows an excellent agreement between the simulation and experimental data in the helix-forming region $\Delta\delta C_\alpha - \Delta\delta C_\beta$ values, a well-established metric of helical (positive values) or sheet (negative values) computed by calculating the difference between the C_α and C_β observed chemical shifts and those expected for a completely random coil disordered protein) at residues 321-330 in TDP-43_{310–350}, an alanine-rich region embedded in the LC domain [15]. However, in the case of FUS LC_{0–43}, simulation data suggest the formation of helical structures in the N-terminal region,

residues 6-13, which is not supported by the experimental data. [25, 68] I also computed the secondary structure propensities from the simulation data directly using the DSSP algorithm [64] and compare these to the predicted values based on the experimental NMR chemical shift data using the $\delta 2d$ Web server [65] (Figure 2.1(b)). Again, there is excellent agreement between simulation and experiment regarding the presence of partial helicity in the alanine-rich segment of TDP-43 but apparent overpopulation of the helix in several regions of the polar-rich FUS LC. Overall, both the force fields display similar behavior with some minor differences.

Robustelli et al. had shown a similar behavior for ff03ws on other IDPs with various partially populated secondary structures and proposed a new force field to provide a better representation of folded and disordered proteins. I also conducted additional simulations using their ff99SB-disp force field [36]. These data are shown in Figure A.1. Consistent with previous findings that partial helical structures may be underpopulated in ff99SB-disp, simulations with ff99SB-disp show underpopulation of helix in TDP-43₃₁₀₋₃₅₀ while lower overall helicity for the disordered FUS LC₀₋₄₃. However, it is interesting that the simulated ensemble using ff99SB-disp also populates α -helical structures, like ff03ws and ff99SBws, in the N-terminal part of FUS LC₀₋₄₃ which are not present in the experimental data.

The analysis above clearly suggests that further fine-tuning of the current state-of-the-art force fields is needed to make them suitable for low-complexity sequences such as FUS LC₀₋₄₃. The observed overpopulation of helical structures in the case of FUS LC₀₋₄₃ is most likely related to the presence of specific amino acids for which the balance in helical and extended structures is not optimal. In previous work, Lifson-Roig helix growth parameter (w) was used to assess residue-level helical propensities with a model Ala-based host-guest peptide that helped us identify residues with significant deviations from the experiment and propose refinements [69]. Moving forward, it will be useful to develop a strategy that can take advantage of the available NMR data on the sequences of interest and does not require us to interpret the secondary structure propensities in terms of a helix-coil transition model in the context of a host-guest peptide.

Here, I calculated the differences between the secondary chemical shifts between the simulation and experimental data, i.e., $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta) = (\Delta\delta C_\alpha - \Delta\delta C_\beta)_{\text{sim}} - (\Delta\delta C_\alpha - \Delta\delta C_\beta)_{\text{expt}}$, for each residue in a variety of IDPs for which NMR data are available. I selected many peptide fragments from a set of prion-like IDPs such as FUS LC, TDP-43 CTD, hnRNPA2 LC, and RNA Pol II C-terminal domain for the proposed comparison (Table A.1). These LC protein sequences are mainly

composed of polar and aromatic amino acids with a low occurrence of nonpolar aliphatic amino acids. The $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta)$ data are shown in Figure 2.2 for each residue type and each simulated peptide (Figure 2.2(a) and (b)) and the average over all the peptides with the two force fields (Figure 2.2(c)). The residues Ser and Thr in ff99SBws and ff03ws, and Gln in ff99SBws, which constitute a significant fraction of the amino acid composition in these sequences, deviate significantly from the experimental values. The positive $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta)$ values reflect the overpopulation of helical structures in these residues. Numerically, the average deviations for threonine are the most significant ($\sim +1.0$ ppm) followed by serine ($\sim +0.5$ ppm) and glutamine ($\sim +0.3$ ppm) in ff99SBws. Also, the observed deviations in different sequences (Figure 2.2(a) and (b)) may reveal sequence dependence to some degree, but the overall trend is qualitatively consistent.

On the basis of the analysis above, we decided to test the effect of changes in the backbone torsion potential parameters of three amino acids (Ser, Thr, and Gln, referred hereafter as the STQ group) in ff99SBws. Previously, Best and Hummer proposed the following form for the backbone dihedral angle ψ correction: $V_1(\psi; k_\psi, \delta_\psi) = k_\psi[1 + \cos(\psi - \delta_\psi)]$, where k_ψ and δ_ψ are the magnitude and phase offset of V_1 , respectively. For ff99SBws, $k_\psi = 2.0$ kJ/mol and $\delta_\psi = 105.4^\circ$ are used to correct the intrinsic bias toward β -sheet structures for the ff99SB force field [29]. After testing for various values, a reduction in k_ψ from 2.0 to 1.0 kJ/mol is sufficient and appropriate to shift the bias away from helical configurations. To make a broader assessment on how changes in the potential parameters for the STQ residues affect the observed secondary structure propensities, I tested several combinations of the modified residues (S, T, ST, TQ, and STQ; Figure 2.3) for three peptides from the data set: FUS LC_{0–43}, TDP-43_{310–350}, and RNA Pol II_{1927–1970}. I conducted PT-WTE simulations for these three peptides with variants of the ff99SBws force field, where the suffix letter(s) represent the amino acids for which the k_ψ is modified. For example, in the case of ff99SBws-ST, the torsion potential is changed for Ser and Thr residues.

The $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta)$ comparison between the ff99SBws and its new variants are shown in the Figure A.2 for ST, TQ, and STQ. A change in only one of the residue types (S or T) k_ψ is not sufficient to observe any meaningful differences from the ff99SBws force field (Table 2.1). For force field variants with two or more residues k_ψ modified, the deviations consistent with overly helical simulations are reduced for FUS LC_{0–43} but are minimally affected for TDP-43_{310–350} and RNA Pol II_{1927–1970}. For reference, an additional comparison between ff99SBws, ff99SBws-STQ, and ff99SB-disp force fields is presented in Figure A.3, showing that ff99SB-disp does not fully capture

the balance of partial helical structure in these sequences. To quantify the overall differences on the whole sequence, I computed the root-mean-squared deviation (RMSD) between the simulation and experimental data and report it in Table 1. The RMSD values highlight the benefits of applying the proposed residue-specific changes for all the peptides, but this is especially true for FUS LC. These changes do not negatively impact the observed $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta)$ behavior of other neighboring residues.

To visualize the local structural changes, I also calculated the secondary structure propensities from the simulated ensembles using the DSSP algorithm for these different force field variants and compare it with experimental values derived from chemical shifts using the $\delta 2D$ method (Figure 2.3). For FUS LC₀₋₄₃, the modified force fields (ST, TQ, and STQ) significantly reduce the helicity in the N-terminal region as compared to the ff99SBws. The helical fraction at all residue positions is below 20%, which might be at or close to the detection limit for interpretation of the NMR data using methods based on chemical shifts such as $\delta 2D$ given the possibility of many different low population states with different structures making up the conformational ensemble. For TDP-43₃₁₀₋₃₅₀, there is no significant change in the helical population in the previously identified region (residues 321-330) among different force fields; a considerable decrease in helicity (from 40% to 20%) is observed between residues 340-345. The latter (STQ) helps improve the agreement with the experiment, which is different than what is seen in the case of ff99SB-disp which shows reduced helicity everywhere for this peptide (Figure A.3). RNA Pol II₁₉₂₇₋₁₉₇₀ shows relatively little changes in the observed behavior presumably since it is almost entirely disordered, which was already captured well in the ff99SBws force field.

To check whether inter-residue contacts are affected by the changes in the force field parameter, I computed contact maps for FUS LC₀₋₄₃ and TDP-43₃₁₀₋₃₅₀, as shown in Figures A.4 and A.5. As expected, a significant reduction in helical contacts ($i, i + 4$) is observed, but otherwise, the contacts are relatively similar between the structural ensemble generated based on ff99SBws and ff99SBws-STQ. The changes in the peptide dimensions due to the STQ dihedral corrections are relatively small (Figure A.6) and therefore not expected to lead to significant changes in the formation of non-local intramolecular contacts.

To assess the convergence of the simulation data in terms of secondary structure propensity, I also plotted residue-level secondary structure as a function of simulation time for each replica moving

through the temperature space (Figures A.7 and A.8). These data highlight many transitions between different states and the transient nature of populated structures, which ultimately contributes to the convergence of the estimated propensities reported in the paper.

I also checked the performance of these new parameters with an additional test case, which is commonly employed to test the suitability of dihedral corrections, i.e., (AAQAA)₃ peptide. As shown in Figure A.9, there is a small reduction in the helical propensity of this peptide, but the overall results are still consistent with the NMR data.

The changes proposed here should improve the accuracy of the force field in general, but more work is needed to see whether certain sequences containing STQ residues do not show improvement in agreement with experimental observables compared to the ff99SBws model. Nonetheless, many recent studies have proposed force fields with residue-specific torsion and non-bonded parameters to improve their accuracy with a particular focus on IDPs. It is possible to incorporate some of these changes in the proposed force field further to improve the resulting IDP properties of arbitrary sequences, but significant work is needed to test the combination of parameters that will yield an optimal behavior.

2.4 Conclusion

In conclusion, the proposed residue-type specific modifications on backbone torsion potentials of ff99SBws improved the accuracy of all-atom simulation of low-complexity IDPs (FUS LC_{0–43}, TDP-43_{310–350}, and RNA Pol II_{1927–1970}) in terms of their agreement with the ¹³C NMR chemical shifts and hence secondary structure propensities. In this study, I adjusted the torsion potential correction parameter for three polar amino acids (S, T, and Q) to illustrate a promising path for all-atom force field optimization for IDPs using NMR spectroscopy data. Most importantly, the new force field (ff99SBws-STQ) is still transferable and does not require any experimental input while simulating a specific protein system of interest. We believe that ff99SBws-STQ will provide a more accurate description of the structural and dynamical properties of IDPs involved in the formation of biomolecular condensates via LLPS and become an indispensable model for future simulations of the many disordered proteins rich in polar residues.

2.5 Tables and Figures

Table 2.1: Root-mean-squared deviation (RMSD) of simulated $\Delta\delta C_\alpha$, $\Delta\delta C_\beta$, $\Delta\delta C_\alpha - \Delta\delta C_\beta$ (in ppm) from NMR ^{13}C chemical shifts of three IDPs for ff99SBws, modified ff99SBws variants (S, T, ST, TQ, STQ) (this study), and ff99SB-disp.

$\Delta\delta C_\alpha - \Delta\delta C_\beta$	FUS _{0–43}	TDP-43 _{310–350}	RNA Pol II _{1927–1970}
ff99SBws	0.90 ± 0.11	0.64 ± 0.09	0.68 ± 0.11
ff99SB-disp	0.90 ± 0.11	0.78 ± 0.07	0.74 ± 0.11
ff99SBws-S	0.86 ± 0.10		
ff99SBws-T	0.98 ± 0.13		
ff99SBws-ST	0.64 ± 0.07	0.71 ± 0.08	0.74 ± 0.08
ff99SBws-TQ	0.62 ± 0.08	0.66 ± 0.06	0.71 ± 0.10
ff99SBws-STQ	0.54 ± 0.07	0.51 ± 0.04	0.74 ± 0.08

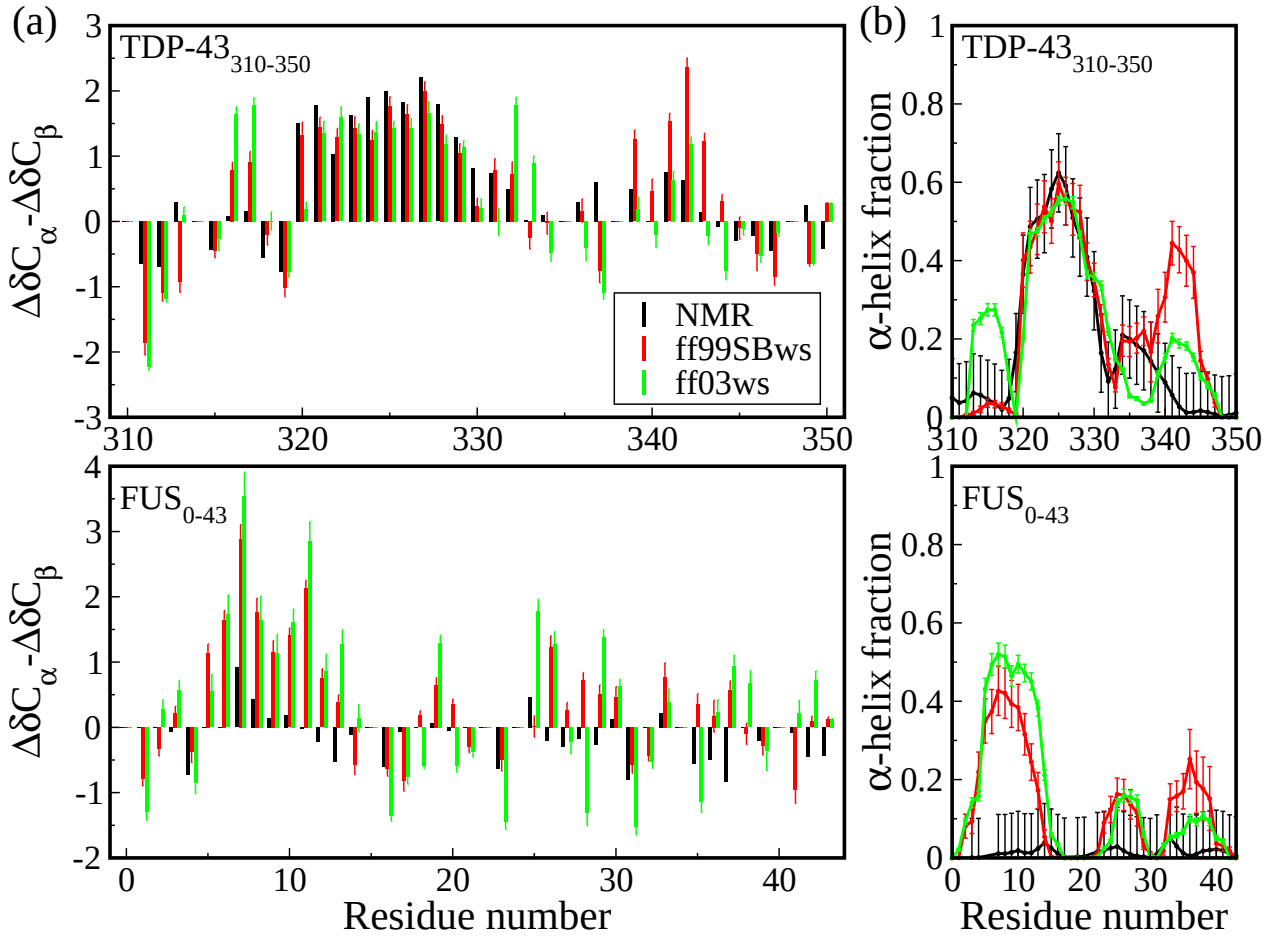


Figure 2.1: Chemical shifts and α -helix propensities of TDP-43₃₁₀₋₃₅₀ and FUS₀₋₄₃ in simulations and experiments. **(a)** Chemical shifts $\Delta\delta C_\alpha - \Delta\delta C_\beta$ and **(b)** α -helix propensities of TDP-43₃₁₀₋₃₅₀ and FUS₀₋₄₃ by using all-atom force fields ff99SBws and ff03ws compared to experimental results. Chemical shifts are calculated from protein structure by using the SPARTA+ algorithm [63]. Secondary structures are assigned by DSSP. Experimental values of secondary structure propensities are calculated from NMR chemical shifts by using the δ 2D program [65].

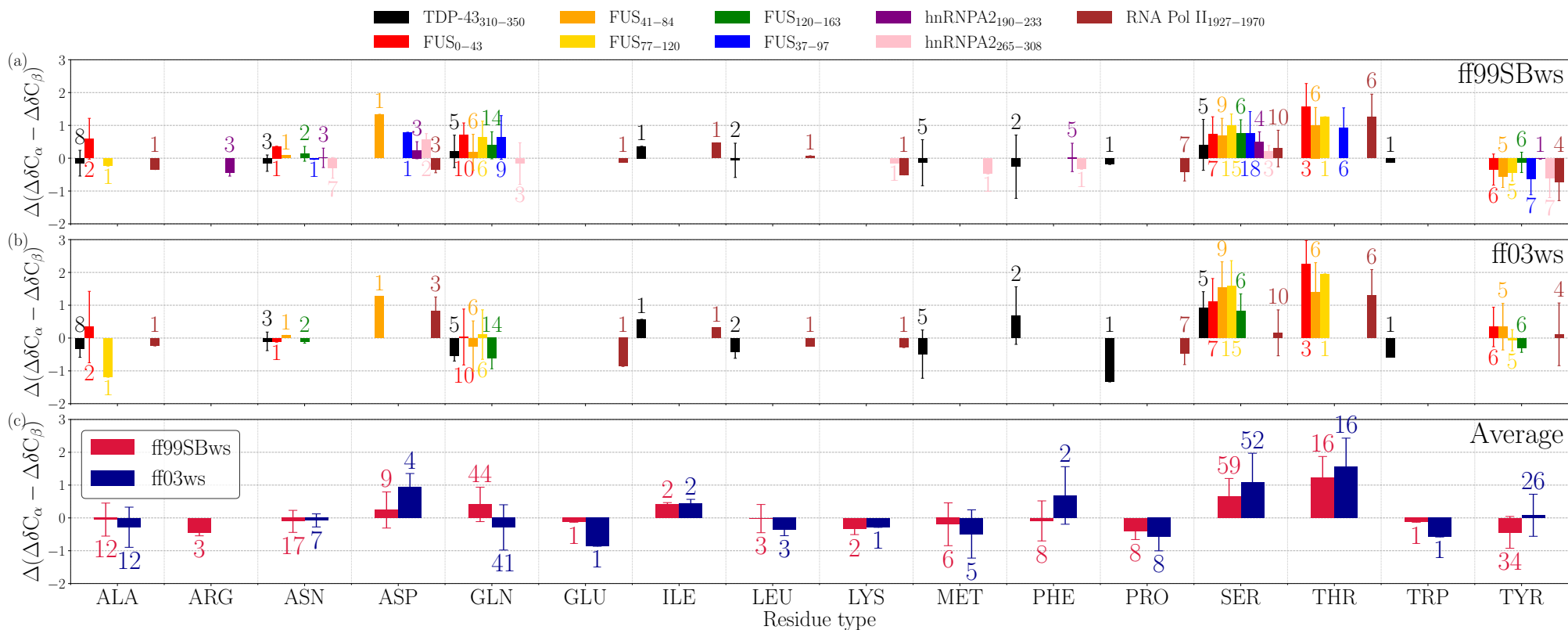


Figure 2.2: Deviation of chemical shifts $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta)$ (in ppm) in simulations from NMR ^{13}C chemical shifts. (a) and (b) show $\Delta(\Delta\delta C_\alpha - \Delta\delta C_\beta)$ of each peptide from ff99SBws and ff03ws simulations, respectively. (c) shows the average of all peptides in the same force field. Error bars are the standard deviations among the sampled residues. Annotated numbers next to the bar are the number of the residues of that type present in the chosen peptide(s).

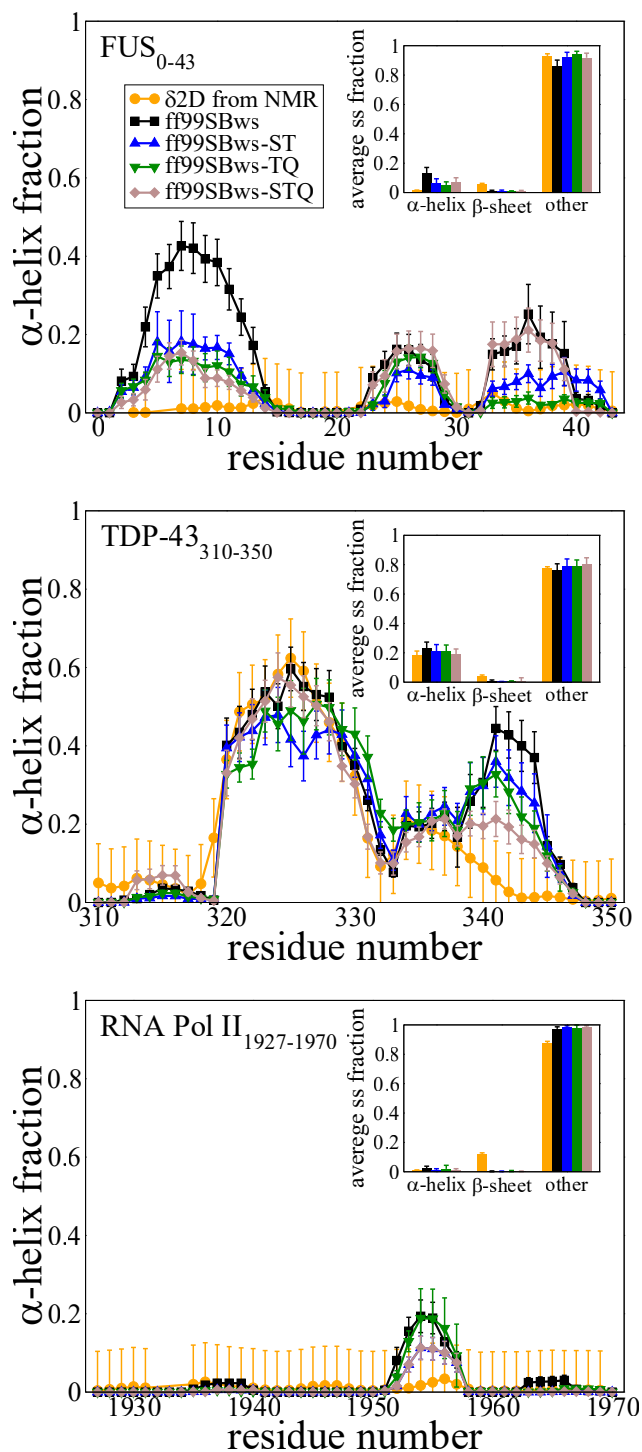


Figure 2.3: Secondary structure propensities of FUS₀₋₄₃, TDP- 43₃₁₀₋₃₅₀, and RNA Pol II₁₉₂₇₋₁₉₇₀ in simulations and experiments. Error bars for DSSP values are deviations between 10 ns segments in the simulation trajectories. Error bars for δ 2d population per residue are 0.1 (10%).

Chapter 3

Quantifying Molecular interactions contributing to liquid-liquid phase separation of intrinsically disordered protein systems

This chapter is adapted from this paper [2] on providing a biophysical molecular model for Liquid-Liquid Phase Separation (LLPS) of the RNA binding proteins FUsed in Sarcoma (FUS) and RNA Polymerase II (RNA Pol II) with nuclear magnetic resonance (NMR) spectroscopy and molecular dynamics (MD) simulations. Mainly, I would like to focus on my contributions on the computational aspect of the study i.e. performing the MD simulations and analysis of the MD trajectories.

I would like to acknowledge all the co-authors, Anastasia C. Murthy, Nina Jovic, Abigail M. Janke, Da Hee Seo, Theodora Myrto Perdikari, Jeetain Mittal, and Nicolas L. Fawzi, for their contributions in this paper [2] which is presented in this chapter. Anastasia Murthy, the first author of this paper, contributed the majority of the experimental work and the writing for the published manuscript.

3.1 Introduction

The nucleus contains several membrane-less organelles, including the nucleolus, heterochromatin, and transcriptional condensates, that form via phase separation of their protein and nucleic acid constituents [70–76]. Found in nuclear and cytoplasmic condensates, Fused in Sarcoma (FUS) is an RNA-binding protein with roles in RNA metabolism and transcription. FUS is among the FET (FUS, EWS, TAF15) family of proteins that have overlapping functions in RNA processing [77–79], and has a multi-domain architecture that facilitates self-association and RNA-binding [80–82]. FUS contains a disordered N-terminal domain of low complexity sequence enriched in serine, tyrosine, glycine, and glutamine residues (SYGQ LC), two folded domains, an RNA recognition motif (RRM) and a zinc-finger, and three arginine-glycine-glycine rich (RGG) domains.

The SYGQ LC domain mediates FUS self-assembly into liquid-like granules, hydrogels, and solid aggregates [68, 83–86]. A previous study by Murthy et al. recently used solution-state NMR combined with molecular simulation and found contacts stabilizing liquid-liquid phase separation (LLPS) of SYGQ LC involves interactions between all enriched residue types via diverse interaction modes such as hydrophobic, hydrogen bonding, and π/π [25]. Furthermore, aromatic residue repeat number and spacing (i.e., the lack of adjacent aromatic residues) are essential determinants of phase separation and aggregation propensity. Solution NMR also showed contacts between aromatic-containing short motifs have been demonstrated in the dispersed phase (before phase separation) of a related protein [87]. Aromatic repeats in FUS LC also form contacts with nuclear import proteins like karyopherin $\beta 2$, which may help prevent FUS aggregation [88]. Containing distinct SYGQ LC and RGG domains, the contribution of LC/RGG interactions and specifically the contribution of tyrosine and arginine contacts to FUS LLPS and aggregation have been established [81, 85, 89–93]. However, the molecular details have not been investigated in detail by direct observation of the contacts in condensed phases. NMR has proved a useful tool to evaluate contacts in the Phe/Arg-rich intrinsically disordered N-terminal region of the related germ-granule protein Ddx4 [94]. Importantly, nuclear Overhauser experiments (NOE) show that contacts between many other residue pairs (not limited to Phe/Arg) are present within Ddx4 disordered domain condensed phases. Furthermore, the shifts in the phase diagrams of Ddx4 and LAF-1 RGG and an engineered variant of hnRNPA1 LC caused by charge residue scrambling suggest that interactions beyond aromatic-arginine contacts contribute to phase separation [26, 95]. Hence, here we seek to probe the specific contacts between

tyrosine, arginine, and the other residues found in the SYGQ LC and RGG domains in the phase separation of FUS.

FET proteins are also involved in the regulation of transcription. FUS is enriched at transcriptional start sites of thousands of genes and directly interacts with and recruits RNA Pol II [81, 96, 97]. In rare leukemias and sarcomas, fusions of the SYGQ LC domain (and sometimes with addition of RGG1) of FUS to DNA-binding domains of certain transcription factors create potent transcriptional activators [98, 99]. Assembly of FUS into transcriptional condensates may be required for the clustering of RNA Pol II [73, 81, 100]. Further, the disordered C-terminal tail of RNA Pol II (RNA Pol II CTD) is rich in aromatic/polar residues and readily incorporates into phase-separated FUS in cells and biochemically [68, 97, 101]. Human RNA Pol II CTD contains 52 heptapeptide repeats with the consensus sequence YSPTSPS and itself undergoes LLPS (with the addition of crowding agents) [102]. Given the importance of LLPS to transcription and potentially to oncogenicity, an understanding of the interactions that contribute to FUS and RNA Pol II co-partitioning could provide a molecular picture of transcriptional activation and hence may facilitate the development of therapies for FET-associated cancers.

Using direct observation of multi-component condensed phases by NMR spectroscopy, MD simulations, and biophysical assays, we tested if a limited or broad set of domains (RGG1, RGG2, or RGG3), contacts (e.g., R and Y), and interaction modes (e.g., cation- π) contribute to full-length FUS LLPS and its co-partitioning with RNA Pol II CTD. We characterized the intermolecular interactions of two condensed phase models: a two-component phase comprised of the SYGQ LC and RGG3 domains of FUS, and a three-component phase of RNA Pol II CTD, FUS SYGQ LC, and FUS RGG3. We tested if co-phase separation of RNA Pol II CTD with FUS requires either the SYGQ LC or the RGG domains and the contribution of particular degenerate residues in RNA Pol II CTD heptads (e.g., lysine) to these interactions. This work highlights the importance of weak, multivalent interactions involving diverse interaction modes that are likely common to many RNA-binding proteins.

3.2 Method for Molecular Dynamic Simulation and Analysis

3.2.1 Metadynamics All-Atom Molecular Dynamics Simulations

For all-atom simulations, we chose fragments from each domain that were forty-four residues long to allow for sufficient conformational sampling within reasonable timescales (Fig. 3.1). We selected a fragment best representing the amino acid distribution of SYGQ LC_{11–54} within the full SYGQ LC domain in addition to the portion that we previously used (SYGQ LC_{120–163}) [25] which lacks some of the less abundant residue types (Fig. 3.1A). Similarly, the fragments for RGG1, RGG2 and RGG3 were chosen such that they were the most representative of each RGG sequence (Fig. 3.1B). We note that in experiments, the length of the sequences for the SYGQ LC and RGG domains are non-equivalent; therefore, the simulations allow us to compare the interactions between similar length fragments. All-atom simulations were conducted on systems containing FUS SYGQ LC_{11–54} or SYGQ LC_{120–163} with RGG1_{220–267}, RGG2_{372–419}, RGG3_{454–501} or RNA Pol II CTD_{1853–1896} and RGG3_{454–501} with RNA Pol II CTD_{1853–1896}. Simulations were conducted using GROMACS 4.6.7 software package [56] with PLUMED 2.4 plugin [57]. We used the Amber99SBws-STQ force field with improved residue-specific dihedral corrections [1], TIP4P/2005 water [46], and improved salt parameters from Luo and Roux [46, 103]. Simulations were conducted using parallel tempering in the well-tempered ensemble (PT-WTE) with 16 temperature replicas from 300 K to 500 K, each with a metadynamic bias on the potential energy to enhance fluctuations [58, 59, 104]. Single chain simulations were conducted using this PT-WTE scheme, while two chain simulations were conducted using this scheme coupled with a well-tempered metadynamics (WT-MetaD) bias on the interchain Van der Waals (VdW) contacts between heavy nonpolar atoms (i.e. $|q| < 0.25$) as we have previously done to improve sampling binding/unbinding events [105].

Initial configurations for two-chain simulations of each chain in each replica were randomly generated by CHARMM 46 tool [106]. Two chains were initially placed 10 Å apart (between any two atomic positions) and solvated in explicit water and ~100 mM NaCl with box size = 8.0 nm. Each replica was equilibrated for 100 ps in NVT and then NPT simulations with parallel tempering for thermal equilibrium. Production simulations were performed for 200 ns, with the first 50 ns discarded as equilibrations, and the last 150 ns used for analysis.

3.2.2 Analysis of Molecular Dynamics Trajectory

All-atom simulations were analyzed at 300 K as in the experiments. For two-chain simulations, we calculated the free energy surface of the two-chain systems from the metadynamics bias using the built-in function (`sum_hills`) in PLUMED v2.7.2, and an alternative time-independent method from Tiwary and Parrinello [107], then subtracted the difference between the two results to generate uncertainties for free energy values. Contact propensities in all-atom two-chain PT-WTE simulations were reweighted based on free energy surface. VdW contacts were considered as any two heavy atoms being within 6 Å of each other. Hydrogen bonds were considered as a donor atom and an acceptor atom being within 3 Å and the donor-hydrogen-acceptor angle being larger than 120°. sp^2/π interactions were calculated as presented by Vernon et al. [108] and considered as any two sp^2 -hybridized groups having at least two pairs of atoms being within 4.9 Å and the angle between the normal axes of the two sp^2 -planes being less than 60°. Cation- π interactions were considered as a cationic atom being within 7 Å of the center of an aromatic ring and less than 60° from the normal axis of the π face. Salt bridges are considered as a cationic atom and an anionic atom being within 6 Å of each other. Uncertainties in contact propensities were calculated using block error with $N = 2$ blocks (75 ns each) with reweighing, calculated using the following formula, where the variance (squared of error) is equal to $\sigma^2 = \Sigma_{i=1}^N [w_i(x_i - \bar{x})^2] / (N^2 \Sigma_{i=1}^N w_i)$ where w_i is the weights of block i , x_i is the contact propensity calculated within block i , $\bar{x}$ is the weighted averaged contact propensity.

Calculation of NMR spin relaxation parameters from simulated trajectories was performed as previously [15, 109]. Briefly, constant temperature MD simulations for each monomer using the same run parameters as above (without enhanced sampling), also in NVT ensemble ($T = 300$ K) were performed, with 12 independent MD simulations for 12 randomly selected initial configurations from the equilibrium ensemble (see above). Each independent simulation is run for 200 ns, yielding a total of 2.4 μ s simulation time. Results from MD trajectories are presented as mean averaged over 12 independent simulations, and uncertainties are the standard error of the mean calculated from the deviation between 12 simulations.

3.3 Results

3.3.1 Contacts between various residue types in LC and RGG domains contribute to FUS LLPS

Several studies suggested that both the SYGQ-rich and RGG domains of FUS contribute to LLPS [90–93]. We directly tested whether the RGG domains can enhance phase separation of the SYGQ LC, looking for the occurrence of LLPS at concentrations below the LLPS saturation concentration of each component. Experimental results from turbidity assay suggest that both SYGQ LC domain and RGG domains contribute to LLPS of full-length FUS, and all three RGG domains (RGG1_{164–267}, RGG2_{372–422}, and RGG3_{453–507}) in FUS are critical for phase separation at physiological concentrations and contribute to LLPS similarly [2]. Furthermore, NMR experiments show that RGG3 and LC remain predominantly disordered and dynamic in the liquid condensed phase [2].

A previous study by Murthy et al. showed that within a condensed phase containing only the SYGQ LC, many residues types (not only tyrosine but, for example, glutamine) contribute to interactions using hydrophobic, hydrogen bonding, and sp^2/π interaction modes stabilizing LLPS [25]. In ^1H - ^1H Nuclear Overhauser Effect (NOE) NMR experiments, observation of NOEs between all major residue types within the SYGQ LC and RGG3 strongly suggests that molecular interactions between the SYGQ LC and RGG domains are distributed across many residue types within the condensed phase and are not concentrated only on certain amino acids (e.g., only arginine and tyrosine) [2]. More elaborated NOESY-HSQC (Nuclear Overhauser Effect Spectroscopy Heteronuclear Single Quantum Coherence) experiments suggest that glycine, proline, methionine, and arginine in RGG3 and threonine, glutamine, tyrosine, glycine, and serines residues in SYGQ LC [2].

In simulations with SYGQ LC_{11–54} or SYGQ LC_{120–163} and RGG1, RGG2, or RGG3, we analyzed the free energy surface as a function of total contacts between the SYGQ LC and RGG fragments (Fig. 3.1E). All three RGG domains show free energy minima above zero contacts with some variability in the location of the minimum, suggesting that all three RGG domains can interact with SYGQ LC_{11–54} or SYGQ LC_{120–163}, albeit to different extents. We note that the amino acid composition between the three RGG domains is slightly different, with RGG1 containing the least amount of arginine residues (Fig. 3.1B). Consistent with the compositional differences within the

RGG domain simulation fragments, we observe differences in the single-chain behavior such as the radius of gyration of the RGG fragments. These differences in single chain properties could account for the difference in contact propensities for certain fragment pairs such as SYGQ LC_{11–54} with RGG2 and SYGQ LC_{120–163} with RGG3 (Fig. 3.1E,F).

To gain insight into the molecular details of the modes of interaction that are important for FUS SYGQ LC and RGG domains, we characterized the interactions between FUS SYGQ LC and RGG domain 44-residue fragments using two chain all-atom simulations, as in Murthy et al. previous work [25]. For all three RGG domains, arginine and glycine have the highest contact propensities with the SYGQ LC, likely because these two residue types are the most abundant in the sequence (Fig. 3.2A-C). Importantly, in the SYGQ LC, several residue types, not just tyrosine, but also in particular glutamine, serine, glycine, and threonine mediate contacts with RGG3 residues, including arginine. To further understand the contribution of specific residues types, we normalized the contact propensities by the amino acid abundance within each sequence. After normalizing, arginine residues remain a “hotspot” for interactions within the RGG domains along with aromatic residues types; however, the contacts involve all residue types within the SYGQ LC and RGG domains (Fig. 3.2A-C). Similar contacts are observed for the 120-163 fragment of the SYGQ LC (Fig. 3.2D). Additionally, the contacts are distributed across the SYGQ LC and RGG domains (Fig. 3.2), as much as we can ascertain within the challenge of sampling a large inter-residue contacts configuration space. These data are consistent with experimental NOEs present between most residue types, suggesting that all residues types can participate in the interaction between the SYGQ LC and RGG domains in the condensed phase.

To characterize the interaction modes between residue pairs in the simulated SYGQ LC and RGG domains beyond non-specific VdW interactions between all pairs, we classified the total interactions, i.e., VdW contacts, into hydrogen bonding, sp^2/π , and/or cation- π interactions. All three interactions modes are represented in contacts between the SYGQ LC and all three RGG domains (Fig. 3.2A). We note that the sp^2/π interactions calculated based on Vernon et al. [108] may only account for planar stacking of the sp^2/π groups; therefore, we also used a broader definition without orientational restrictions to include other possible sp^2/π geometries (e.g. perpendicular stacking), where the overall trend for the interactions remains the same (Fig. 3.3A,B). To test if results depend on choice of simulated fragment, we compared the interactions of SYGQ LC_{120–163} with each FUS RGG domain and found a similar distribution of interactions modes (Fig. 3.2A,B).

We note that RGG2 has highest number of interactions with the compositionally representative fragment SYGQ LC_{11–54}, while RGG3 has the most contacts with the previously studied [25] fragment SYGQ LC_{120–163} (Fig. 3.2A), suggesting that certain regions of FUS may interact more favorably. However, if we normalize the hydrogen bonds, sp^2/π and cation- π interactions by the total number of contacts within each dataset, we find that the relative contribution of each interaction mode is the same regardless of the SYGQ LC fragment used (Fig. 3.3B). Therefore, these data suggest that all three RGG domains interact with the SYGQ LC via similar interaction modes. Finally, we characterized the contribution of the interaction modes between residue pairs within the SYGQ LC and RGG3 domain. We find that, for residue pairs in VdW contact, the most frequent of these interaction modes involves hydrogen bonding interactions (Fig. 3.3C,D and Fig. 3.2E,F). Interestingly, glycine and serine pairings show significant hydrogen bonding in addition to the previously suggested backbone sp^2/π interactions [73, 108]. Importantly, the interactions between tyrosines in the SYGQ LC and arginines within the RGG3 domain involve all three interaction modes. We note that the agreement of current all-atom models to study IDPs with the available experimental data from NMR, SAXS, FRET provides confidence in their ability to capture structural and dynamical properties of these proteins reasonably well [52, 110, 111]. On the other hand, these models have not been tested for their ability to provide insight on the precise balance between different interaction modes (especially π - π , cation- π). Further, the definitions of these contacts we adopted [108] may require additional refinement based on input from quantum calculations. Therefore, the percentage of each interaction type should be treated with caution and should not be interpreted as providing insight on the relative energetic contribution of each mode. Taken together, while cation- π interactions between arginine and tyrosine have been proposed as the major contributor SYGQ LC and RGG-rich sequences, glycine and other residue types within the RGG-rich sequences also contribute through hydrogen bonds and sp^2/π interactions as well as non-specific VdW contacts.

3.3.2 FUS contacts with the C-terminal tail of RNA polymerase II

Previous studies demonstrated that the degenerate repeat half of the C-terminal tail of RNA Pol II (CTD) interacts with and partitions into FET protein liquid-like phases and interacts with fibrillar FET LC hydrogels [66, 68, 97]. Arginine residues within TAF15 are important for mediating co-partitioning of TAF15 and RNA polymerase [100], yet the molecular details of these interactions are poorly understood. We hypothesized that the degenerate repeats 27-52 of RNA Pol II interact

not only with FUS SYGQ LC but also with the arginine-rich domains within FUS. Experimental results in NMR experiments and turbidity assays shows that RNA Pol II CTD interacts and phase separates with FUS SYGQ LC and RGG3 [2].

I conducted all-atom simulations with RNA Pol II CTD_{1853–1896} and FUS SYGQ LC_{11–54} and RGG3 fragments and analyzed the free energy surface as a function of total contacts. Both the pairs of protein fragments have free energy minima at non-zero contacts, suggesting that both the SYGQ and RGG3 interact with RNA Pol II CTD in simulations consistent with the experimental data (Fig. 3.5B). Taken together, these data suggest that the C-terminal heptads of RNA Pol II can interact with the RGG domains within FUS in addition to the SYGQ LC to promote co-phase separation.

3.3.3 Lysines in RNA polymerase II CTD form LLPS contacts

NOE NMR experiments demonstrated interactions between all residue types enriched in RNA Pol II CTD (i.e., serine, tyrosine, proline, and lysine) to threonine and glutamines within SYGQ LC as well as arginines in RGG3 [2].

To probe the contribution of particular residue types to the interactions between RNA Pol II CTD and FUS SYGQ LC or RGG3, we analyzed the contact propensities between residue pairs within two-chain all-atom simulations. The interactions between RNA Pol II CTD_{1853–1896} and FUS SYGQ LC_{11–54} involve all enriched residue types (Fig. 3.4A, left); however, the normalized contact propensities indicate that lysines and tyrosines within RNA Pol II CTD_{1853–1896} are hotspots for interactions with FUS SYGQ LC_{11–54} (Fig. 3.4A, right). In contrast, the interactions between RNA Pol II CTD_{1853–1896} and FUS RGG3 are dominated by arginine and aromatic residue contacts (Fig. 3.4B, right). Both pairs exhibit contacts that are relatively distributed across the protein fragments used except for some enhancement over two to three adjacent residues (Fig. 3.4B,C). Given the extreme length of simulations needed for convergence of simulated ensembles to reliably quantify interactions at the level of each residue [109], we focused instead on interactions between each residue type, which represents information averaged over many occurrences of these residues within the sequence.

Finally, we characterized the contribution in the simulations of the interaction modes between residue pairs in FUS SYGQ LC or RGG3 with RNA Pol II CTD. In both datasets we find that the contacts between residue pairs show significant presence of hydrogen bonding interactions (3.6A)

though contacts between tyrosine residues in FUS SYGQ LC and lysine residues in RNA Pol II CTD show significant cation- π interactions. Conversely, in both datasets, glycine and lysine pairs interact by hydrogen bonding interactions rather than sp^2/π . Although we noted the caveats of molecular simulation approaches above, our analysis of the interaction modes formed between residue pairs in multiple systems suggests that hydrogen bond interactions may be more important for LLPS, in general, than currently appreciated in the literature. Both experimental data [2] and simulation data suggest that lysine residues play a role in determining the interactions and phase partitioning of RNA Pol II CTD with FUS.

3.4 Discussion

Cation- π interactions between arginine and tyrosine (and phenylalanine) are often assumed to be the major driving force of FUS and other disordered domain LLPS based on studies mutating all instances of several residue types [90, 92, 93, 112, 113]. However, mutagenic studies replacing single residue types cannot directly probe which residue pairs form interactions, much less tease apart synergistic contacts. Here, we demonstrate that other residue types and interaction modes also make contributions to FUS LLPS (Fig. 3.2, 3.3, and 3.4). In all-atom simulations, we analyzed the contribution of various contact modes, including hydrogen bonding and π -interactions, on FUS SYGQ LC and RGG LLPS. Several studies have suggested that the number or positioning of key residues may make larger contributions to LLPS than the bulk of the sequence, suggesting a pecking order among interactions that induce LLPS [26, 87, 114]. While we did not evaluate the interactions on an individual residue position level, our data are consistent with the view that certain residue types, including tyrosine and arginine, form primary interactions essential for robust physiological LLPS, stemming at least in part from the additive effect of multiple potential interaction modes; however, these residues do not interact only with each other. In fact, many residue types, such as serine and glycine, that are limited in potential interaction modes but enriched in the sequence, form interactions that also contribute to LLPS. In this way, it is difficult to precisely assign a relative weight for individual residues and interaction modes. These results are consistent with previous NMR work demonstrating tyrosine residues in FUS LC and phenylalanine and arginine residues in Ddx4 make contacts with many residue types within the condensed phases [25, 94]. Addressing the relative contribution of SYGQ LC-LC, RGG-RGG and SYGQ LC-RGG interactions

to phase separation within full-length FUS, as well as the effect of RNA on these interactions, will be important next steps to understanding LLPS of FUS.

Previous computational work on protein LLPS mostly used coarse-grained models, which provide important information on the sequence-dependent LLPS behavior of proteins [115, 116], but cannot provide guidance on the underlying interactions. Recent studies have used atomistic simulation data to obtain insights into the prevalence of different interactions modes within the folded protein structures [108] or interactions between short disordered peptides [117, 118]. Here, we used the Amber99SBws-STQ force field [1] (Chapter 2), with water model used in its design, TIP4P/2005 [46], due to its ability to capture NMR-validated conformational properties of disordered proteins with or without transiently populated secondary structure elements combined with a correction for helical bias observed in three residue-types (Ser, Thr, and Gln) present in low-complexity sequences including FUS LC [1]. Recent papers discuss how these models provide molecular details of IDP structure and dynamics on par or better than other existing all-atom force fields [119, 120]. Further work is needed to establish if these force fields also capture the relative importance of different interactions modes, as significant questions remain [121].

Biomolecular condensates are complex mixtures of components. RNA Pol II CTD is recruited to SYGQ LC liquid condensates and hydrogels [68, 97]. The degenerate half of RNA Pol II CTD is enriched in lysines. Lysine-rich sequences are enriched in liquid condensates called P bodies and are important for tau LLPS [122–124]. Previous studies reported that the lysine residues partially mediate incorporation of RNA Pol II CTD into TAF15 hydrogels [66, 97]; we find that lysine residues within RNA Pol II CTD are also important for co-partitioning into SYGQ LC liquid condensates (Fig. 3.6). While there are some parallels between the arginine-rich RGG3 domain and lysine-rich RNA Pol II CTD, arginine-tyrosine pairs can form cation- π , π^2/π and hydrogen bonding interactions simultaneously, while lysine is limited to cation- π , hydrogen bonding and salt bridge interactions in this context (Fig. 3.3) and 3.4) Indeed, co-occurrence of all three interaction modes between arginine-tyrosine pairs may contribute to the role of these residues as interaction hotspots. Finally, SYGQ LC is not required for co-partitioning of RNA Pol II CTD into FUS liquid condensates and RGG domains contribute to the interaction with RNA Pol II CTD [2]. Other arginine-rich proteins have been shown to associate with RNA Pol II CTD such as the splicing factors SRSF1/SRSF2 [125]. Arginine and aspartate residues within TAF15 are important for RNA polymerase clustering [100]. Residue charge may contribute to the interaction between the RGG

domains and RNA Pol II CTD, as mutation of lysines in RNA Pol II CTD increased partitioning into RGG-rich droplets (Fig. 3.6B). Hence, phosphorylation of RNA Pol II CTD may alter the balance of interactions with FET SYGQ LC and RGG domains, potentially leading to changes in RNA Pol II and FET protein co-localization associated with the stages in transcription [96, 100, 102, 125, 126].

3.5 Tables and Figures

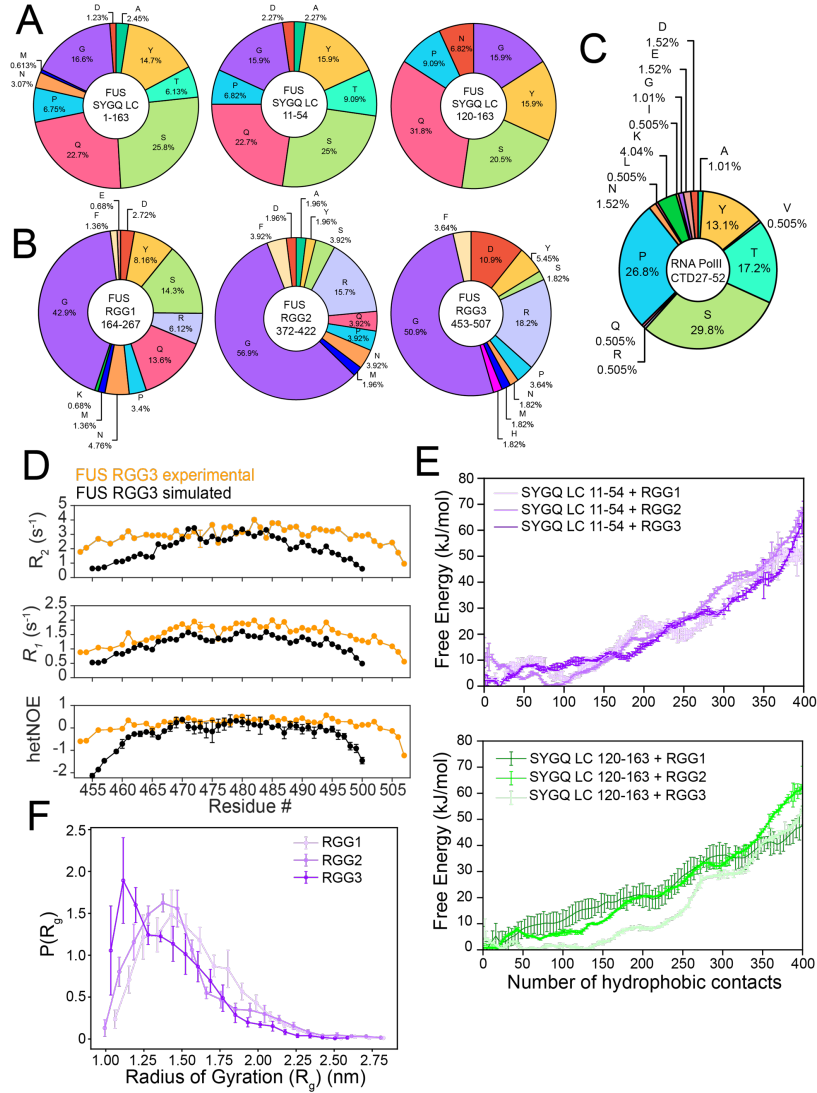


Figure 3.1: Composition of fragments used for all-atom simulations. **(A)** Amino acid content of the SYGQ LC and fragments 11-54 and 120-163 used for all-atom simulations. **(B)** Amino acid content of the FUS RGG domains. Fragments used for all-atom simulations (RGG1_{220–267}, RGG2_{372–419} and RGG3_{454–501}) contain similar amino acid compositions to their experimental counterparts. **(C)** Amino acid content of RNA polymerase C-terminal tail heptads 27-52. **(D)** ¹⁵N spin relaxation parameters for FUS RGG3 in the dispersed phase from experiment and simulations. The segments used for simulations are shorter, explaining the discrepancies at the termini. Experimental data are plotted as mean $\pm$ propagated best-fit parameter confidence interval equal to 1 s.d. in one representative data set of two independent experiments. Simulated data are plotted as mean $\pm$ s.e.m of $n = 12$ independent trajectories launched from randomly selected equilibrated ensemble members. **(E)** Free energy landscape as a function of van der Waals contacts formed between hydrophobic atoms in FUS_{11–54} or FUS_{120–163} and RGG1, RGG2 or RGG3 from simulations. The use of different 44-amino acid fragments of FUS LC in the simulations produces differences in the energy landscapes, suggesting that the amino acid variation between the fragments used can have an impact on the number of contacts. Data are plotted as mean $\pm$ s.e.m of $n = 5$ equal divisions of the total data set from one data set with $n=16$ independent replicas using PT-WTE. **(F)** Radius of gyration distribution of three different 44-residue long RGG fragments in single-chain simulations. The differences in compaction within the simulation system reflects the differences in amino acid composition of each RGG fragment. Data are plotted as mean $\pm$ s.e.m of $n = 5$ equal divisions of the total data set as in (E).

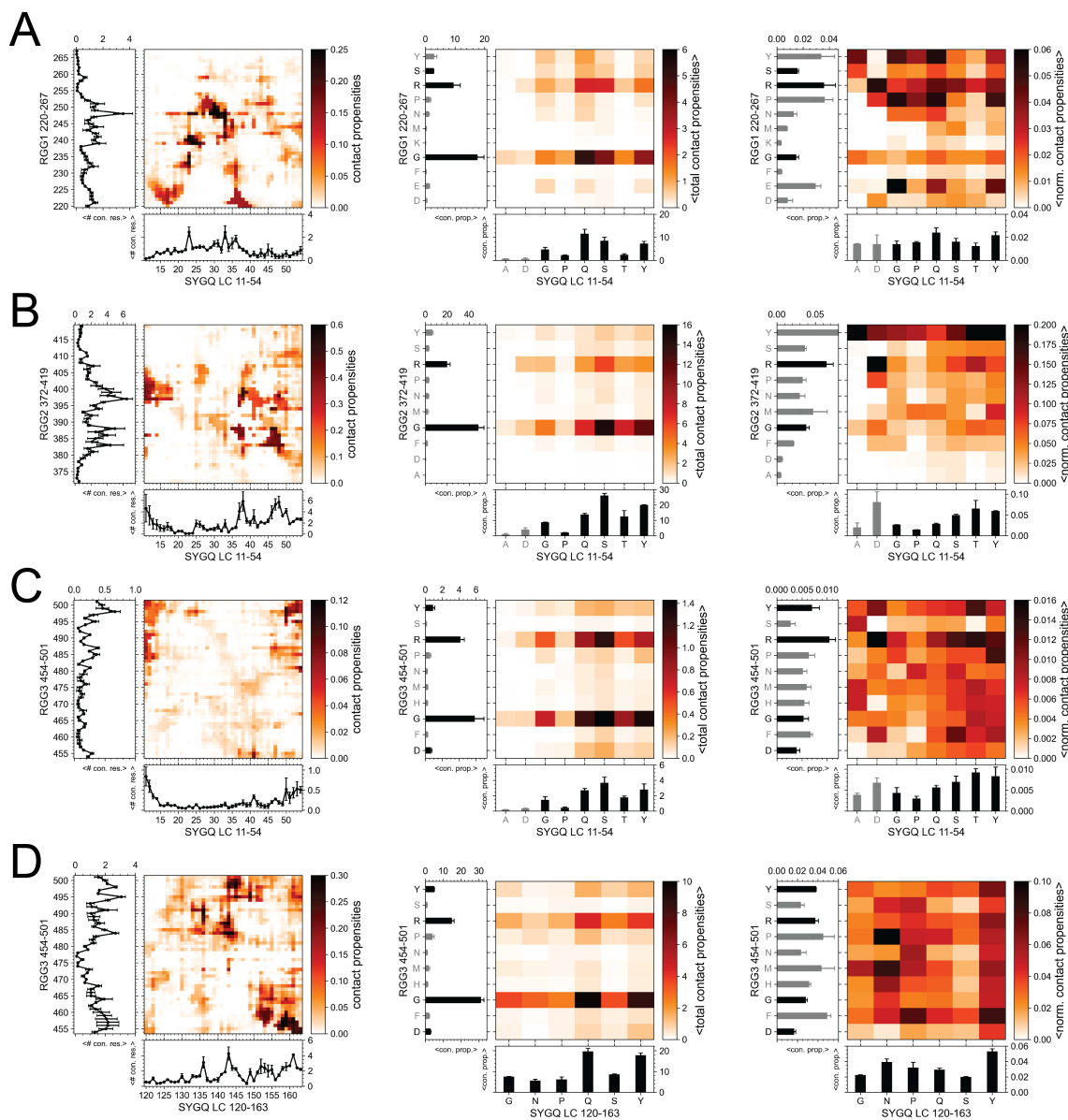


Figure 3.2: Contacts between FUS SYGQ LC and RGG domains. Inter-molecular contact propensities from two-chain simulations of SYGQ LC_{11–54} with (A) RGG1_{220–267}, (B) RGG2_{372–419}, and (C) RGG3_{454–501}, respectively, and (D) SYGQ LC_{120–163} with RGG3_{454–501}, binned by residue position (left), binned by residue type (center), and binned by residue type and normalized by residue frequency (right). Curve plots represent the total number of contacts for a particular residue position. Bars represent the total number of contacts for a particular residue type. Gray bars represent residue types that occur less than three times in the sequence. Data are plotted as mean $\pm$ s.e.m. of (left) $n = 5$ equal divisions of the total 16 replica PT-WTE data set, (middle) total contact propensities, or (right) normalized total contact propensities from one representative data set out of two independent experiments.

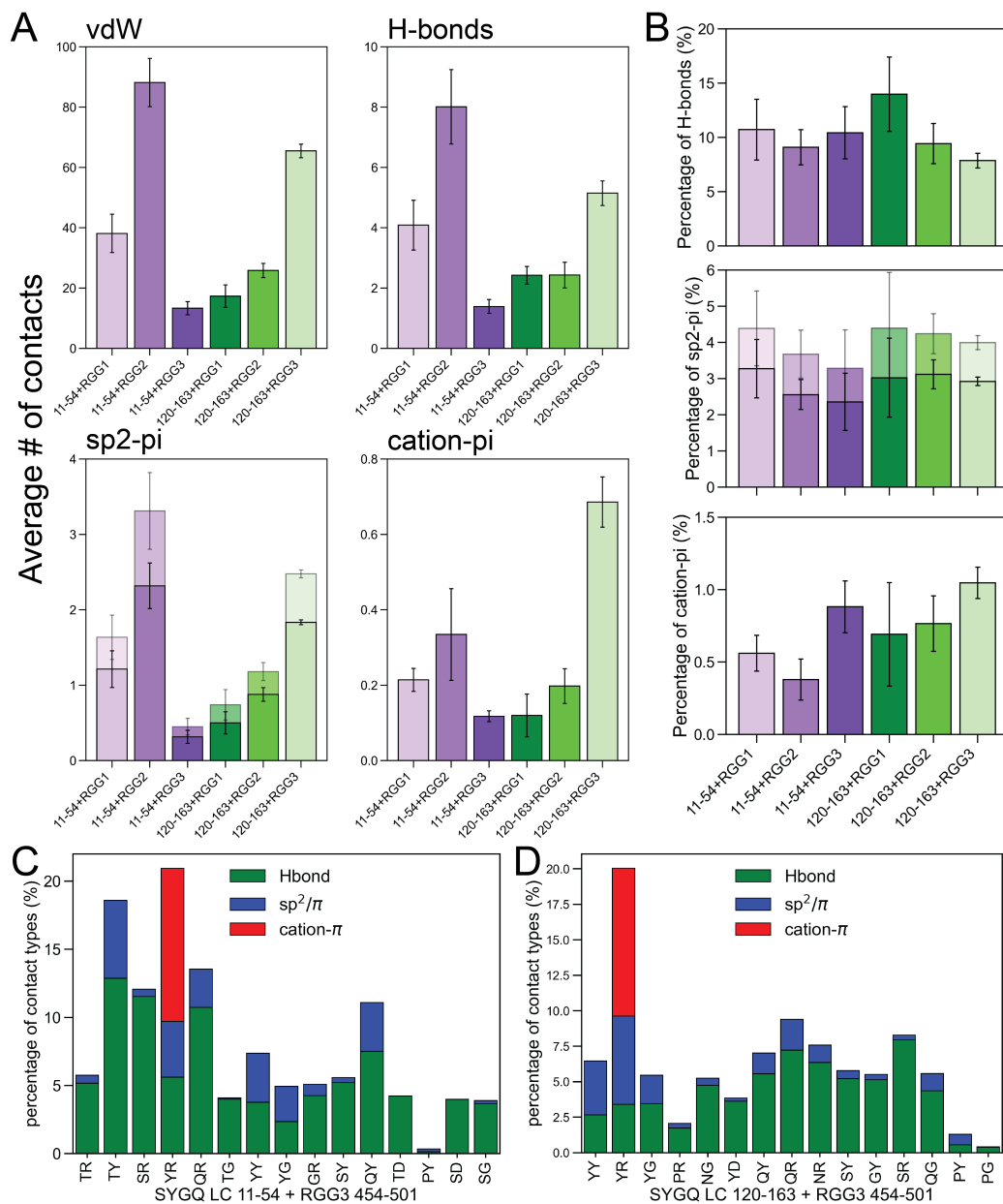


Figure 3.3: Diverse interaction modes contribute to FUS SYGQ LC and RGG3 interactions within the condensed phase. **(A)** Average number of intermolecular van der Waals contacts, hydrogen bonds, sp^2/π , and cation- π interactions from two-chain simulations of SYGQ LC₁₁₋₅₄ or SYGQ LC₁₂₀₋₁₆₃ and RGG1, RGG2 or RGG3. For sp^2/π contacts the data are binned for π - π (top, lightest), sp^2 - π (middle, lighter) and sp^2 - sp^2 (bottom) contacts. Data are plotted as mean $\pm$ s.e.m of $n = 5$ equal divisions of the total data set from one PT-WTE simulation with $n = 16$ independent replicas. **(B)** Percentage of hydrogen bond, sp^2/π , and cation- π interactions out of the total intermolecular contacts from two-chain simulations of SYGQ LC₁₁₋₅₄ or SYGQ LC₁₂₀₋₁₆₃ and RGG1, RGG2 or RGG3. See (A) for sp^2/π shading definition and statistics information. **(C-D)** The percentage of hydrogen bonds, sp^2/π , and cation- π contacts out of the total van der Waals interactions for interacting amino acid type pairs. Data are shown only for top fifteen interacting amino acid type pairs normalized for their frequency and selected from residue types occurring more than 2 times in the respective sequence in order of highest to lowest contact frequency (left to right) SYGQ LC₁₁₋₅₄ or SYGQ LC₁₂₀₋₁₆₃ with RGG3₄₅₄₋₅₀₁.

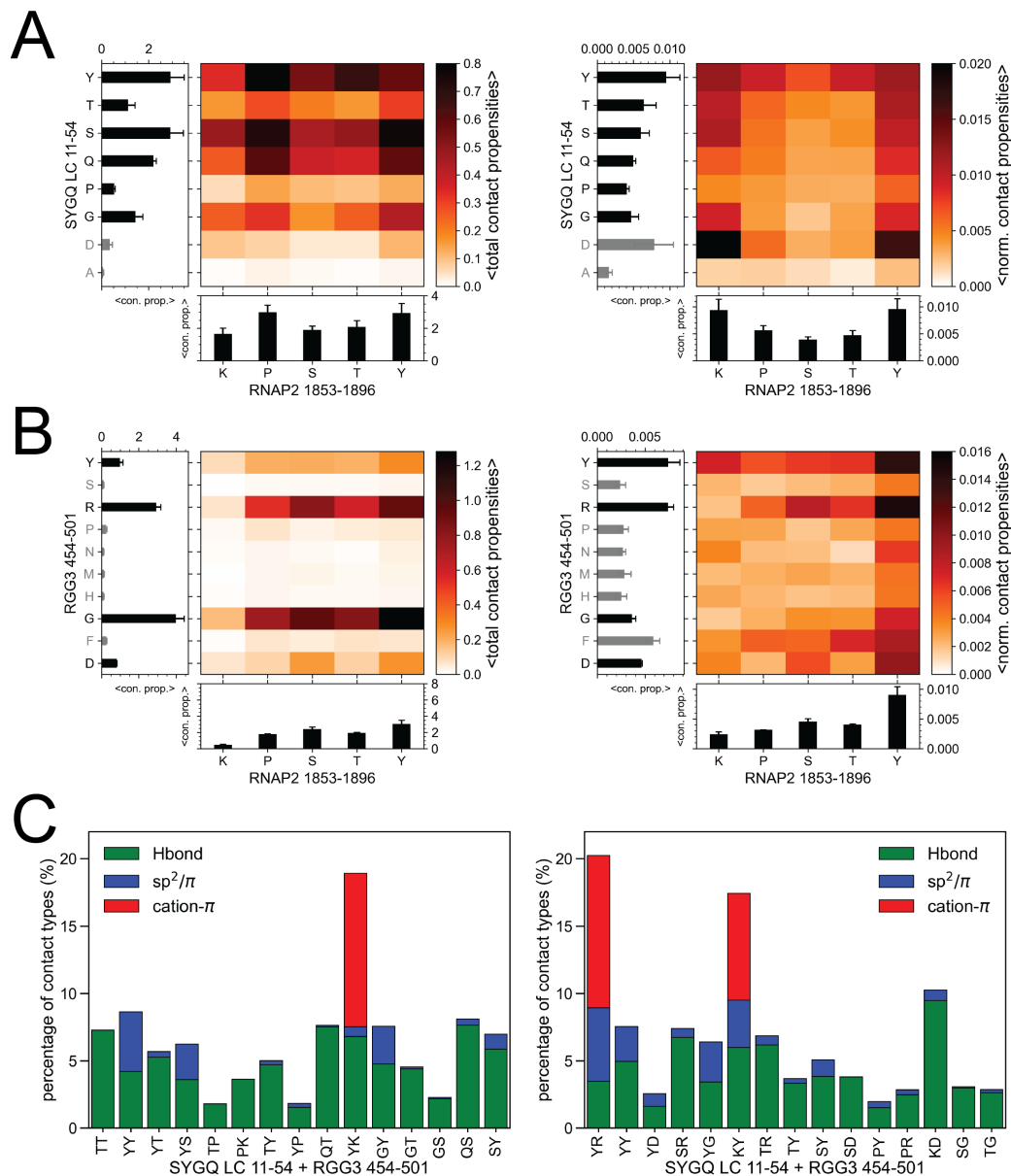


Figure 3.4: Lysine residues within RNA Pol II CTD contribute to interactions with FUS within the condensed phase. Inter-molecular contact propensities from two-chain simulations of RNAP2_{1853–1896} with (A) SYGQ LC_{11–54}, (B) RGG3_{454–501}, respectively, binned by residue type (left), and binned by residue type and normalized by residue frequency (right). Curve plots represent the total number of contacts for a particular residue position. Bars represent the total number of contacts for a particular residue type. Gray bars represent residue types that occur less than three times in the sequence. Data are plotted as mean $\pm$ s.e.m of (left) $n = 5$ equal divisions of the total 16 replica PT-WTE data set. (C) Comparison of the dominant interaction modes between residues within SYGQ LC_{11–54} and RNAP2_{1853–1896} (left) or RGG3_{454–501} and RNAP2_{1853–1896} (right) in decreasing order of normalized contact frequency among residue types present 3 or more times in the sequences.

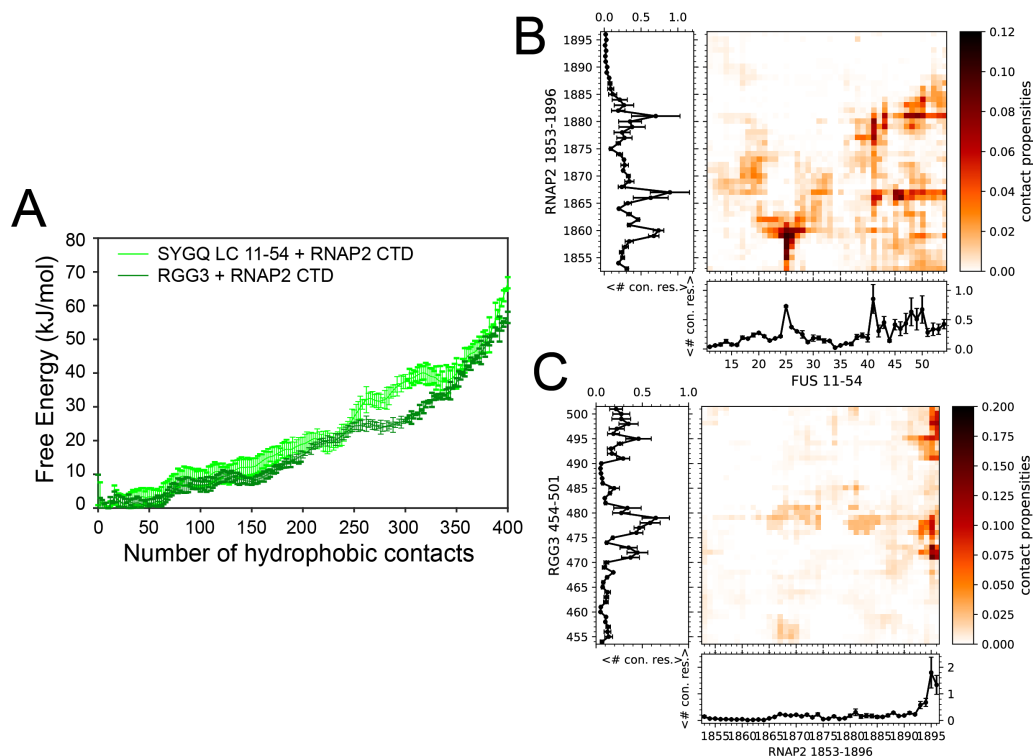


Figure 3.5: Contacts within a three-component phase containing FUS SYGQ LC and RGG3 and RNAP2 CTD. **(A)** Free energy landscape of van der Waals contacts between hydrophobic atoms between RNAP2 CTD and FUS SYGQ LC or RGG3 from two-chain simulations. Data are plotted as mean $\pm$ s.e.m of $n=5$ equal divisions of the total data set from one representative data set. **(C)** ^1H - ^{15}N HSQC of RNA Pol II CTD in the dispersed (orange) and condensed (green) phases. Inter-residue contact propensities from two-chain simulations of RNAP2 CTD_{1853–1896} and **(B)** FUS SYGQ LC_{11–54} or **(C)** RGG3, respectively, binned by residue position. Plots represent the total number of contacts for a particular residue position. Data are plotted as mean $\pm$ s.e.m of $n = 5$ equal divisions of the total data set from one representative data set.

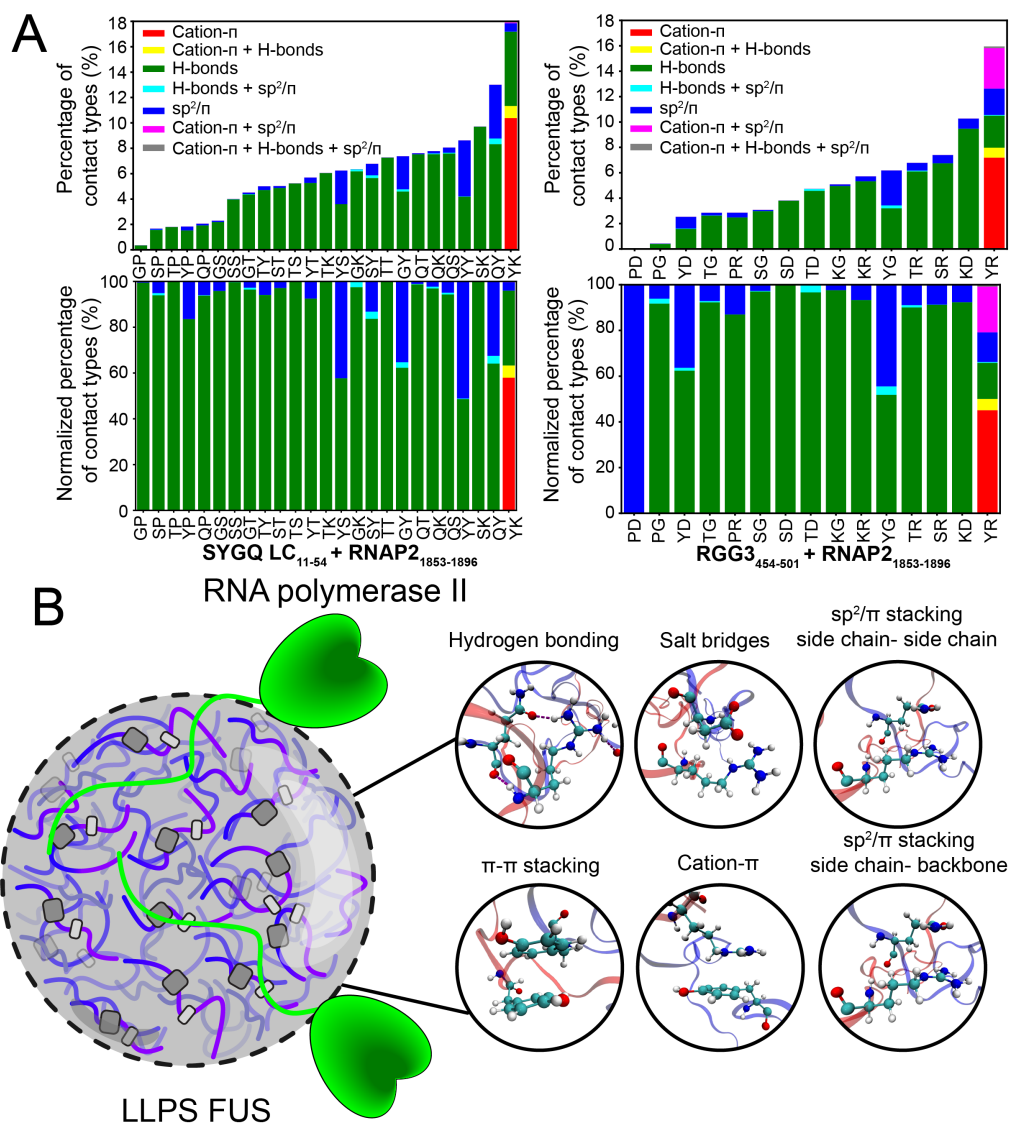


Figure 3.6: Multiple interaction modes contribute to multi-component condensed phases of FET proteins and RNA Pol II. **(A)** Comparison of the dominant interaction modes between residues within SYGQ LC₁₁₋₅₄ and RNAP2₁₈₅₃₋₁₈₉₆ or RGG3₄₅₄₋₅₀₁ and RNAP2₁₈₅₃₋₁₈₉₆. **(B)** The interactions between FUS and RNA Pol II CTD that stabilized LLPS involve hydrogen bonding, salt bridges, π - π stacking as well as cation- π interaction modes.

Chapter 4

SINATRA Pro: A Topological Data Analytic Approach for Discovering Biophysical Signatures in Protein Dynamics

This chapter is adapted from the paper [3]. I would like to acknowledge all the -authors, Gabriel Monteiro da Silva, Henry Kirveslahti, Erin Skeens, Bibo Feng, Timothy Sudijono, Kevin K. Yang, Sayan Mukherjee, Brenda Rubenstein, and Lorin Crawford, for their contributions in this paper [3]. My co-first author, Gabriel Monteiro da Silva, conducted all of the MD simulations for this work and contributed the majority of the writing for the published manuscript regarding the biological backgrounds of this study.

4.1 Introduction

Identifying structural features associated with macromolecular dynamics is crucial to our understanding of the underlying physical behavior of proteins and their broader impact on biology and

health. Structural and dynamical properties of proteins often serve as signatures of their functions and activities [127]. Subtle topological changes in protein conformation can lead to dramatic changes in biological function [128, 129], thus highlighting the importance of being able to accurately characterize protein conformational dynamics.

Conventionally, the structural dynamics of proteins have been modeled using molecular dynamics (MD) simulations, which work by sampling structural ensembles from conformational landscapes. In infinite timescales, such structural ensembles are expected to represent all physical states such that their ensemble-averaged observables converge to true physical values and are thus physically meaningful. While MD simulations have provided key insights into the atomistic motions that underpin many protein functions [130], biologically-relevant structural changes can be overshadowed by spurious statistical noise caused by the thermal fluctuations that naturally arise during the course of these simulations [131]. In practice, this can often make important structural features difficult to identify and robustly interpret from MD trajectories. Data from MD simulations are often analyzed in a strictly goal-dependent manner by using computational methods that quantify and assess specific protein characteristics. For example, geometric changes that arise as a result of ligand binding, point mutations, or post-translational modifications are usually inferred by analyzing the root mean square fluctuations (RMSF) of atomic positions or the per-domain radius of gyration with respect to a reference structure [132]. Unfortunately, these standard approaches are less powerful when the relevant changes in protein structure are overshadowed by fluctuations irrelevant to the biological process of interest.

Recently, more sophisticated methods have aimed to overcome these challenges by taking advantage of correspondences between the atomic positions on any two given proteins. For example, per-residue distance functions or contact maps can be calculated on each frame of a trajectory for clustering [133] or principal component analyses (PCA) [134, 135], which project complex conformations onto a lower-dimensional space for ease of comparison. However, the downside to these methods is that they require diffeomorphisms between structures (i.e., the map from protein A to protein B must be differentiable). There are many scenarios in protein dynamics where no such transformation is guaranteed because atomic features can be gained or lost during the evolution of the system [136]. Indeed, there are 3D shape algorithms that construct more general “functional” correspondences and can be applied even across shapes having different topology [137, 138]; however, previous work has shown that the performance of these algorithms drops significantly when

the assumed functional mapping input is even slightly misspecified [139].

In this chapter, we introduce SINATRA Pro: a topological data analytic pipeline for identifying biologically-relevant structural differences between two protein structural ensembles without the need for explicit contact maps or atomic correspondences. Our algorithm is an extension of a previous framework, SINATRA, which was broadly introduced to perform variable selection on physical features that best describe the variation between two groups of static 3D shapes [139]. Using a tool from integral geometry and differential topology called the Euler characteristic (EC) transform [140–143], SINATRA was shown to have the power to identify known morphological perturbations in controlled simulations and robustly identify anatomical aberrations in mandibular molars associated within four different suborders of primates. SINATRA Pro is an adaptation of the SINATRA framework for protein dynamics. Here, we develop a simplicial complex construction step to specifically model both 3D geometric and topological relationships between atomic positions on protein structures. We also utilize a new set of statistical parameters which we calibrate for complex protein systems.

In the following sections, we demonstrate SINATRA Pro’s ability to identify key structural and dynamical features in a hierarchy of proteins with increasingly challenging features to statistically resolve. The five proteins studied, TEM β -lactamase, the Abelson Kinase (Abl1), the HIV-1 protease, Elongation Factor Thermo Unstable (EF-Tu), and Importin- β , undergo structural changes in response to a wide range of well-studied biological phenomena, including mutations, interactions with partners, and small molecule binding. We show that SINATRA Pro outperforms standard analytic techniques including RMSF and PCA in consistently pinpointing physical locations of biologically-relevant conformational changes. Overall, we find that SINATRA Pro holds great promise for extracting topological differences between two sets of protein structures from meaningless statistical noise.

4.1.1 Pipeline Overview

The SINATRA Pro pipeline involves five key steps (see Fig. 4.1). First, the algorithm begins by taking aligned structures from two all-atom protein MD simulation trajectories of different phenotypic states (e.g., wild-type versus mutant) as inputs (Fig. 4.1(a)). In the second step, SINATRA Pro uses the 3D Cartesian coordinates of each atom of the proteins to create mesh representations of their 3D structures (Fig. 4.1(b)). Here, atoms within a predetermined physical distance cutoff (e.g.,

~ 6 Ångströms (Å) apart) are connected by “edges” and then triangles enclosed by the connected edges are filled to create “faces.” In the third step, we convert the resulting triangulated meshes to a set of topological summary statistics using an invariant called the “differential Euler Characteristics (DEC)” transform (Fig. 4.1(c)). In the fourth step, SINATRA Pro implements a nonlinear Gaussian process model to classify the protein structures using the topological summary statistics, with which association measures are computed for each topological feature to provide a statistical notion of “significance” (Fig. 4.1(d)). In the last step of the pipeline, SINATRA Pro maps the association measures back onto the original protein structures (Fig. 4.1(e)), which produces “evidence scores” that reveal the spatial locations that best explain the variance between two protein ensembles.

The software for SINATRA Pro requires the following inputs: (i) 3D Cartesian coordinates corresponding to the atomic positions in each protein structure; (ii) $\mathbf{y}$, a binary vector denoting protein class or phenotype (e.g., $y_i = 0$ for wild-type or $y_i = 1$ for mutant); (iii) r , the cutoff distance for simplicial construction (i.e., constructing the mesh representation for every protein); (iv) c , the number of cones of directions; (v) d , the number of directions within each cone; (vi) θ , the cap radius used to generate directions in a cone; and (vii) l , the number of sublevel sets (i.e., filtration steps) used to compute the differential Euler characteristic (DEC) curve along any given direction. In addition to real data analyses, we implement a controlled benchmark simulation study designed to assess SINATRA Pro’s performance at identifying structurally-perturbed regions in protein dynamics relative to other methods. Results for the controlled benchmark simulations were done using free parameters $\{r = 1.0 \text{ Å}, c = 20, d = 8, \theta = 0.80, l = 120\}$, and results for the real data analyses were done using parameters $\{r = 6.0 \text{ Å}, c = 20, d = 8, \theta = 0.80, l = 120\}$. All values were chosen via a grid search. Guidelines for how to choose the free parameters for the software are given in Table 4.1. Tables detailing the scalability of the current algorithmic implementation of SINATRA Pro can also be found in Supporting Information (see Tables B.1).

4.1.2 Converting Protein Structure Data to 3D Mesh Representations

To convert protein structures into a mesh representation, in the first step of the SINATRA Pro pipeline, we make use of a technique which we refer to as a “simplicial construction” (Fig. 4.1(b)). In this procedure, we treat the atomic positions for the protein as vertices on a 3D shape or surface. First, we draw an edge between any two atoms if their Euclidean distance is smaller than some radius cutoff r , namely $\text{dist}[(x_1, y_1, z_1), (x_2, y_2, z_2)] < r$. Next, we fill in all of the triangles (or faces)

formed by these connected edges. The resulting triangulated meshes are then normalized to the unit sphere, which means that the coordinates for all atoms are scaled with respect to the mesh with the largest radius. We treat the normalized meshes as simplicial complexes which we then use to compute topological summary statistics.

4.1.3 Topological Summary Statistics for Protein Mesh Representations

Adopted from its predecessor [139], the second step of the SINATRA Pro pipeline uses a tool from integral geometry and differential topology called the Euler characteristic (EC) transform [140–143]. As a brief overview of this approach, given the mesh representation $\mathcal{M}$ of a protein structure, the Euler characteristic is an accessible topological invariant defined as

$$\chi = \#V(\mathcal{M}) - \#E(\mathcal{M}) + \#F(\mathcal{M}), \quad (4.1)$$

where the collection $\{\#V(\mathcal{M}), \#E(\mathcal{M}), \#F(\mathcal{M})\}$ denotes the number of vertices (atoms), edges (connections between atoms), and faces (triangles enclosed by edges) of the mesh, respectively. An EC curve $\chi_\nu(\mathcal{M})$ tracks the change in the Euler characteristic with respect to a given filtration of length l in some direction ν . Theoretically, this is done by first specifying a height function $h_\nu(\mathbf{x}) = \mathbf{x}^\top \nu$ for some atomic position $\mathbf{x} \in \mathcal{M}$ in direction ν . This height function is then used to define sublevel sets (or subparts) of the mesh $\mathcal{M}_\nu^a$ in direction ν , where $h_\nu(\mathbf{x}) \leq a$. In practice, the EC curve is $\chi(\mathcal{M}_\nu^a)$ computed over a range of l filtration steps in direction ν . The corresponding EC transform is defined as the collection of EC curves across a set of $\nu = 1, \dots, m$ directions, and maps onto a 3D protein structure as a concatenated $J = (l \times m)$ -dimensional feature vector to be used for statistical analyses.

In previous studies, it has been observed that the Euler characteristic can be a less-than-optimal shape summary statistic when inter-class variation between 3D objects is high and driven by local fluctuations in morphology [139, 142–145]. Given that this situation can be quite common in molecular dynamics, we introduce a new topological invariant which we refer to as the differential Euler characteristic (DEC) (see Fig. 4.1(c)). As an alternative to Eq. 4.1, the DEC is computed as the following

$$\Delta\chi = \Delta V(\mathcal{M}) - \Delta E(\mathcal{M}) + \Delta F(\mathcal{M}), \quad (4.2)$$

where, for some lag parameter t , we define $\Delta V(\mathcal{M}) = \#V_l(\mathcal{M}) - \#V_{l-t}(\mathcal{M})$, $\Delta E(\mathcal{M}) = \#E_l(\mathcal{M}) - \#E_{l-t}(\mathcal{M})$ and $\Delta F(\mathcal{M}) = \#F_l(\mathcal{M}) - \#F_{l-t}(\mathcal{M})$. In this study, we set $t = 1$ such that, intuitively,

the DEC tracks the changes (i.e., the local appearance or disappearance of topological features) in the number of vertices, edges, and faces from one sublevel set to the next. Much like with the original Euler characteristic, the DEC curve is $\Delta\chi(\mathcal{M}_\nu^a)$ computed over a range of l filtration steps in a given direction ν and the DEC transform is similarly defined as the collection of DEC curves across a set of $\nu = 1, \dots, m$ directions. Overall, for each dataset with N total proteins, an $N \times J$ design matrix $\mathbf{X}$ is statistically analyzed, where the columns denote the differential Euler characteristic computed at a given filtration step and direction. Each sublevel set value, direction, and set of atomic positions used to compute a DEC curve are stored by the algorithm for the association mapping and projection phases of the pipeline.

4.1.4 Choosing the Number of Directions and Filtration Steps.

In this paper, we use a series of simulations and sensitivity analyses to develop an intuition as to how to set the granularity of sublevel filtrations l and choose the number of directions m for real protein structure data (Fig. 4.1, Fig. B.1, and Table 4.1). Since the structural changes that a protein class exhibits can occur on both a global and local scale, depending on its biophysical and chemical properties, we recommend choosing the former parameter l via cross validation or a grid-based search. For the latter, the SINATRA Pro software defines the total number of directions m as the union of c sets of cones of directions $\mathcal{D} = \bigcup \mathcal{C}_k(\theta)$, where each cone $\mathcal{C}_k(\theta) = \{\nu_{k,1}, \dots, \nu_{k,d} \mid \theta\}$ for $k = 1, \dots, c$ is parameterized by a cap radius θ from which equidistant vectors are generated over the unit sphere. We use cones because local shape information matters most when determining reconstructed manifolds and it has been shown that topological invariants that are measured in directions of close proximity contain similar local information [139, 141, 146, 147]. This naturally leads to the construction of sets $\mathcal{C}_k(\theta)$ where the angle θ between them is relatively small (again see Table 4.1). In general, we use sufficiency results for topological transforms (see Theorem 7.14 in Curry et al. [141]) to motivate the notion that considering larger numbers of $m = c \times d$ directions will lead to a more robust summary of 3D shapes and surfaces. Hence, ideally, one would select an effectively large number of c cones (and d directions within these cones) to ensure that SINATRA Pro is summarizing all relevant structural information about the variance between phenotypic classes (e.g., mutants versus wild-type).

4.1.5 Probabilistic Model for Protein Structure Classification

In the third step of the SINATRA Pro pipeline, we use (weight-space) Gaussian process probit regression model to classify protein structures based on their topological summaries generated by the DEC transformation via Eq. 4.2. Here, we specify the following probabilistic hierarchical model (Fig. 4.1(d)) [148–152]

$$\mathbf{y} \sim \mathcal{B}(\boldsymbol{\pi}), \quad g(\boldsymbol{\pi}) = \Phi^{-1}(\boldsymbol{\pi}) = \mathbf{f}, \quad \mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad (4.3)$$

where $\mathbf{y}$ is an N -dimensional vector of Bernoulli distributed phenotypic class labels (e.g., mutants versus wild-type), $\boldsymbol{\pi}$ is an N -dimensional vector representing the underlying probability that a shape is classified as a “class” (e.g., $y = 1$ if “mutant”), $g(\cdot)$ is a probit link function with $\Phi(\cdot)$ the cumulative distribution function (CDF) of the standard normal distribution, and $\mathbf{f}$ is an N -dimensional vector estimated from the data. We take a classic kernel regression approach [151, 153–155] where we posit that $\mathbf{f}$ lives within a reproducing kernel Hilbert space (RKHS) defined by some (nonlinear) covariance function, which implicitly accounts for higher-order interactions between features, leading to more complete classifications of structural data [156–158]. To this end, we assume $\mathbf{f}$ is normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{K}$ with elements defined by the radial basis function $\mathbf{K}_{ii'} = \exp\{-\vartheta\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2\}$ with bandwidth ϑ set using the “median criterion” approach to maintain numerical stability and avoid additional computational costs [159]. Here, $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ denote the collection of topological features for the i -th and i' -th observation in $\mathbf{X}$, respectively. The full model specified in Eq. 4.3 is commonly referred to as “Gaussian process classification” (GPC).

Given the complete specification of the GPC, we use Bayesian inference to draw samples from the posterior distribution of the latent variables, which is proportional to $p(\mathbf{f} | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{f}) \times p(\mathbf{f})$. Here, $p(\mathbf{y} | \mathbf{f})$ denotes the likelihood of the observed binary labels given the functions (i.e., the Bernoulli distribution), and $p(\mathbf{f})$ is the prior distribution for the latent variables (i.e., the multivariate normal distribution). The probit likelihood in Eq. 4.3 makes it intractable to estimate the posterior distribution $p(\mathbf{f} | \mathbf{y})$ via a closed-form solution. We instead use a Markov chain Monte Carlo (MCMC) method called “elliptical slice sampling” to conduct posterior inference (see Data and Software Dependencies) [160].

4.1.6 Feature Selection of Topological Summary Statistics

After implementing the elliptical slice sampling algorithm to estimate the posterior distribution of the latent variables $\mathbf{f}$ in Eq. 4.3, we define a nonparametric effect size for each topological summary statistic via the following standard projection [157, 161]

$$\boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{f}, \quad (4.4)$$

where $\mathbf{M}^+$ is used to denote the generalized inverse of a matrix $\mathbf{M}$, and each element in $\boldsymbol{\beta}$ details the nonlinear relationship between the DEC topological summary statistics and the variance between protein structures. In order to determine a statistical rank ordering for these effect sizes, we assign an information theoretic-based measure of relative centrality to each j -th topological feature using Kullback-Leibler divergence (KLD) [158]

$$\text{KLD}(\beta_j) := \text{KL} [p(\boldsymbol{\beta}_{-j}) \parallel p(\boldsymbol{\beta}_{-j} | \beta_j = 0)] = \int_{\boldsymbol{\beta}_{-j}} \log \left(\frac{p(\boldsymbol{\beta}_{-j})}{p(\boldsymbol{\beta}_{-j} | \beta_j = 0)} \right) p(\boldsymbol{\beta}_{-j}) d\boldsymbol{\beta}_{-j}. \quad (4.5)$$

for $j = 1, \dots, J$ topological features. Finally, we normalize to obtain an association metric (Fig. 4.1(d)),

$$\gamma_j = \text{KLD}(\beta_j) / \sum \text{KLD}(\beta_l). \quad (4.6)$$

There are two key takeaways from this scaled formulation. First, the KLD is non-negative, and it equals zero if and only if the posterior distribution of $\boldsymbol{\beta}_{-j}$ is independent of the effect β_j . Intuitively, this is equivalent to saying that removing an unimportant topological feature has no impact on explaining the variance between different protein structure. Second, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_J)$ is bounded on the unit interval $[0, 1]$ with the natural interpretation of providing relative evidence of association for each DEC statistic (where values close to 1 suggest greater importance). From a classical hypothesis testing point-of-view, the null hypothesis for Eq. 4.6 assumes that every DEC feature equally contribute to the total variance between proteins, while the alternative hypothesis proposes that some DEC features are better associated with biophysical changes in protein structures than others [139, 158].

4.1.7 Closed Form Solution for Atomic-Level Association Measures

For simplicity, we assume that the implied posterior distribution of $\boldsymbol{\beta}$ (deterministically given in Eq. 4.4) is approximately multivariate normal with an empirical mean vector $\boldsymbol{\mu}$ and positive semi-definite covariance/precision matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$ [139, 158]. Given these values, we iteratively partition

such that, for each j -th topological feature:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_j \\ \boldsymbol{\beta}_{-j} \end{pmatrix}; \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_j \\ \boldsymbol{\mu}_{-j} \end{pmatrix}; \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_j & \boldsymbol{\sigma}_{-j}^\top \\ \boldsymbol{\sigma}_{-j} & \boldsymbol{\Sigma}_{-j} \end{pmatrix}; \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_j & \boldsymbol{\lambda}_{-j}^\top \\ \boldsymbol{\lambda}_{-j} & \boldsymbol{\Lambda}_{-j} \end{pmatrix}. \quad (4.7)$$

Under normality assumptions, Eq. 4.5 has the following closed form solution

$$\text{KLD}(\beta_j) = \frac{1}{2} \left[-\log|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}| + \text{tr}(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - J + \alpha_j(\beta_j - \mu_j)^2 \right], \quad (4.8)$$

where $\log|\cdot|$ represents the matrix log-determinant function, and $\text{tr}(\cdot)$ is the matrix trace function. Importantly, the term $\alpha_j = \boldsymbol{\lambda}_{-j}^\top \boldsymbol{\Lambda}_{-j}^{-1} \boldsymbol{\lambda}_{-j}$ characterizes the linear (and non-negative) rate of change of information when the effect of any topological feature is absent from the analysis [158]. By symmetry in the notation for elements of the sub-vectors and sub-matrices, we simply permute the order of the variables in $\boldsymbol{\beta}$ and iteratively compute the KLD to measure the centrality of each DEC transform.

4.1.8 Performance of SINATRA Pro on Benchmark Simulations

We implemented a controlled simulation study designed to assess SINATRA Pro’s performance at identifying structurally-perturbed regions in protein dynamics relative to other methods. Here, the premise behind “controlled simulations” is that topological artifacts (i.e., perturbations of atomic positions in a certain region) are manually introduced to a set of protein structures to establish a ground truth and statistically evaluate the concept of power. The original and perturbed structures represent two phenotypic classes and are fed into SINATRA Pro to assess whether it can reliably identify the perturbed regions of interest.

To generate data for these controlled simulations, we use real structural data of wild-type β -lactamase (TEM), an enzyme widely implicated in microbial resistance that has evolved numerous mutations of clinical relevance. In the first phenotypic group (set A), original structures are drawn at 1 nanosecond (ns) intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Next, a comparable set of perturbed structures (set B) are drawn at 1 ns intervals but shifted by 0.5 ns with respect to the set A structures (e.g., $t_{\text{MD}} = [0.5, 1.5, 2.5, 3.5, \dots, 99.5] \text{ ns} + \delta$) to allow for thermal noise to be introduced. Here, we displace the atomic positions of each atom in the Ω -loop (i.e., the region of interest or ROI) in each perturbed structure within set B by

- a constant Cartesian vector set to (i) 0.5 Å, (ii) 1.0 Å, and (iii) 2.0 Å in each (x, y, z) direction;
- a spherically uniform random vector where each (x, y, z) direction is first drawn from a standard Gaussian distribution $\mathcal{N}(0, 1)$ and then the vector is normalized to be of length (iv) 0.5 Å, (v) 1.0 Å, and (vi) 2.0 Å.

These simple, artificial control cases are designed to represent two different forms of structural changes that can happen within protein dynamics. Namely, scenarios (i)-(iii) involve a displacement of atoms by a constant amount in a constant direction, which emulates a static structural change; while, scenarios (iv)-(vi) displace the atoms by a constant amount in a (spherically uniform) random direction, which emulates a dynamic or stochastic structural change. Altogether, we use datasets of $N = 1000$ protein structures per simulation scenario: 100 ns intervals $\times$ 5 different choices of $\delta = \{0.0, 0.1, 0.2, 0.3, 0.4\}$ ns $\times$ 2 phenotypic classes (wild-type versus perturbed). We evaluate all competing methods' abilities to correctly identify perturbed atoms located within the Omega-loop region (Material and Methods). Here, we use receiver operating characteristic (ROC) curves that plot true positive rates (TPR) against false positive rates (FPR) (Fig. 4.2). This is further quantified by assessing the area under the curve (AUC). The results presented in the main text reflect using SINATRA Pro with parameters set to $\{r = 1.0 \text{ Å}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. Note that additional figures assessing how robust SINATRA Pro is to different free parameter value settings can be found in the Supporting Information (see sensitivity analysis in Fig. B.1).

Lastly, we want to point out that we do not realign the proteins after structural perturbation has occurred. Realigning the structures after introducing a perturbation poses a slightly different and notably less controlled simulation study. For example, in the constant displacement case, realigning the structures will shift the whole structure against the perturbing vector and result in an unintentional displacement on the opposite side of the structure. A true positive in this case will not be as well-defined as when we keep the unperturbed structure in place and define the perturbed structure as the ground truth for the positive signal. An example of this can be seen in Fig. B.2.

4.1.9 Overview of Competing Baselines.

In this section, we compare SINATRA Pro to four competing approaches: root mean square fluctuation (RMSF) calculations, principal component analysis (PCA), Elastic Net classification, and

Neural Network classification. The first baseline is RMSF which computes $\sqrt{\|\mathbf{x} - \langle \mathbf{x} \rangle\|^2}$, where $\mathbf{x} = (x, y, z)$ denotes the positions of the protein’s atoms for each frame and $\langle \mathbf{x} \rangle$ is the average position of that corresponding atom over the entire MD simulation. The difference in the RMSF values between the original and perturbed structures is taken as the score for feature selection. The second baseline performs PCA (based on singular value decomposition) over the Cartesian (x, y, z) -coordinates for the atoms using `scikit-learn` [162], which reduces the sample space into principal components. We sum the components (weighed by their singular values) for the original wild-type and perturbed data separately, and then determine the magnitudes of the change in the component sum between the two protein classes as the score for feature selection. The last two baselines concatenate the coordinates of all atoms within each protein and treats them as features in a dataframe. The Elastic Net uses a regularized linear classification model via stochastic gradient descent in `scikit-learn` to assign sparse individual coefficients to each coordinate of every atom, where the free regularization parameter is chosen with 90% training and 10% validation set splits. We assess the power of the Elastic Net by taking the sum of the coefficient values corresponding to each atomic position. The Neural Network uses the following architecture with Rectified Linear Unit (ReLU) nonlinear activation functions [163]: (1) an input layer of Cartesian coordinates of all of the atoms; (2) a hidden layer with $H = 2048$ neurons; (3) a second hidden layer with $H = 512$ neurons; (4) a third hidden layer with $H = 128$ neurons; and (5) an outer layer with a single node which uses a sigmoid link function for protein classification. Batch normalization was implemented between each layer and a normalized saliency map to rank the importance of each atom [164]. The simplest saliency map attributes the partial derivatives $\partial y_i / \partial \mathbf{x}_{ij}$ as the importance of the coordinates for the j -th atom in the i -th protein structure; here, y_i denotes the neural network output after the sigmoid link function for the i -th protein structure. We then assign global importance to each atom by $\sum_{i=1}^{N_c} |\partial y_i / \partial \mathbf{x}_{ij}| / N_c$, where N_c denotes the number of protein structures in a given class. For the Neural Network, we assess power by taking the sum of the saliency map values corresponding to each atomic position. To ensure that both PCA and the Neural Network are evaluated with their optimal parameter configurations, we also performed variants of these approaches in their own sensitivity analysis. The results for these variants of PCA and the Neural Network can be found in Supporting Information (see Figs. B.3 and B.4, respectively).

It is important to note that, while SINATRA Pro is implemented over the entire protein structure, the four baselines that we consider are limited to only assessing structural differences between

atoms with correspondences between the two datasets — that is, for the competing baselines, we use all “pairable” atoms and omit mutated residues which be considered “unpairable” between structures. The main reason for this is that atomic features can be gained or lost due to mutations or phylogenetic variations that introduce heterogeneity in protein sequences, thus creating a lack of a one-to-one correspondence between any two given 3D structures. Without this explicit mapping between structures, none of the four coordinate-based competing approaches are able to be fully implemented as they all rely on (in some way or another) equal dimensionality across all proteins. Therefore, when assessing MD trajectories, the corresponding or “pairable” atoms represent consistent “landmarks” that summarize the global geometry of the protein structure. Ultimately, we recognize that these method comparisons with SINATRA Pro are not equivalent; however, they do highlight a key and practical advantage of the topological data analytic approach used in SINATRA Pro which maintains its utility even when such atom-by-atom correspondences between protein structures are not available.

4.1.10 Method Comparisons.

The overall performance of each competing method to identify is dependent on two factors: (1) whether the structural changes are reproduced by static or stochastic conformations, and (2) the underlying statistical assumptions of the methods. For example, RMSF had the most difficulty identifying constant displacements in the protein structures (Figs. 4.2(a)-(c)). In these scenarios, RMSF was effectively a random classifier with an average AUC ≈ 0.5 and diagonal ROC curves showing no signal detected. These results are explained by that fact that RMSF effectively measures how much each atomic coordinate $\mathbf{x}$ deviates from the average atomic position in the ensemble $\langle \mathbf{x} \rangle$. When conformations are constantly shifted, $\mathbf{x}$ and $\langle \mathbf{x} \rangle$ are scaled by the same factor and, as a result, their differences remain unchanged. Therefore, static structural changes are essentially undetectable by RMSF. On the other hand, RMSF is perfectly well suited for stochastic structural changes because, when atomic displacements are caused by a random spherical vector, the scaling factors between each $\mathbf{x}$ and $\langle \mathbf{x} \rangle$ are noticeably different (AUC ≥ 0.89 in Figs. 4.2(d)-(f)). Note that PCA follows a similar trend, but with much less power likely due to the fact that we only consider the top 10 PCs of each protein trajectory in isolation for these analyses. One potential improvement for the PCA approach (at least for the larger perturbation scenarios) could be to first treat the original and perturbed structures as a single dataset and compute the top PCs over the concatenated protein

trajectories. We leave this exploration to the reader.

A slightly different intuition can be followed when looking at the results for the Elastic Net and Neural Network classifiers. When atomic positions are shifted equally by a constant Cartesian vector, the atoms in the ROI for the perturbed proteins become (in some cases) completely separable from those in the original structures. Therefore, an Elastic Net and Neural Network have no trouble assigning the true causal atoms non-zero effect sizes ($\text{AUC} \geq 0.85$ for both approaches in Figs. 4.2(a)-(c)). This observation is similar to previous works which show coordinate-based regularization to be most effective when variation between 3D structures occurs on a global scale and in the same direction on the unit sphere [139]. In the cases of random spherical perturbations, the variance of the distribution of atoms in the ROI widens; hence, the Elastic Net and Neural Network have a more difficult time identifying features that differentiate two protein classes, unless those variations happen on a global scale (again see Figs. 4.2(d)-(f)).

Most notably, SINATRA Pro performs consistently well in all simulation scenarios, identifying both static and dynamic differences better than most of the competing baselines that we considered ($\text{AUC} \geq 0.96$ in Fig. 4.2). Although SINATRA Pro is not as adept as the Elastic Net ($\text{AUC} = 1.00$ in Figs. 4.2(a)-(c)) at detecting static changes, it is able to robustly select significant features that are ignored by RMSF. In addition, SINATRA Pro is much better than the Elastic Net and Neural Network at identifying significant spherical perturbations that arise dynamically between protein structures. We hypothesize that summarizing atomic positions with Euler statistics is what enables SINATRA Pro to robustly capture both varying topology and geometry, unlike its coordinate-based counterparts, regardless of whether those differences occur in a constant or stochastic way.

4.2 Material and Methods

4.2.1 Molecular Dynamics Simulations

The protein structure data used in the current study are a result of molecular dynamic (MD) simulations. For large systems (i.e., IBB-bound Importin- β , unbound Importin- β) and those containing small-molecule ligands (i.e., wild-type HIV Protease, Ile50Val HIV protease, GTP-bound EF-Tu, and GDP-bound EF-Tu), we used Schrödinger’s *Desmond* (release 2020-1) [165] to run three independent 100 nanosecond (ns) simulations for each system. This decision is rooted in Desmond’s high

performance when dealing with hundreds of thousands of atoms, and the extensive validation of the small-molecule parameters contained in the OPLS3e force-field [166]. The systems were built within a dodecahedron box extending 1 nanometer (nm) beyond the solute in all three dimensions and solvated with water molecules using the SPC model [167]. Charges were neutralized by replacing a varying number of solvent molecules with sodium and chloride ions. Before production dynamics, all systems were relaxed and equilibrated with Desmond’s standard relaxation protocol, which first performs energy minimization with 50 kcal/mol/Å² restraints on the protein’s heavy atoms, followed by an extensive equilibration protocol. This protocol is detailed below:

1. NVT equilibration at 10 K for 12 ps
2. NPT equilibration at 10 K for 12 ps
3. NPT equilibration at 300K with harmonic restraints on the protein’s heavy atoms for 120 ps
4. NPT equilibration at 300 K, unrestrained, for 240 ps,

where NVT denotes constant temperature and volume and NPT denotes constant temperature and pressure. After equilibration, unrestrained NPT production simulations were conducted at 300 K and 1 atm for 100 ns for each system, to ensure consistency among the trajectories. Time steps for all simulations were set to their default values: 2:2:6 fs (bonded:near:far).

For comparatively small systems without ligands (i.e., wild-type TEM, Arg164Ser TEM, wild-type Abl1, and Met290Val Abl1), we used GROMACS (release 2018-2) [168] to run three independent 100 ns simulations for each system (150 ns for Abl1). Simulations were conducted with a 2 fs time step using the Amberff14SB force-field [39] and the TIP3P water model [167]. As with the Desmond simulations, the systems were built within a dodecahedron box and charges neutralized by replacing a number of solvent atoms with sodium and chloride ions. For each system, energy was minimized using a steepest-descent algorithm until the maximum force on any given atom was less than 1000 kJ/mol/min. Solvent atoms were equilibrated in sequential 0.5 ns NVT and NPT simulations with solute heavy atoms restrained by a spring constant of 1,000 kJ/mol/nm² using the LINCS algorithm [169]. After equilibration, production dynamics were conducted sans the position restraints. All simulations were conducted at 300 K and 1 atm. Lastly, using Visual Molecular Dynamics (VMD) (version 1.9.3) [170], we converted all trajectories employed in this study to a DCD file format and stripped solvent atoms to facilitate downstream computational analyses.

4.2.2 Protein Structure Alignment

In the current study, protein structures are aligned in one of two ways. In the first procedure, we align protein structures by minimizing the root-mean-square distance (RMSD) between the atoms on their backbone alpha-carbons (which we denote in shorthand by C_α). The first frame of the MD simulation is chosen as the reference structure. Next, all other frames (i.e., the mobile and fluctuating structures) in the dataset are aligned to this reference frame by (i) first superposing the center of mass of the C_α atoms to the same origin and then (ii) minimizing the RMSD rotation matrix. This calculation is performed using the `MDAnalysis` software package in `Python` (see Data and Software Availability) [132, 171–173]. For inter-class alignment when comparing protein class A to class B (e.g., mutants versus wild-type), all frames in the trajectory of class A and B are aligned to the first frame in the trajectory of class A . Note that, in our current study, there is only one point mutation in each protein. In this case, determining correspondences between sequences in the trajectories of class A and B is trivial because mutated residues in the mutant structures just correspond to the un-mutated residue in the wild-type (e.g., Arg164 in WT versus Ser164 in Arg164Ser mutant in β -lactamase). In the controlled simulation experiments, perturbed structures were obtained by directly modifying the atomic coordinates of the pre-aligned proteins; therefore, the perturbed structures do not need further alignment since their unperturbed regions remain aligned after the controlled modifications.

In the second alignment procedure, we perform sequence-independent structural alignment where we *implicitly* normalize the 3D protein structures by rotationally aligning their topological summary statistics. To carry out this alignment procedure, we first take each pair of protein structures and superpose the center of mass of the C_α atoms to the same origin. Next, we compute topological summary statistics over the mesh representation of each structure in $m = 500$ spherically uniformly distributed directions (see the remaining sections of Material and Methods). We take the squared Euclidean distance between any two directions to be the cost needed to align structures via their topological summaries; and we determine the “optimal” direction alignment by finding the rotation that minimizes the cumulative cost of aligning all directional pairs between proteins. Here, we use the random sample consensus (RANSAC) method to determine the rotational matrix that aligns the angle between any two directions to be within an error threshold of 0.9 [174]. More specifically, we require that the dot product between two directions has to be larger than 0.9 to be considered

aligned in RANSAC. Figures comparing the correlation between SINATRA Pro results on the five real protein systems using the RMSD pre-alignment versus using the topological rotational alignment can be found in the Supporting Information (Figs. B.25 and B.26). There are large agreements between both alignment procedures in almost all protein systems. The one exception being HIV-1 protease where we believe that the difference in results is driven by the fact that the small dynamic topology of protein makes it difficult to identify optimal stable reference signatures for alignment in summary statistic space.

4.2.3 Approximate Computation

In practice, we use a few approximations to scale the otherwise computationally expensive steps in Eq. 4.8. The first approximation involves computing the log determinant. With a dataset of reasonably dense meshes, the number of topological features is expected to be large (i.e., $J \gg 0$). In this setting, the term $-\log(|\Sigma_{-j}\Lambda_{-j}|) + \text{tr}(\Sigma_{-j}\Lambda_{-j}) + (1 - p)$ remains relatively equal for each feature j and makes a negligible contribution to the entire sum. Thus, we simplify Eq. 4.8 to

$$\text{KLD}(\beta_j) \approx \alpha_j(\beta_j - \mu_j)^2/2. \quad (4.9)$$

This approximation of the KLD still relies on the full precision matrix Λ . For a large number of topological features J , this calculation is expensive; however, it is only done once and can be done with efficient matrix decomposition. The rate of change parameter $\alpha_j = \Lambda_{-j}^\top \Lambda_{-j}^{-1} \Lambda_{-j}$, on the other hand, depends on the partitioned matrix Λ_{-j}^{-1} for every j -th topological feature. This requires inverting a $(J - 1) \times (J - 1)$ matrix J times. Fortunately, we can reduce this computational burden by taking advantage of the fact that any Λ_{-j}^{-1} is formed by removing the j -th row and column from the precision matrix Λ . Therefore, given the partition in Eq. 4.7, we can use the Sherman-Morrison formula [175] to efficiently approximate these quantities using the following rank-1 update for each topological feature

$$\Omega^{(j)} = \Lambda - \Lambda \sigma_j \sigma_j^\top \Lambda / (1 + \sigma_j^\top \Lambda \sigma_j) \quad j = 1, \dots, J. \quad (4.10)$$

Here, σ_j is the j -th column from the posterior covariance matrix Σ , and each Λ_{-j}^{-1} is approximated by removing the j -th row and column from $\Omega^{(j)}$. Ultimately, this reduces the computational complexity of Equation 4.9 to just J -independent $O(J^2)$ operations which can be parallelized.

4.2.4 Reconstruction and Visualization of Biophysical Signatures

After obtaining association measures γ for each topological feature computed via Eq. 4.6, in the fourth step of the SINATRA Pro pipeline, we map this information back onto the original structures to visualize topological differences between the protein classes. The main idea is that we want to select or prioritize atoms that correspond to the topological features with the greatest association measures. To do this, we perform a criterion-based *reconstruction* algorithm [139]. In each direction, each atom (i.e., vertex) lies along a filtration step that corresponds to a γ_j value. Therefore, each atom corresponds to $m = (c \times d)$ values in γ . To perform the reconstruction, we sort the values in γ from smallest to largest and continuously increase a threshold. If all of the γ values corresponding to an atom are larger than the threshold, the atom is considered “alive”. As the threshold is increased, when the criterion is no longer satisfied, the atom is considered “dead” and that minimum value below the threshold (which we will denote by $\hat{\gamma}$) is assigned the atom as its evidence score. This calculation is repeated for each frame in the dataset. For atomic-level evidence scores (e.g., Figs. B.5 and B.6), the $\hat{\gamma}$ values are ranked among all atoms and scaled from 0 (lowest) to 100 (highest) to facilitate the visualization and interpretation of structural and biophysical enrichment. To compute residue-level evidence scores (e.g., Fig. 4.3-4.6, and Figs. B.7, B.9, B.10, 4.5, B.12, B.14, B.16, and 4.7), we take the average of the $\hat{\gamma}$ values for all atoms within a residue which are then also ranked and scaled from 0 to 100.

4.2.5 Performance Assessment for Controlled Simulation Study

We demonstrate the power of the SINATRA Pro pipeline for identifying biophysical signatures in protein dynamics via multiple controlled simulations studies using the sequential procedure:

1. Fit the Gaussian process classification (GPC) model using elliptical slice sampling and compute relative centrality association measures γ_j for each j -th topological feature (i.e., differential Euler characteristic or DEC per sublevel set filtration). Recall, the total number of features $J = c \times d \times l$ is a product of (i) c , the number of cones of directions; (ii) d , the number of directions within each cone; and (iii) l , the number of sublevel sets (i.e., steps in the filtration) used to compute the DEC along a given direction.
2. Sort the topological features from largest to smallest according to their association measures

$$\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_p.$$

3. By iteratively moving through the sorted measures $T_k = \gamma_k$ (starting with $k = 1$), we reconstruct the atoms corresponding to the topological features with $\{j : \gamma_j \geq T_k\}$.

An atom is “detected” when the sublevel set in which it resides is selected across all of the directions within a particular cone. We form a union of the set of detected atoms across all cones to construct the set of reconstructed vertices at a given level T_k . Using this set of vertices, we compute the true positive rate (TPR) and false positive rate (FPR) by assessing overlap with the set of truly associated (i.e., perturbed) atoms used to generate the protein classes:

$$\text{TPR} = \frac{\sum \text{TP}}{\sum \text{P}}, \quad \text{FPR} = \frac{\sum \text{FP}}{\sum \text{N}} \quad (4.11)$$

where TP is the number of correctly detected true atoms, P is the total number of causal atoms, TN stands for the true negatives detected by the SINATRA Pro pipeline, and N stands for the total number of non-causal atoms. In this manner, we obtain a receiver operating characteristic (ROC) curve for the simulation studies (see Fig. 4.2 and Fig. B.1).

4.2.6 ROI Null Experiment and Statistical Assessment

To statistically assess whether SINATRA Pro is identifying the known regions of interest (ROI) in proteins by chance (see Table 4.2), we use a previously developed null-based scoring method [139]. The goal of this analysis is to estimate the probability of obtaining a result from SINATRA Pro under the assumption that the null hypothesis H_0 of there being no structural differences between mutant and wild-type proteins is true. Here, we treat the K atoms located within each ROI of every mutant protein as a landmark. We construct a test statistic τ^* for each ROI by summing the association metric scores of every atom it contains. To construct a “null” distribution and assess the strength of any score τ^* , we randomly select T “seed” atoms across the mesh outside the ROI for each mutant protein and uniformly generate T -“null” regions that are also K -atoms wide. We then compute similar (null) scores $\tau_1, \dots, \tau_T$ for each randomly generated region. A “ P -value”-like quantity (for the i -th mutant protein) is then generated by:

$$P_i = \frac{1}{T+1} \sum_{t=1}^T \mathbb{I}(\tau_i^* \leq \tau_t), \quad i = 1, \dots, N \quad (4.12)$$

where $\mathbb{I}(\cdot)$ is an indicator function, and a smaller P_i means more confidence in either method’s ability to find the desired paraconid landmark. To ensure the robustness of this analysis, we generate the

N -random null regions using a K -nearest neighbors (KNN) algorithm on each of the T -random seed vertices [176]. We also use a calibration formula to transform each P -value to an approximate Bayes factor (BF) [177], which is defined as the ratio of the marginal likelihood under the alternative hypothesis H_1 (i.e., that there is indeed a structural difference between phenotypic classes) versus the null hypothesis H_0 :

$$\text{BF}(P_i)_{10} = [-e P_i \log(P_i)]^{-1}, \quad i = 1, \dots, N \quad (4.13)$$

for $P_i < 1/e$ and $\text{BF}(P_i)_{10}$ is an estimate of $\Pr(H_1 | \mathcal{M}) / \Pr(H_0 | \mathcal{M})$, where $\mathcal{M}$ is again used to denote the protein meshes. We take the median of the P_i and $\text{BF}(P_i)_{10}$ values in Eqs. 4.12 and 4.13 across all mutant proteins, respectively, and report them in Table 4.3.

4.2.7 Data and Software Dependencies

Code for implementing the SINATRA Pro pipeline is freely available at <https://github.com/lcrawlab/SINATRA-Pro>, and is written in Python (version 3.6.9). As part of this procedure: (i) inference for the Gaussian process classification (GPC) model is based on an elliptical slice sampling algorithm adapted from the R package **FastGP** (version 1.2) [178] and (ii) the computation of nonlinear effect sizes and association measures for the differentiated Euler characteristic (DEC) curves was done by adapting the “RelATive cEntrality (RATE)” source code originally written in R (version 1.0.0; <https://github.com/lorinanthony/RATE>) [158]. Visualizing the reconstructed protein regions outputted by SINATRA Pro was done using the extensive molecular modeling system software **Chimera** (version 1.14) [179]. Molecular dynamic simulations were performed using Schrödinger’s **Desmond** (release 2020-1) [165] and **GROMACS** (release 2018-2) [168]. Furthermore, preprocessing steps for the protein structures resulting from MD simulations examined in the study were performed using **Visual Molecular Dynamics (VMD)** (version 1.9.3) [170] and the Python library **MDAnalysis** (version 1.1.1) [132, 171–173].

4.3 Results

4.3.1 Detecting Conformational Changes in Real Protein Systems

To examine SINATRA Pro’s ability to identify known structural changes of biological significance in real data, we consider the following five protein systems (Table 4.2): (1) the wild-type and

Arg164Ser mutant of TEM β -lactamase; (2) the wild-type and Ile50Val mutant of HIV-1 protease; (3) the guanosine triphosphate (GTP) and guanosine diphosphate (GDP) bound states of EF-Tu; (4) the wild-type and Met290Ala mutant of the Abl1 tyrosine protein kinase; and (5) unbound and IBB-bound states of importin- β . We choose to analyze these particular systems because they undergo varying degrees of conformational changes that have been well-studied in the literature (again see Table 4.2). Here, we will treat these previously identified features as ROIs, where the assumed “difficulty” for SINATRA Pro to statistically resolve structural signatures will be based on the stochasticity observed within each protein system. Namely, it will be more difficult to perform feature selection on structural ensembles that are highly dynamic as spurious fluctuations can interfere with detecting signal from the ROI. For each protein system, three MD trajectories were generated. From each of these trajectories, ten different series of structures are sampled, where each series is separated by 0.1 ns and each structure in a series is separated by 1 ns. SINATRA Pro is then implemented on the ten different series of structures extracted from the same trajectory data to mitigate the impact of variability between MD simulations (see Material and Methods).

Atomic enrichments are illustrated in Figs. B.5 and B.6, while residue-level structural enrichments are shown in Figs. 4.3-4.6 and Figs. B.7-B.19. To quantitatively assess the probability that SINATRA Pro is identifying any given ROI by chance, we implement a null region hypothesis test to estimate a P -value and an approximate Bayes factor (BF) corresponding to our power to reliably and robustly select certain features (Material and Methods). Reported results for the P -values and BF calculations are based on all MD simulated structures and can be found in Table 4.3. For comparison, we again implement the RMSF (Figs. 4.3-4.6 and Figs. 4.5, 4.7, and B.18-B.19) and Elastic Net (Figs. B.8, B.11, B.13, B.15, and B.17) baselines on all atomic positions with correspondences within these same protein systems. Here, we use scatter plots to illustrate the correlation between how each of these methods and SINATRA Pro rank the variable importance of the “pairable” atoms. All results presented in the main text reflect using SINATRA Pro with parameters set to $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search.

Note that we also provide additional figures assessing how robust SINATRA Pro is to different configurations of protein meshes for these data in the Supporting Information. For example, results analyzing SINATRA Pro with protein meshes constructed under different radius cutoffs r values can be found in Figs. B.20-B.24. To ensure sampling consistency of the 100 ns results, we performed sensitivity analyses based on protein structures drawn from different lengths up to 200 ns MD

simulation trajectories (Figs. B.7, B.10, B.12, B.14, and B.16). Lastly, we test the robustness of SINATRA Pro when the protein structures have been aligned in two different ways (Material and Methods). In the first, we align protein structures by minimizing the root-mean-square distance (RMSD) between their atoms (which we focus on in the main text); while, in the second, we perform sequence-independent structural alignment where we *implicitly* normalize the 3D protein structures by rotationally aligning their topological summary statistics (i.e., alignment without taking into correspondence between structures). Figures comparing the correlation between SINATRA Pro results on the RMSD pre-aligned protein structures versus using the topological rotational alignment can be found in Figs. B.25 and B.26.

4.3.2 Conformational Changes in the Active Site and Regulatory Ω -Loop of Arg164Ser TEM β -lactamase

Previous studies suggest that the Arg164Ser mutation in β -lactamase (TEM) induces structural changes in a highly plastic region known as the Ω -loop (residues 163-178), which plays a major role in the regulation of enzymatic activity [180, 181]. In wild-type β -lactamase, Arg164 makes a salt bridge with Asp179 that “pins down” the Ω -loop. Mutating Arg164 to serine breaks this salt bridge and disrupts a vast network of electrostatic and hydrogen interactions, dramatically affecting the dynamical behavior of the area surrounding the loop, parts of the active site, and potentially other protein domains [182]. These dynamical rearrangements confer multi-drug resistance to bacteria expressing TEM Arg164Ser, allowing them to hydrolyze a large number of cephalosporin antibiotics such as ceftazidime, cefixime, and cefazolin, in lieu of hydrolyzing ampicillin [183]. Given the enormous burden of multiresistant bacteria on public health, it is important that we understand the molecular mechanisms behind the structural rearrangements responsible for the transition to the cephalosporinase phenotype in order to orient future antibiotic design. Although previous studies have probed these rearrangements with varying approaches [182], the full mechanism remains elusive, highlighting the need for novel sampling and analytical methods that can detect the very slight changes in TEM’s active site topology that lead to drug resistance in the Arg164Ser and similar mutants. To help bridge this gap in understanding, we ran all-atom MD simulations of unbound TEM-1 and its Arg164Ser mutant, generated by homology modeling, and analyzed the results using SINATRA Pro, RMSF, and the Elastic Net baselines. Here, we expect SINATRA Pro to reveal

new insights about the molecular mechanisms underlying the specificity shift precipitated by the Arg164Ser mutation, due its ability to detect both minute static and stochastic changes in topology that elude traditional methods.

We compare the MD trajectories of wild-type and mutant TEM using aligned structures of the whole protein (Fig. 4.3(a) and Fig. B.5(a)), residues 65-230 (Fig. 4.3(b) and Fig. B.5(b)), and residues 65-213 (Fig. 4.3(c) and Fig. B.5(c)). In all three cases, statistical association measures from SINATRA Pro suggest that there are indeed significant structural changes in the Ω -loop (residues 163-178) relative to the rest of the regions in the protein (Fig. 4.3(d)), especially on residues 164 and 176-179, which are involved in the electrostatic interaction networks disrupted by the arginine to serine substitution. This ROI is not as prominently identified by the RMSF (see scatter plots in Fig. 4.3(e)-(g)) or the Elastic Net baselines (Fig. B.8(b)-(d)). Alternatively, all three approaches were able to identify the region harboring residues 213-230, which undergoes a noticeable dynamic shift over the course the MD trajectory. These results are consistent with our controlled simulations, which showed that the only time that RMSF and the Elastic Net both exhibit relatively decent power for stochastic changes is when large structural deviations are introduced (e.g., see power comparisons in Fig. 4.2(f)).

To more thoroughly assess if the Arg164Ser mutation contributes to the detected changes, we removed the Arg/Ser164 sidechain, as well as the whole residue (backbone and side chain), from our analyses. With the Arg/Ser164 atoms removed, association metrics of Arg/Ser164 and residues 176-179 diminished, which implies that signals pertaining to the dynamical contributions from the electrostatic interaction networks mediated by the side-chains of Arg164 and Glu179 are lost due to the missing topology. However, enrichment in the Ω -loop persisted, affirming that the identified topological differences are not just due to changes in these atoms. The null region test showed that the Ω -loop is indeed a robust significant structural feature in TEM, with $P = 5.63 \times 10^{-2}$ and BF = 2.27 when the whole TEM protein is analyzed (Table 4.3 with $r = 6.0$ Å), $P = 6.85 \times 10^{-2}$ and BF = 2.00 when residues 65-230 are analyzed, and $P = 7.22 \times 10^{-4}$ and BF = 70.4 when residues 65-213 are analyzed. We hypothesize that the ROI P -value is larger than the nominal 0.05 level for analyses with the whole structure and residues 65-230 because movement in the Ω -loop occurs jointly with moderate fluctuations in the region harboring residues 210-230. Overall, when we limit our scope to just residues 65-213, the region test robustly rejects the null hypothesis of the Ω -loop being identified by chance. It is worth noting that these results also highlight that there can

be variations in the strength of signal detected by SINATRA Pro depending on if one carries out analyses on whole protein structures versus on different fragments. These variations are most likely due to bias caused by different structural segments having different centers of mass and reference points during alignment. This will cause differences in the topological summary statistics that are computed for each structure and will lead to slightly varying signals in the biophysical signatures detected in downstream SINATRA Pro analyses.

Our results are particularly interesting for the TEM β -lactamase example because they highlight the importance of codon positions 164 and 179 in controlling Ω -loop dynamics, which contributes to modulating activity. Moreover, SINATRA Pro correctly captures the topological effects of the disruption of the electrostatic network formed by Arg164, Arg178, and Asp179 due to the Arg164Ser mutation. In addition to reaffirming previously observed phenomena, SINATRA Pro also identified meaningful shifts in the 210-230 segment in response to the resistance-granting Arg164Ser mutation. This suggests that the topology of the 210-230 segment, which forms the upper boundaries of the active site, is tightly correlated with shifts in the Ω -loop. Our results suggest an additional potential mechanism for activity modulation by Ω -loop fluctuations, where topological changes propagate from regulatory loops to parts of the active site, suggesting potential allosteric couplings between the Ω -loop and the 210-230 segment. These results function as a testament to SINATRA Pro’s capacity for distinguishing meaningful topological differences from the random fluctuations introduced by disorder-inducing mutations such as Arg164Ser, which obfuscates traditional analyses pipelines.

4.3.3 Changes in the Flap Region of HIV-1 Protease Driven by the Ile50Val Mutation

Our next analysis focuses on the HIV-1 protease, an enzyme that is essential for viral reproduction and is a well-established target for controlling HIV infections [184]. *In vivo*, the protease cleaves the HIV polyproteins Gag and Gag-Pol at multiple sites, creating the mature protein components of an HIV virion [185]. Over the past 25 years, ten HIV protease inhibitors have been approved for human use by the Food and Drug Administration (FDA), with many more undergoing clinical trials [186]. Similar to TEM, point mutations in the protease gene lead to products that are considerably less susceptible to inhibition by current drugs, generating drug-resistant HIV variants that pose a considerable risk [187]. Many hypotheses have been proposed for the molecular mechanisms

underlying the most common resistance-granting mutations, and recent studies have used sophisticated geometric analyses to classify conformational ensembles of mutant structures based on their influences on the dynamics [188]. Structurally, the HIV protease forms a homodimer with highly ordered domains [189]. Most resistance-granting mutations, such as the Ile50Val substitution, are thought to mainly affect the cross-correlated fluctuations of the flaps (residues 47-55), imparting minute changes to the fulcrum and lateral topology [188, 190, 191]. These findings suggest that mutations such as Ile50Val effectively rewire residue communication networks, significantly reducing its affinity for binding the inhibitor. These structural rearrangements lead to surprisingly nuanced changes to the topology, which as discussed previously, require refined quantitative methods to be detected. To test SINATRA Pro’s performance in detecting these small changes, we ran all-atom molecular dynamics simulations of “protein and ligand complex in water” systems containing HIV Protease or its Ile50Val mutant complexed with the antiviral drug Amprenavir [192]. We then followed that with analysis using either SINATRA Pro, RMSF, or the Elastic Net, with the objective of measuring each routine’s capacity for detecting and reporting the topological changes induced by the mutation.

Even though MD simulations are performed on the protein’s native dimeric form, chains A and B were separately selected and aligned before being input into each statistical method to avoid alignment bias due to inter-chain orientation. This focuses SINATRA, RMSF, and the Elastic Net on identifying the structural differences within each chain (e.g., Fig. 4.4(a)-(b) and Fig. B.5(d)-(e)). However, note that results for the analyses on the dimeric structures can be found in the Supporting Information (Fig. B.9). Overall, our analyses reveal that chains A and B seem to respond asymmetrically to the backbone effects of the mutation within the timeframe of the simulations (Table 4.3). This is not unexpected, as during the course of the simulations, the inhibitor Amprenavir affects dynamics asymmetrically by interacting more significantly with residues of Chain A. The change in RMSF for most of the residues in the flap are shown to be greater than 0.2 Å for chain A and smaller than 0.2 Å for chain B, indicating that the flap became more dynamic in the MD simulations when the Ile50Val mutation was introduced into chain A (Fig. 4.4(d)-(e) and Fig. B.18(b)). Meanwhile, the Elastic Net shows larger nonzero coefficients in the fulcrum for chain A than in chain B (Fig. B.11(a)).

While association metrics from SINATRA Pro identify structural changes in the flap for both chains (Fig. 4.4(c)), they also capture the geometric shifts within the fulcrum for chain A. We

hypothesize that the coexistence of the two changes (flap and fulcrum) in chain A contributes to a smaller peak (i.e., a weaker signal) in the association metrics produced by SINATRA Pro in the flap for chain A than in chain B. This asymmetry is confirmed by the null test, as topological changes in the flap appear to be less statistically significant in chain A ($P = 2.95 \times 10^{-1}$ and $\text{BF} = 1.022$) than in chain B ($P = 8.14 \times 10^{-4}$ and $\text{BF} = 63.554$) for this MD simulation data. Similar to β -lactamase, we assess if the Ile50Val mutation contributes to these detected topological changes. Upon removing the Ile/Val50 side-chain, the signal observed by SINATRA Pro in the flap drops with the missing topology, but still displays a significant peak relative to the rest of the protein, which implies that the change in association scores is not solely due to the structural differences upon introducing the Ile50Val mutation (Fig. 4.4(c)). Although SINATRA Pro clearly identified the effects of the isoleucine to valine substitution in flap topology and fulcrum dynamics, SINATRA Pro did not detect other previously elucidated structural signatures of the mutation, such as lateral extension [188]. As the baseline approaches also failed to identify these features, it is likely that their absence stems from sampling limitations inherent to the brute-force and relatively short production dynamics used to generate the conformational datasets.

The HIV protease system presents a good test case for SINATRA Pro due to its relative structural simplicity and the symmetry of the dimer. Encouragingly, SINATRA Pro’s results closely match those observed in previous studies that sought to characterize deltas in the backbone dynamics of the HIV protease in response to resistance-granting mutations [188].

4.3.4 Domain 2 in EF-Tu Undergoes Structural Changes upon GTP Hydrolysis

In our third analysis, we focus on EF-Tu (elongation factor thermo unstable), which is a G-protein that is responsible for catalyzing the binding of aminoacyl-tRNAs to the ribosome in prokaryotes. After binding GTP and a given aa-tRNA, EF-Tu strongly interacts with the ribosomal A site [193]. Following productive aa-tRNA binding, EF-Tu is released upon GTP hydrolysis [194]. The resultant GDP molecule is exchanged for GTP with EF-Ts (elongation factor thermo stable), allowing elongation to continue. Structurally, EF-Tu is composed of a Ras-like catalytic domain (RasD), common to G-proteins, and two beta-barrel domains (D2 and D3) [195]. Previous studies probing dynamic fluctuations of GTPases have identified that, after hydrolysis (in the GDP-bound state),

EF-Tu shows considerably increased flexibility of backbone atoms belonging to Domains 2 and 3, which are downstream of the nucleotide binding site in RasD [196]. These fluctuations are thought to be correlated with conformational rearrangements required for the exchange of GDP for GTP [196]. The conformational rearrangements are thought to occur on multiple millisecond timescales [196], presenting an obstacle for their study using all-atom molecular dynamics simulations. While the full relaxations associated with the change in ligand chemistry are challenging to sample effectively, we expect SINATRA Pro to have the ability to detect subtle topological differences arising from rearrangements of residue interacting networks in response to the chemical change in the ligand.

To compare our method’s performance to that of alternative techniques (Figs. 4.5 and B.13), we run SINATRA Pro, RMSF, and the Elastic Net on the whole structure (Figs. B.5(f) and 4.5(a)) and fragment windows limited to residues 208-308 (Figs. B.5(g) and 4.5(b)) and 311-405 (Figs. B.5(h) and 4.5(c)). Note that all figures displaying the enrichment of structural features are projected onto the GTP-bound structures. The evidence scores from SINATRA Pro reveal significant structural changes at Domain 2, with minimal structural changes in the majority of the Ras-like Domain, which agrees with findings in previous studies [196]. The null region test shows that Domain 2 is indeed an important structural feature in EF-Tu identified by SINATRA Pro with $P = 9.30 \times 10^{-4}$ and $BF = 56.657$, which robustly rejects the null hypothesis of the ROI being identified by chance (Table 4.3).

The chemical changes associated with the substitution of GTP with GDP in the EF-Tu system are thought to have significant impacts on backbone topology, making this a particularly interesting use case for SINATRA Pro. Despite the challenges associated with the considerable noise inherent to the complex EF-Tu system, SINATRA Pro succeeded in identifying the meaningful topological deltas that are thought to be important for function and that were elucidated in previous studies.

4.3.5 N-pocket Enlargement and α C Helix Displacement in Met290Ala Abl1

Protein tyrosine kinases (TKs) such as Abl1 and Src play significant roles in eukaryotic life, as phosphorylation of tyrosine residues in key proteins act as on/off switches that regulate a plethora of cellular processes and allow for efficient message passing [197]. Deregulation of the activity of these enzymes due to mutations is usually associated with severe forms of cancer and other chronic

diseases, posing a grave public health problem [198]. Due to their physiological importance, the enzymatic activity of tyrosine kinases is tightly regulated by a series of structural elements that fluctuate among metastable conformations between the active and inactive states [199]. This highly dynamic behavior has been exploited for the development of TK inhibitors, such as the widely-known anticancer drug Imatinib, which exclusively targets the “DFG out” state of Abl1 [199–201]. In this conformation, the phenylalanine residue of the region known as the DFG motif (comprised of Asp381, Phe382, and Gly383) occupies Abl1’s ATP binding site, preventing substrate binding and inactivating the enzyme [202]. Other TK inhibitors such as Dasatinib are capable of binding to Abl1’s “DFG-in” conformation, in which the positions of the aspartic acid and phenylalanine side-chains are inverted with respect to their positions in DFG-out conformations, activating the enzyme (i.e., making it capable of productive phosphorylation) [201]. The transition from the DFG-in to the DFG-out state is thought to happen on the multi-millisecond timescale, which presents a challenge for capturing it with unbiased atomistic MD simulations [202]. As a workaround, previous studies have used an engineered Abl1 mutant, Met290Ala, in which the energy barrier for the DFG flip is considerably reduced, as the steric effect presented by the bulky methionine is removed [128]. Although the sampling of the entire DFG flip is a rare event that happens in the millisecond timescale and thus outside of the scope of this work, we hypothesize that the Met290Ala mutation should induce minute topological changes around the DFG motif even in shorter simulations, due to the removal of the steric hindrances associated with the bulky methionine and that SINATRA Pro is capable of detecting these changes in contrast to the wild-type dynamics. To test this hypothesis and further measure our method’s capacity for detecting localized topological changes, we ran molecular dynamics simulations on the TK domain of the unbound state of Abl1 and its Met290Ala mutant. Specifically, we run SINATRA Pro, RMSF, and the Elastic Net on the whole structure (Fig. 4.6(a) and Fig. B.6(a)), fragments limited to residues 242-502 (Fig. 4.6(b) and Fig. B.6(b)), and the N-lobe spanning residues 242-315 (Fig. 4.6(c) and Fig. B.6(c)).

Since we only simulated the kinase domain of Abl1, both the N-terminal and C-terminal domains are shown to be highly dynamic as they are no longer stabilized by the mass of the entire protein. As a result, whole structural changes are overshadowed by large and noisy fluctuations, and competing methods (Elastic Net and RMSF) have a difficult time identifying enrichment in the DFG motif (Figs. B.15(a) and B.19(a)). Nonetheless, SINATRA Pro is able to identify the enrichment in the DFG motif regardless of the inclusion or exclusion of the N- and C-terminal (Figs. 4.6). The signal

in the DFG motif ROI becomes better statistically resolved when we remove some of the structural noise and concentrate on regions spanning residue fragments 242-502 and 242-315 (i.e., the N-lobe). The null region test results for SINATRA Pro show that the DFG motif is indeed an important structural feature in Abl1: $P = 8.86 \times 10^{-3}$ and BF= 8.783 for the whole structure analysis (i.e., including the termini) and $P = 2.50 \times 10^{-4}$ and BF= 178 for the analysis on residues 242-502 (i.e., excluding the termini), both of which reject the null hypothesis of the ROI being identified by chance (Table 4.3). In these analyses, the structural differences around the DFG motif between unbound Abl1 and its Met290Ala mutant were large enough for both RMSF and the Elastic Net to have power. As a comparison, SINATRA Pro not only robustly identifies residues associated with the greater N-pocket cleft as being statistically significant (i.e., the DFG motif), but also the α C helix spanning residues 281-293 as a moderately enriched region (Fig. 4.6).

From the SINATRA Pro output, we can postulate hypotheses regarding the involvement of specific codon positions outside of the DFG motif in the concerted motions that culminate in the flip, such as the two peaks of signal surrounding it (residues 350-360 and 390-400). These interesting results show that even short simulations can prove useful for gaining mechanistic insights regarding long-timescale macromolecular relaxations, as long as the heuristics employed to analyze the resulting trajectories are capable of detecting the often minute signals associated with these topological shifts.

4.3.6 Opening of Superhelix Differentiates Unbound and IBB-bound Importin- β

Our last analysis focused on the karyopherin Importin- β , an essential member of the nuclear import complex in eukaryotes, as it mediates the transportation of cargo from the cytosol to the nucleus [203]. Molecular recognition by Importin- β often requires the cooperative binding of molecular adaptors that recognize and bind to nuclear localization sequences (NLS)—structural motifs present in cargo destined for the nucleus [203]. Structurally, Importin- β is organized as a superhelix composed of up to 20 tandem HEAT repeats, each of which contain two antiparallel alpha helices linked by a turn [204]. This highly ordered structure is further stabilized by interactions with Importin- β -binding (IBB) domains of transport adaptors such Importin- α or Snurportin 1, which attach very strongly to Importin- β [205]. The release of IBB peptides after successful transport across the nuclear pore leads to large structural rearrangements and fluctuations that are propagated across most

of Importin- β 's backbone [206]. Although not difficult to detect with traditional analysis pipelines, such as calculating per-residue root-mean-square fluctuations or the backbone's radius of gyration, the pseudo-global nature of these rearrangements is diametrically opposite to most of the previously explored examples, presenting an important test for SINATRA Pro. Considering this, we ran MD simulations of unbound and IBB-bound Importin- β and, as with the previous examples, analyzed the resulting trajectories with SINATRA Pro to compare against standard methods. The structural features identified are projected onto the IBB-bound form (Figs. B.6(d) and 4.7(a)). Association metrics from SINATRA Pro, RMSF, and the Elastic Net all indicate large-scale conformational changes occur upon IBB release that involve the majority of the importin- β structure (Figs. 4.7(b)-(c) and B.17).

Since Importin- β functions as a molecular spring due to its supercoiled structure and extensive interactions with targets for transport, the sudden removal of the bound IBB domain to generate the unbound structure leads to extensive and drastic fluctuations across most of the backbone during production dynamics, originating from multiple highly-correlated nodes in each HEAT repeat. These drastic rearrangements translate to significant deltas in the topology and per-residue fluctuations that are readily detected by all tested methods. Importantly, the SINATRA Pro output replicates the expected results for the IBB bound/unbound Importin- β system, demonstrating the method's capacity for picking up relevant structural determinants not only for localized changes, but also for backbone-wide large-scale fluctuations.

4.4 Discussion

There is a growing library of computational methods that leverage geometry and topology to study aspects of both protein structure and function. For example, recent work has used shape retrieval techniques from computer vision to identify distant protein homologs [207, 208]; while others are actively studying the utility of topology-based and tessellation-based protein representations in various deep learning tasks for accurate protein structure prediction [209]. In this paper, we introduced SINATRA Pro: a topological data analytic approach designed to extract biologically-relevant structural differences between two protein ensembles. Through an extensive benchmark simulation study, we assessed the utility and statistical properties of SINATRA Pro against commonly used methods in the field. Here, we showed that our proposed framework can robustly identify both static

and dynamic structural changes that occur between protein ensembles. We also highlighted that, unlike other standard approaches in the field, SINATRA Pro does not require atom-by-atom correspondences between structures and thus can be implemented using all atomic information that is available, rather than being limited to atomic features that are conserved over side-chain substitutions. With real MD data, we used SINATRA Pro to analyze five different protein systems and demonstrated its ability to identify known regions of interest that have been validated by previous experimental and computational studies, as well as reveal novel structural features that are also likely to be biologically-relevant according to evidence in the literature. Overall, these results show the promise of SINATRA Pro as a hypothesis generation tool that practitioners can use to design more informed experiments for answering downstream scientific questions (e.g., whether a mutation or chemical change “induces” a specific structural change).

There are many potential extensions to the SINATRA Pro pipeline. First, in its current form, SINATRA Pro treats all atomic features as being equally important *a priori* to the phenotype of interest. One particularly interesting extension of the method would be to up- or down-weight the contributions of different types of atomic features (e.g., carbons, hydrogens, or oxygens) or residues (e.g., serine versus arginine) to more accurately represent the topology of specific inter-atomic connections such as hydrogen and covalent bonds. In practice, this would require making such annotations and deriving topological summary statistics of protein structures based on a weighted Euler characteristic transform [210]. Another natural extension would be to apply the SINATRA Pro pipeline to other data types used to study variation in 3D protein structures such as cryogenic electron microscopy (cryo-EM), nuclear magnetic resonance (NMR) ensembles, and X-ray crystallography (i.e., electron density) data. Previous work has already shown that topological characteristics computed on tumors from magnetic resonance images (MRIs) have the potential to be powerful predictors of survival times for patients with glioblastoma multiforme (GBM) [143, 210] and other cancer subtypes [145, 211, 212]; however, it has also been noted that the efficacy of current topological summaries decreases when heterogeneity between two phenotypic classes is driven by minute differences [139]. For example, cryo-EM images can look quick similar even for two proteins harboring different mutations. SINATRA Pro’s improved ability to capture inter-class variation is driven by local fluctuations in shape morphology, so it would be interesting to see if our proposed pipeline could offer more resolved insights for these types of applications.

4.5 Tables and Figures

Notation	Description	Range	General Guidelines
r	Radius cutoff (Å) for simplicial reconstruction	$[0, \infty)$	Use smaller $r \leq 2.0$ Å for rigid proteins and $r \in [2.0 \text{ Å}, 6.0 \text{ Å}]$ for flexible proteins.
c	Number of cones of directions	$[1, \infty)$	Set much greater than 1 as more power is generally achieved by taking filtrations over multiple directions
d	Number of directions per cone	$[1, \infty)$	Set much greater than 1 as more power is generally achieved by taking filtrations over multiple directions
θ	Cap radius used to generate directions within a cone	$(0, 2\pi]$	Set between $[0.1, 0.8]$ since cones should be defined by directions in close proximity
l	Number of sublevel sets (filtration steps)	$[1, \infty)$	Optimal choice depends on the size of protein molecule being analyzed so use grid search

Table 4.1: General guidelines for choosing values for the free parameters in the SINATRA Pro pipeline software. The guidelines provided are based off of intuition gained through the simulation studies provided in the main text and Supporting Information. In practice, we suggest specifying multiple cones $c > 1$ and utilizing multiple directions d per cone (see monotonically increasing power in Fig. B.1 in Supporting Information). While the other two parameters (θ and l) do not have monotonic properties, their effects on SINATRA’s performance still have natural interpretations. Selection of $\theta \in [0.1, 0.8]$ supports previous theoretical results that cones should be defined by directions in close proximity to each other [139, 141]; but not so close that they explain the same local information with little variation. Note that our sensitivity analyses suggest that the power of SINATRA Pro is relatively robust to the choice of θ . Optimal choice of l depends on the size of the protein molecules that are being analyzed. Intuitively, for rigid proteins, coarse filtrations with too few sublevel sets cause SINATRA Pro to miss or “step over” structural shifts that occur locally during the course of a molecular dynamic (MD) trajectory. In practice, we recommend choosing the angle between directions within cones θ and the number of sublevel sets l via cross validation or some grid-based search.

Protein	PDB ID	Chemical Change	Structural Signature	Difficulty	Ref(s)
β -lactamase (TEM)	1BTL	Arg164Ser	Increased dynamics of Ω -Loop (Residues 163-178)	Easy	[180, 181]
HIV-1 Protease	3NU3	Ile50Val	Reduced stability in the flaps (Residues 47-55)	Medium	[188, 190, 191]
EF-Tu	1TTT	GTP Hydrolysis	Increased flexibility of Domain 2 (Residues 208-308)	Easy	[196, 213, 214]
Abl1	3KFA	Met290Ala	Fluctuations in the DFG motif and displacement of helix α C	Hard	[128, 202, 215–217]
Importin- β	2P8Q	IBB Release	Uncoiling in the conformation of the superhelix	Hard	[204–206]

Table 4.2: Detailed overview of the different protein systems analyzed in this study. The columns of this table are arranged as follows: (1) the name of each protein studied; (2) the corresponding Protein Data Bank (PDB) ID for each molecule [218]; (3) the known chemical change or mutation type that is considered; (4) the specific structural signatures that are known to be associated with each chemical change or mutation type; (5) the presumed difficulty level for SINATRA Pro to detect each structural signature based on the homogeneity in shape variation between the wild-type and mutant proteins; and (6) references that have previously suggested some level of association or enrichment between each structural change and the mutation of interest.

Protein	ROI	Fragment	$r = 2 \text{ \AA}$		$r = 4 \text{ \AA}$		$r = 6 \text{ \AA}$	
			P -value	Bayes Factor	P -value	Bayes Factor	P -value	Bayes Factor
TEM	Ω -Loop	Whole	5.95×10^{-1}	—	3.35×10^{-4}	137.121	5.63×10^{-2}	2.270
		65-230	1.20×10^{-1}	1.447	4.16×10^{-2}	2.783	6.85×10^{-2}	2.004
		65-213	7.22×10^{-4}	70.438	7.22×10^{-4}	70.438	7.22×10^{-4}	70.438
HIV-1	Flap	Chain A	2.33×10^{-1}	1.084	4.03×10^{-2}	2.841	2.95×10^{-1}	1.022
		Chain B	8.14×10^{-4}	63.554	8.14×10^{-4}	63.554	8.14×10^{-4}	63.554
EF-Tu	Domain 2	Whole	9.30×10^{-4}	56.657	9.30×10^{-4}	56.657	9.30×10^{-4}	56.657
Abl1	DFG Motif	Whole	1.94×10^{-1}	1.157	5.38×10^{-1}	—	8.86×10^{-3}	8.783
		242-502	1.54×10^{-1}	1.279	2.50×10^{-4}	177.614	2.50×10^{-4}	177.614

Table 4.3: Null hypothesis experiment to evaluate SINATRA Pro’s ability to find regions of interest (ROI) in each of the proteins analyzed in this study. Here, we assess how likely it is that SINATRA Pro finds the region of interest (ROI) by chance. These ROIs include: (i) the Ω -loop (residues 163-178) in TEM; (ii) the flap region (residues 47-55) in HIV-1 protease; (iii) Domain 2 (residues 208-308) in EF-Tu; and (iv) the DFG motif (residues 381-383) in Abl1. Note that protein structures were only analyzed if they contained an entire ROI. For example, in the context of Importin- β , the superhelix includes the entire structure and so we do not conduct a null analysis. In this experiment, to produce the results above, we generate “null” regions on each protein using a K -nearest neighbors (KNN) algorithm on different atoms as random seeds [176], and exclude any generated regions that overlap with the ROI. Next, for each region, we sum the association metrics of all its atoms. We compare how many times the aggregate scores for the ROI are higher than those for the null regions. These “ P -values,” and their corresponding calibrated Bayes factors (BF) when the computed $P < 1/e$, are provided above. Note that P -values less than the nominal size 0.05 and BFs greater than 2.456 are in bold. Results above are based on SINATRA Pro using parameters $\{c = 20, d = 8, \theta = 0.80, l = 120\}$ while varying the radius cutoff parameter r for mesh construction on each protein structure.

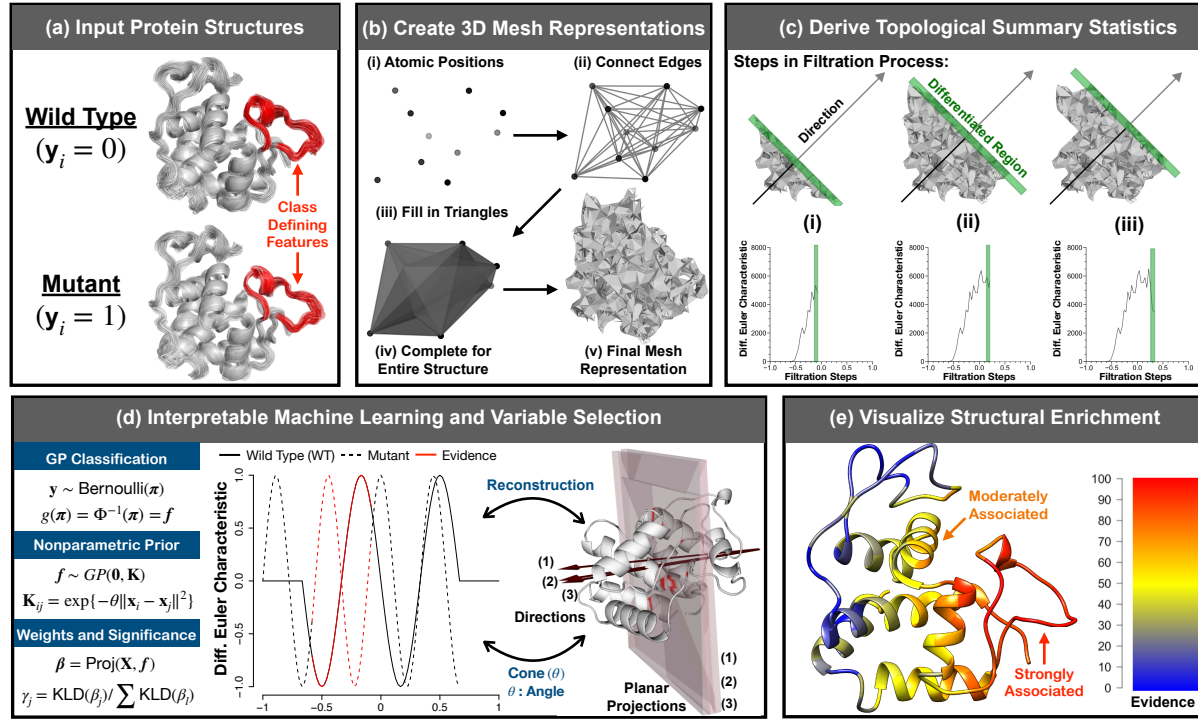


Figure 4.1: Schematic overview of SINATRA Pro: a novel framework for discovering biophysical signatures that differentiate classes of proteins. **(a)** The SINATRA Pro algorithm requires the following inputs: (i) (x, y, z) -coordinates corresponding to the structural position of each atom in every protein; (ii) y , a binary vector denoting protein class or phenotype (e.g., mutant versus wild-type); (iii) r , the cutoff distance for simplicial construction (i.e., constructing the mesh representation for every protein); (iv) c , the number of cones of directions; (v) d , the number of directions within each cone; (vi) θ , the cap radius used to generate directions in a cone; and (vii) l , the number of sublevel sets (i.e., filtration steps) used to compute the differential Euler characteristic (DEC) curve along a given direction. Guidelines for how to choose the free parameters are given in Table 4.1. **(b)** Using the atomic positions for each protein, we create mesh representations of their 3D structures. First, we draw an edge between any two atoms if the Euclidean distance between them is smaller than some value r , namely $\text{dist}[(x_1, y_1, z_1), (x_2, y_2, z_2)] < r$. Next, we fill in all of the triangles (or faces) formed by these connected edges. We treat the resulting triangulated mesh as a simplicial complex with which we can perform topological data analysis. **(c)** We select initial positions uniformly on a unit sphere. Then for each position, we generate a cone of d directions within angle θ using Rodrigues’ rotation formula [219], resulting in a total of $m = c \times d$ directions. For each direction, we compute DEC curves with l sublevel sets. We concatenate the DEC curves along all the directions for each protein to form vectors of topological features of length $J = l \times m$. Thus, for a study with N -proteins, an $N \times J$ design matrix is statistically analyzed using a Gaussian process classification model. **(d)** Evidence of association measures for each topological feature vector are determined using relative centrality measures. We reconstruct corresponding protein structures by identifying the atoms on the shape that correspond to “statistically associated” topological features. **(e)** The reconstruction enables us to visualize the enrichment of biophysical signatures that best explain the variance between the two classes of proteins. The heatmaps display atomic (or residue-level, which we define as a collection of atoms) evidence potential on a scale from $[0 - 100]$, with a score of 100 meaning most enriched.

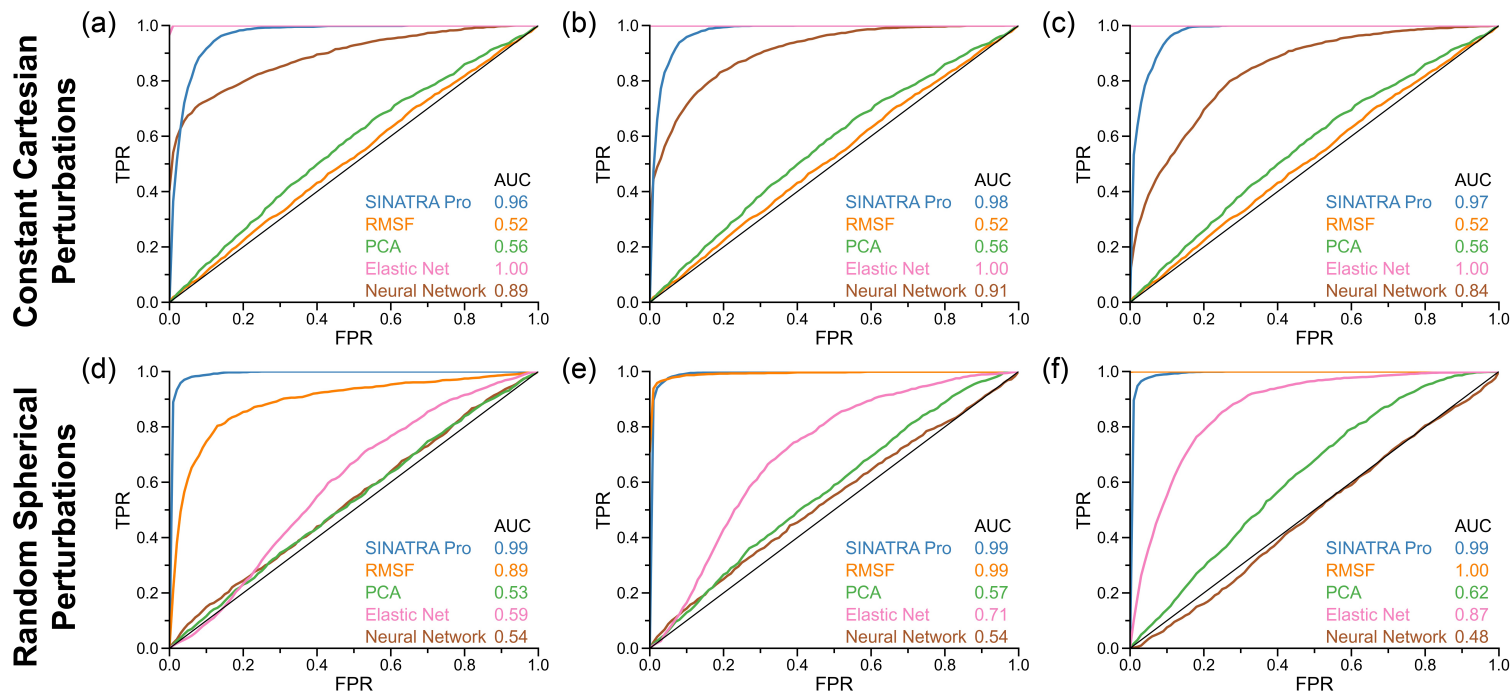


Figure 4.2: Receiver operating characteristic (ROC) curves comparing the power and robustness of SINATRA Pro to competing 3D mapping approaches in controlled molecular dynamic (MD) simulations. To generate data for these simulations, we consider two phenotypic classes using the real structural data of wild-type β -lactamase (TEM). In the first phenotypic class, structural protein data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). In the second phenotypic class, proteins are drawn at 1 ns intervals shifted 0.5 ns with respect to the first set (e.g., $t_{\text{MD}} = [0.5, 1.5, 2.5, 3.5, \dots, 99.5] \text{ ns} + \delta$) to introduce thermal noise, and then we displace the atomic positions of each atom in the Ω -loop region by **(top row)** a constant Cartesian vector of **(a)** 0.5 Ångströms (Å), **(b)** 1.0 Å, and **(c)** 2.0 Å, or **(bottom row)** by a spherically uniform random vector of **(d)** 0.5 Å, **(e)** 1.0 Å, and **(f)** 2.0 Å. Altogether, we have a dataset of $N = 1000$ proteins per simulation scenario: 100 ns interval $\times$ 5 different choices $\delta = \{0.0, 0.1, 0.2, 0.3, 0.4\} \text{ ns} \times 2$ phenotypic classes (original wild-type versus perturbed). The ROC curves and corresponding area under the curves (AUC) depict the ability of SINATRA Pro to identify “true class defining” atoms located within the Ω -loop region using parameters $\{r = 1.0 \text{ Å}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. We compare SINATRA Pro to four methods: root mean square fluctuation (RMSF) (orange); principal component analysis (PCA) (green); Elastic Net classification (pink); and a Neural Network (brown). For details on these approaches, see Materials and Methods. A sensitivity analysis exploring the optimal parameter configurations for PCA and Neural Network can be found in Supporting Information (see Figs. B.3 and B.4, respectively).

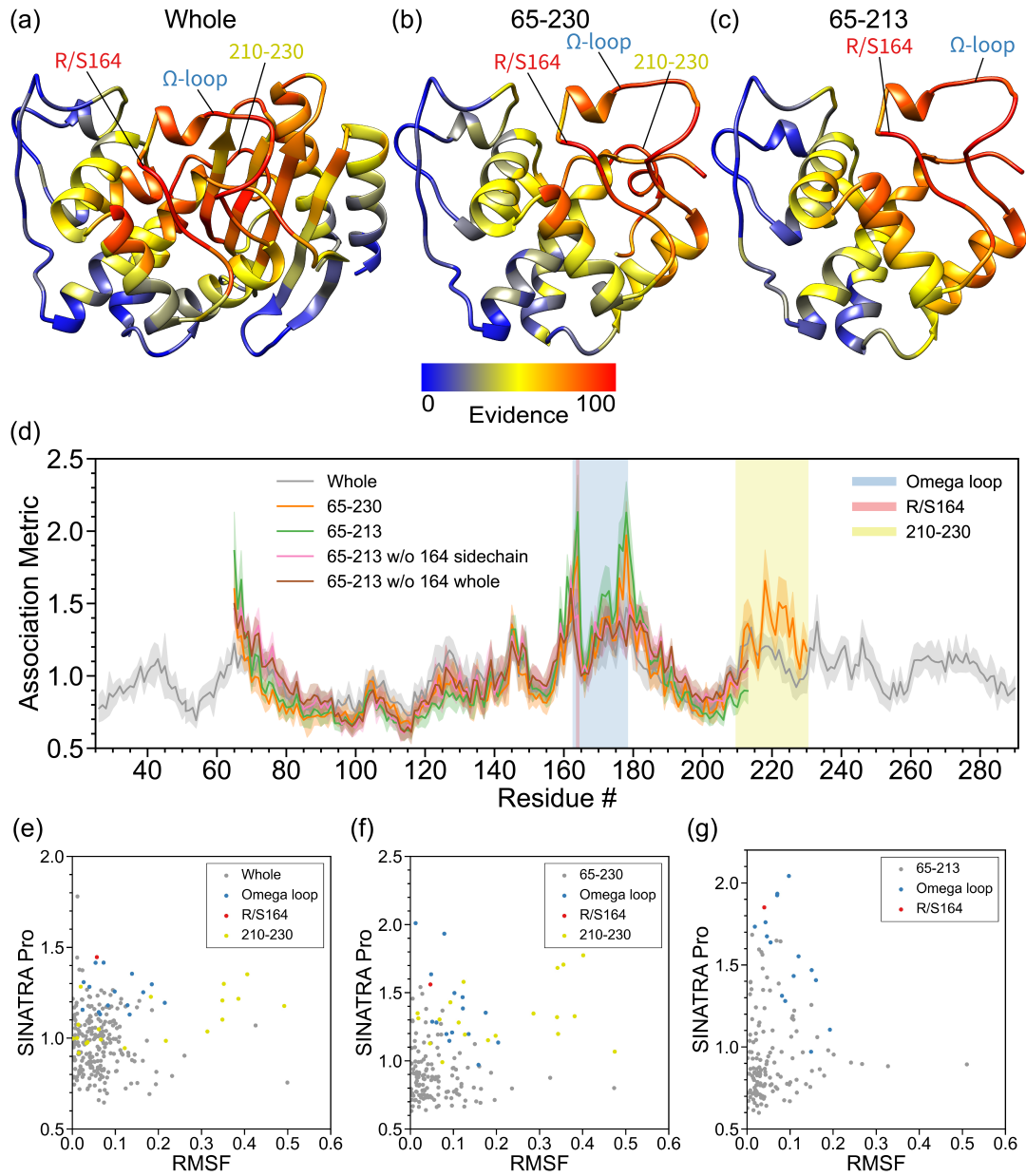


Figure 4.3: Real data analyses aimed at detecting structural changes in the Ω -loop of β -lactamase (TEM) induced by an R164S mutation. In this analysis, we compare the molecular dynamic (MD) trajectories of wild-type β -lactamase (TEM) versus R164S mutants [180, 181]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ protein structures in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). This figure depicts results after applying SINATRA Pro using parameters $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. The heatmaps in panels (a)-(c) highlight residue evidence potential on a scale from [0 – 100]. A maximum of 100 represents the threshold at which the first residue of the protein is reconstructed, while 0 denotes the threshold when the last residue is reconstructed. Panel (a) shows the residue-level evidence potential when applying SINATRA Pro to the whole protein, while panels (b) and (c) illustrate results when strictly applying the SINATRA Pro pipeline to atoms in residues 65-230 and 65-213, respectively. Annotated regions of interest are color coded and correspond to the shaded residue windows in panel (d). Panel (d) shows the mean association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods). Here, the overlap shows the robustness of SINATRA Pro to identify the same signal even when it does not have access to the full structure of the protein. The final row plots the correlation between the SINATRA Pro association metrics and the root mean square fluctuation (RMSF) for all atoms with correspondences in the (e) whole protein, (f) fragment 65-230, and (g) fragment 65-213.

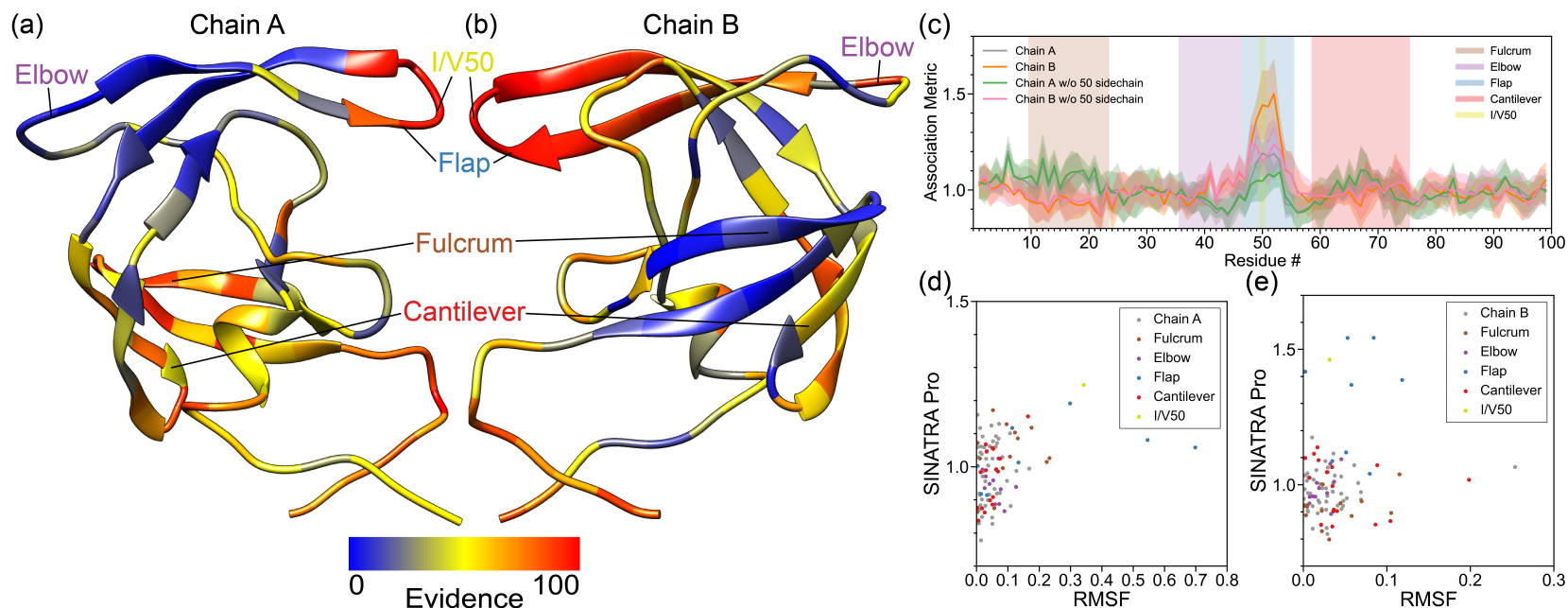


Figure 4.4: Real data analyses recover structural changes in the flap region of HIV-1 protease driven by a Ile50Val mutation. In this analysis, we compare the molecular dynamic (MD) trajectories of wild-type HIV-1 protease versus Ile50Val mutants (i.e., within residues 47-55). For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). This figure depicts results after applying SINATRA Pro using parameters $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. The heatmaps in panels (a) and (b) highlight residue evidence potential on a scale from [0 – 100]. A maximum of 100 represents the threshold at which the first residue of the protein is reconstructed, while 0 denotes the threshold when the last residue is reconstructed. Panel (a) shows residue-level evidence potential when applying SINATRA Pro to chain A, while panel (b) depicts results for chain B. Annotated regions of interest are color coded and correspond to the shaded residue windows in panel (c). Panel (c) shows the association metrics (and their corresponding standard errors) computed for each residue in chains A and B, with and without the 50th residue’s side chain being included in the analysis (see Material and Methods). Here, the overlap shows the robustness of SINATRA Pro for identifying the same signal even when it does not have access to the full structure of the protein. The final row plots the correlation between the SINATRA Pro association metrics and the root mean square fluctuation (RMSF) for all atoms with correspondences in (d) chain A and (e) chain B, respectively. Highlighted are all atoms with correspondences found in regions of the protein corresponding to the fulcrum (brown), elbow (purple), flap (blue), cantilever (red), and I/V50 (yellow) [188, 190, 191].

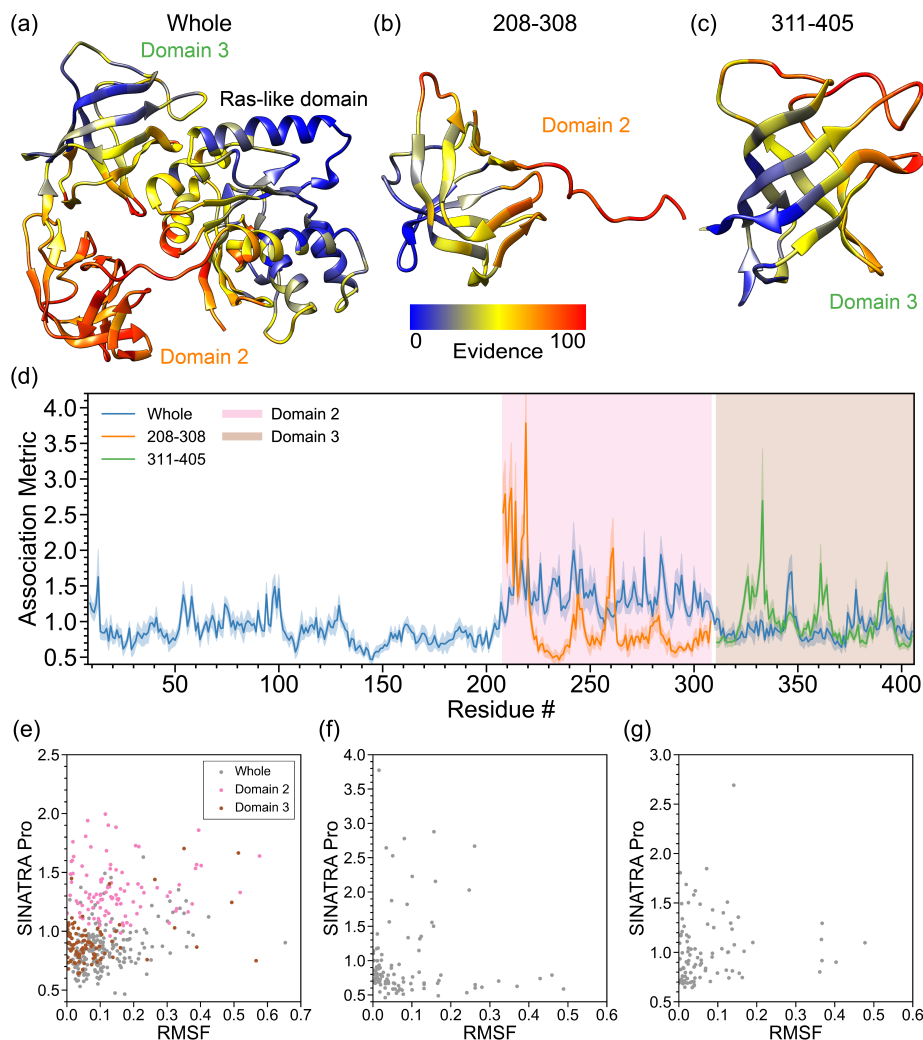


Figure 4.5: Real data analyses aimed at detecting structural changes in Domain 2 of the elongation factor EF-Tu upon guanosine triphosphate (GTP) hydrolysis. In this analysis, we compare the molecular dynamics (MD) trajectories of GTP-bound EF-Tu versus GDP-bound EF-Tu [196, 213, 214]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). This figure depicts results after applying SINATRA Pro using parameters $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. The heatmaps in panels (a)-(c) highlight residue evidence potential on a scale from [0 – 100]. A maximum of 100 represents the threshold at which the first residue of the protein is reconstructed, while 0 denotes the threshold when the last residue is reconstructed. Panel (a) shows residue-level evidence potential when applying SINATRA Pro to the whole protein, while panels (b) and (c) illustrate results when strictly applying the SINATRA Pro pipeline to atoms in residues 208-308 and 311-405, respectively. Panel (d) shows the association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods). Here, the overlap shows the robustness of SINATRA Pro to identify the same signal even when it does not have access to the full structure of the protein. The final row plots the correlation between the SINATRA Pro association metrics and the root mean square fluctuation (RMSF) for all atoms with correspondences in the (e) whole protein, (f) fragment 208-308, and (g) fragment 311-405.

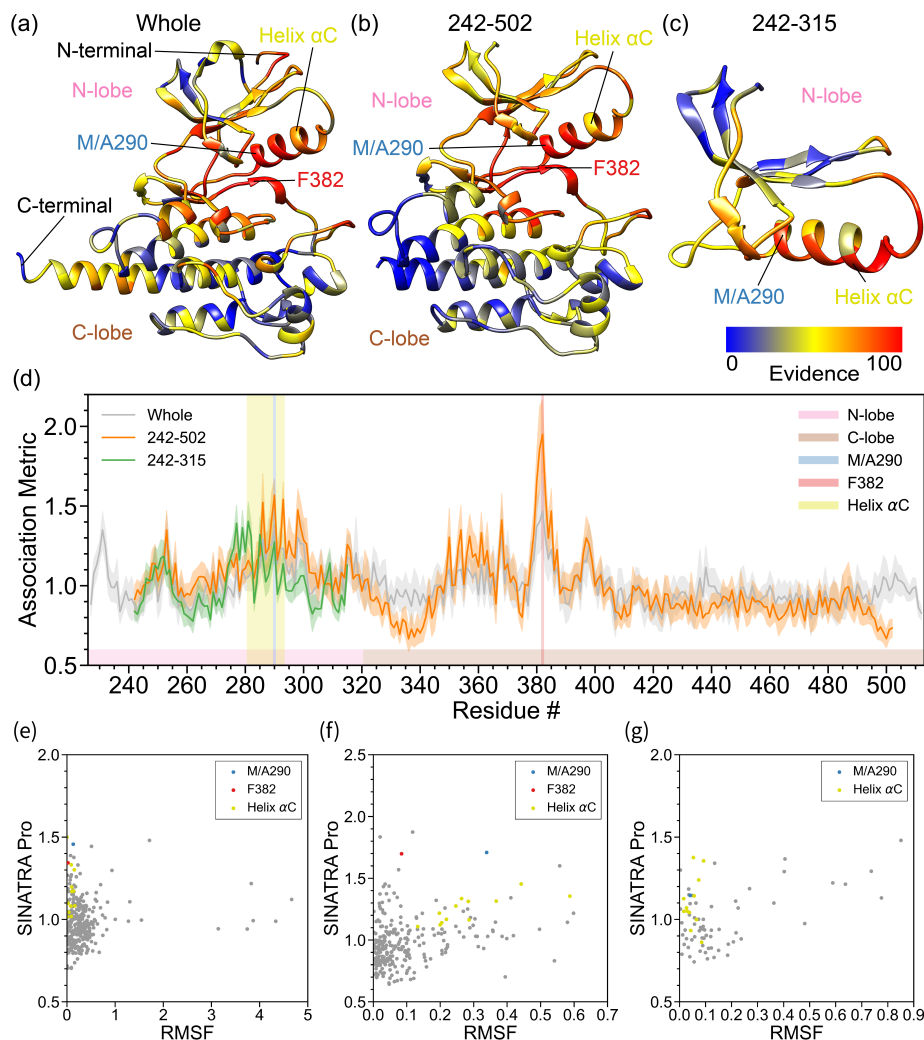


Figure 4.6: Real data analyses identify enrichment in the N-terminal pocket of the Abl1 Tyrosine protein kinase due to a M290A mutation in the α C helix. In this analysis, we compare the molecular dynamic (MD) trajectories of wild-type Abl1 versus M290A mutants [128, 202, 215–217]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 150 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \times 1.5 \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 3000$ proteins in the study: 150 ns long interval $\times$ 15 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 1.4\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). This figure depicts results after applying SINATRA Pro using parameters $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. The heatmaps in panels (a)–(c) highlight residue evidence potential on a scale from [0 – 100]. A maximum of 100 represents the threshold at which the first residue of the protein is reconstructed, while 0 denotes the threshold when the last residue is reconstructed. Panel (a) shows residue-level evidence potential when applying SINATRA Pro to the whole protein, while panels (b) and (c) illustrate results when strictly applying the SINATRA Pro pipeline to atoms in residues 242–502 and 242–315, respectively. Annotated regions of interest are color coded and correspond to the shaded residue windows in panel (d). Panel (d) shows the association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods). Here, the overlap shows the robustness of SINATRA Pro to identify the same signal even when it does not have access to the full structure of the protein. The final row plots the correlation between the SINATRA Pro association metrics and the root mean square fluctuation (RMSF) for all atoms with correspondences in the (e) whole protein, (f) fragment 242–502, and (g) fragment 242–315.

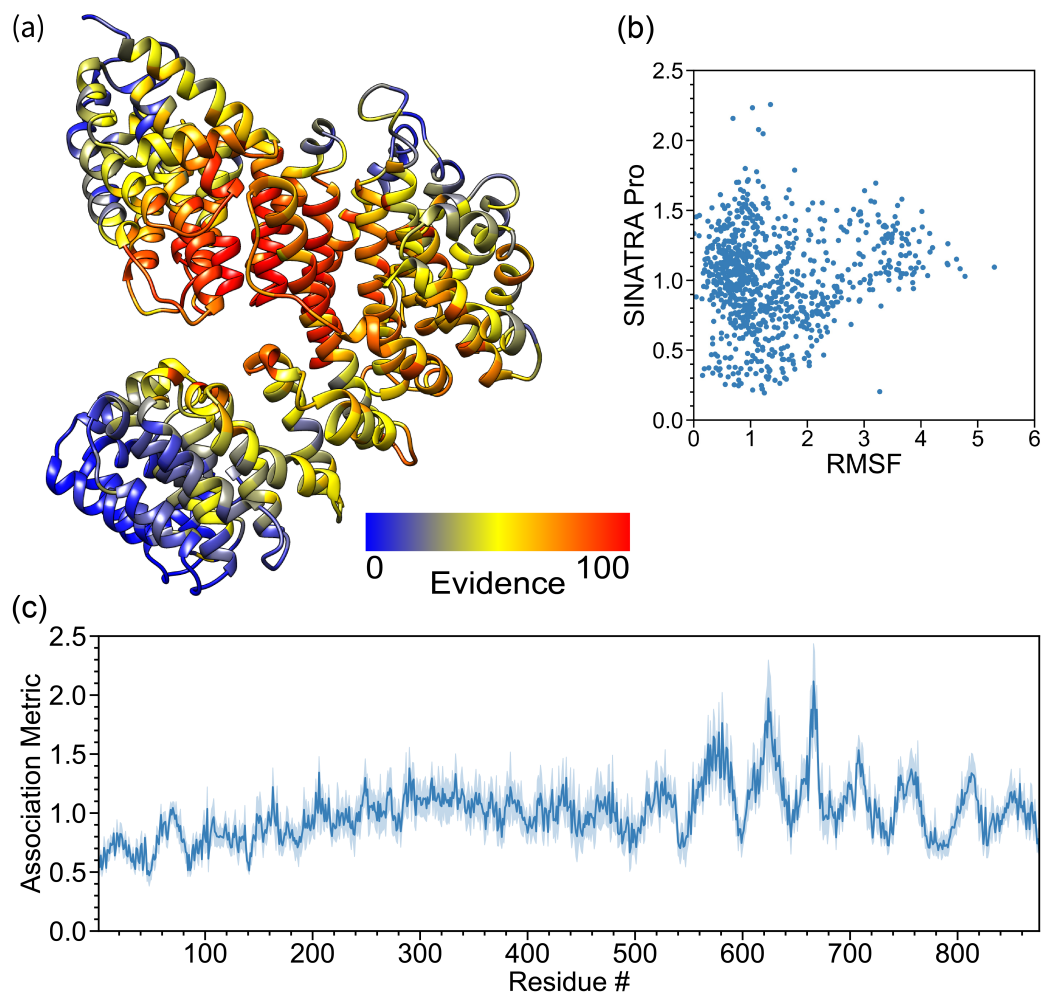


Figure 4.7: Real data analyses aimed at detecting uncoiling of the superhelix in Importin- β upon release of an IBB peptide. In this analysis, we compare the molecular dynamics (MD) trajectories of IBB-bound Importin- β versus unbound Importin- β [204–206]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). This figure depicts results after applying SINATRA Pro using parameters $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. The heatmap in panels (a) highlights residue evidence potential on a scale from $[0 - 100]$. A maximum of 100 represents the threshold at which the first residue of the protein is reconstructed, while 0 denotes the threshold when the last residue is reconstructed. Panel (b) plots the correlation between the SINATRA Pro association metrics and the root mean square fluctuation (RMSF) for all atoms with correspondences. Panel (c) shows the SINATRA Pro association metrics (and their corresponding standard errors) computed for each residue within the analysis (see Material and Methods in the main text for more details).

Chapter 5

Future directions and Preliminary results

5.1 A Topological Data Analytic Approach for Structural alignment for protein structural ensembles and distant homologs

5.1.1 Introduction

Motivated by the SINATRA Pro project, in this study, I attempted to apply similar topological data analytic approach to align biological structures, specifically protein structures, using topological summary statistics. While I also aim to generalize this approach to other biological structures such as bone, tissue, and tumor, here, I focus on our effort to optimize such strategy for protein structural alignment, which is a fundamental problem in computational structural biology and has broad applications to protein sequence, structure and functional study [220]. Pairwise structure alignment could be applied to measure the similarity between two proteins, which is the first step for homology search and fold space construction. Meanwhile, the multiple structure alignment informs about evolutionary conservation and divergence from a family of protein structures [221]. A breakthrough in the computational method for protein structure prediction at near-experimental

quality by AlphaFold2 [222] has expanded the database of protein structures to an unprecedented scale, leading to a surge in the necessity for structural alignment methods that can efficiently and accurately search through these protein structures and identify biologically meaningful structural similarity.

In the past three decades, many structural alignment algorithms have been developed for pair-wise alignment [223–233] and multiple alignment [234–242]. However, the alignment accuracy and computational efficiency of these programs are still low when judged by manually-curated structure alignments, especially on structurally similar but evolutionarily distant proteins [243]. Recently, novel approaches in the structural alignment algorithms have been developed to generate alignments that are biologically meaningful from the evolutionary perspective [244], or to search through large protein structural datasets with state-of-the-art accuracy while being computationally incredibly efficient [245].

In this section, I present a novel approach to the protein alignment problem by extending the idea of topological data analysis. Using the Euler-characteristic-based topological summary statistics, I quantify the similarity among local and global topology of the protein backbone structures. I then use this information to identify relevant structural components for comprehensive and precise alignments. Experimental results indicate that the proposed approach, I named the “EC Align”, recreates manual alignments in human-curated protein distant homolog dataset MALISAM [246].

5.1.2 Alignment scheme using directional DEC curves

To accomplish rotational alignment between the protein structures, I quantitatively describe the topology of the protein structures using the directional DECT and align the directions with similar DEC curves, which infer similar directional topological profiles. Fig. 5.1 shows the alignment scheme using DECT. As the first step of the scheme, I first take each pair of protein structures and superpose the center of mass of the C_α atoms to the same origin. The input protein structures are converted in the simplicial complexes by the procedure as described in Chapter 4. However, I only construct the vertices and edges for the simplicial complexes and focus on aligning the backbone atoms (C_α , C, O, and N atoms). These modifications allow the resulting simplicial complexes to include sufficient topological information for alignment while significantly reducing the computational cost.

Next, I compute topological summary statistics over the mesh representation of each structure in spherically uniformly distributed directions. The topology of the complexes in various directions

(uniformly distributed over a sphere) is summarized by DECT as described in Eq. 4.2. These DEC curves describe the topological profile of the protein structure in each of the $m = 500$ directions.

To quantify the topological similarity between the directions, I calculate the Euclidean distance between the DEC curves to identify pairs of directions with similar topological profiles between the two structures. Directions with similar topological profiles have small Euclidean distances and vice versa. I incorporated a Gaussian kernel of width $0.5 \text{ \AA}$ to smooth out the DEC curves and the distance metric between the DEC curves and to emulate the electronic volume of the atoms. Then the algorithm minimizes the linear cumulative cost, i.e., the sum of Euclidean distances for all pairs of directions between the two protein structures, to assign directional correspondence. Directions with smaller Euclidean distance, i.e., similar topological profiles, are more likely to be paired up.

After assigning the directional correspondences, I use the random sample consensus (RANSAC) [174] to identify the rotational matrix that “optimally” aligns most directional correspondences. RANSAC method [174] iteratively estimates a mathematical model from a set of observed data that contains outliers while minimizing the influence of the outliers on the values of the estimates. In this algorithm, RANSAC determines the rotational matrix that aligns most directional correspondences while disregarding the pairs of directions that fail to align with the others. In the RANSAC scheme, I consider any two directions to be aligned as the angle between the two directions to be within an error threshold of 0.9. More specifically, the dot product between two directions $\mathbf{u}$ and $\mathbf{v}$ larger than 0.9 i.e., $\mathbf{u} \cdot \mathbf{v} \geq 0.9$ are considered aligned in RANSAC.

For translational alignment, I take advantage of the mathematical property that DECT reacts linearly to translational shifting of the topology, i.e., assume a shape M offsets translationally by vector α , then for each DEC curve in direction $\mathbf{v}$ is translated by $\alpha \cdot \mathbf{v}$, i.e.,

$$\begin{aligned} \text{DEC}_M(\mathbf{v}, h) &= \int_{\mathbb{R}^d} 1_M(x + \alpha) \cdot 1((x + \alpha) \cdot \mathbf{v} \leq h) \chi(dx, \mathbf{v}, h) \\ &= \int_{\mathbb{R}^d} 1_M(x + \alpha) \cdot 1(x \cdot \mathbf{v} \leq h - \alpha \cdot \mathbf{v}) \chi(dx, \mathbf{v}, h) \\ &= \text{DEC}_M(\mathbf{v}, h - \alpha \cdot \mathbf{v}). \end{aligned} \tag{5.1}$$

Therefore the problem of translational alignment can be treated as finding the offset of translation in the DEC curves. The translational offset between two DEC curves is estimated by cross-correlation using the Fast Fourier Transform, where the position of the maximum cross-correlation is the estimated linear offset between the two curves in the same direction. These functional offset $\{d_v\}$ correspond to $h - \alpha \cdot v$ for each direction v . Then I solve for α by linear regression on $d = h - \alpha \cdot v$,

in which the slope of the linear regression corresponds to the linear translation offset vector α . Translating the structure by $-\alpha$ undo this offset and translationally aligns the protein structures.

I applied this approach to align input protein structures for SINATRA Pro. Figures comparing the correlation between SINATRA Pro results on the five protein systems using the RMSD alignment versus using the topological rotational alignment are shown in the Supporting Information (Figs. B.25 and B.26). There are good agreements between both alignment procedures in almost all protein systems. The one exception is HIV-1 protease, where it is believed that the small, flexible protein makes it challenging to identify stable reference signatures for alignment in summary statistic space.

5.1.3 Coarse Alignment scheme to align protein homologs and analogs

Protein structure alignment is not always trivial. Protein structures vary in size and often contain part of the structure that does not correspond to any part of the structure in their aligning counterpart. For example, when aligning structures in the same MD trajectory, all structures belong to the same protein and have the same size. In this case, assigning atomic correspondence is trivial. However, proteins differ in size and sequence. For example, protein distance homologs that inherit similar features from their common ancestor have similar sequences but different sizes [247]. Protein analogs that converge to similar energetically-favored structures have similar shapes but different sequences since they inherit from different ancestors [247].

These protein homologs and analogs pose a slightly different challenge to protein alignment algorithms. Proteins of low sequence similarity are challenging for sequence-based protein alignment algorithms since the correspondence between the sequence cannot be trivially assigned, especially when few sequences are available for the structure. Proteins of different sizes contain structural outliers, i.e., irrelevant structural parts, which often interfere with the alignment algorithm and cause false alignment. For example, these structural outliers disrupt the distances between directional DEC, leading to faulty alignments. Therefore, the challenge for the protein structural alignment algorithm for protein homologs and analogs is to identify the analogous structural parts in the two matching proteins and disregard structural outliers from the alignment. To overcome this challenge, I introduced a coarse alignment scheme based on matching atoms with similar local topology.

To capture the local topological environment around each residue, I introduced a new variant of the DECT I named the radial DECT (RDECT) prior. Instead of performing the filtration in the directional, linear manner, I radially perform the filtration, i.e., a sphere expands from a centroid.

Mathematically, similar to Eq. 4.2, I define the RDEC as,

$$\Delta_r \chi = \Delta_r V(\mathcal{M}) - \Delta_r E(\mathcal{M}) + \Delta_r F(\mathcal{M}). \quad (5.2)$$

Similar to EC and DEC, the collection $\{\#V(\mathcal{M}), \#E(\mathcal{M}), \#F(\mathcal{M})\}$ denotes the number of vertices (atoms), edges (connections between atoms), and faces (triangles enclosed by edges) of the mesh, respectively. An RDEC curve $\Delta_r \chi(\mathcal{M})$ tracks the change in the Euler characteristic over an expanding sphere centered at centroid $\mathbf{r}_0$ with respect to a given filtration of length l i.e. $\Delta_r V(\mathcal{M})$, $\Delta_r E(\mathcal{M})$, and $\Delta_r F(\mathcal{M})$ are the numbers of vertices, edges, and faces born at distance $r' = |\mathbf{x} - \mathbf{r}_0|$ between radius r and $r + \Delta r$ i.e. $r \leq r' \leq r + \Delta r$, where $\mathbf{x}$ and $\mathbf{r}_0$ is the Cartesian coordinates of the vertices, and the centroid respectively, with C_α atom of each residue being the centroid. This RDEC metric allows us to capture the local topological profile around each of the C_α atoms, i.e., the “center” of each residue. I then use this metric to assign residual correspondence between the two protein structures. Fig. 5.3 shows the overview of this RDECT-based coarse alignment scheme.

I then compute the Euclidean distance between RDEC curves to quantify the similarity between the local topological profile of each residue. To further improve the quality of the alignment in the evolutionary context, I include the BLOSUM62 score (BLOCKS SUBstitution Matrix) [248], a sequence-based distance metric derived from close homologs which measure the evolutionary distance between the amino acids, i.e., amino acids that are less distant has higher BLOSUM score. I multiply the topological distances by the normalized negative BLOSUM score to scale down the cost of amino acid pairs that are evolutionarily proximal. The resulting cost matrix for input to Needle-Wunsch (NW) algorithm [249], a dynamic programming algorithm for comparing protein and gene sequences, to assign atomic correspondence with similar local topology. I then apply RANSAC [174] to neglect outliers and identify the rotational matrix that aligns the most relevant correspondences.

5.1.4 Benchmarking using Protein distance homologs datasets

To assess the ability of this alignment scheme to reproduce alignments in manually curated data sets, I apply the alignment scheme on the MALISAM (Manual ALIGNments of Structurally Analogous Motifs), a database of pairwise, structure-based alignments for structurally analogous motifs in proteins [246]. Results of the assessment is shown in and Table 5.1 and Fig. 5.3. While EC align recreates most manual alignments in MALISAM, its alignment accuracy is still not up to par compared to current structural alignment methods.

5.2 Topological Data Analytic Framework for Evaluating and Predicting Protein Conformational Transition

In Chapter 4, I proposed SINATRA Pro, a topological data analytic pipeline for identifying biologically-relevant structural differences between two protein structural ensembles. SINATRA Pro robustly detects subtle topological shifts in protein structural ensemble. To ensure sampling consistency, I performed sensitivity analyses based on protein structures drawn from different lengths up to 200 ns MD simulation trajectories (Figs. B.7, B.10, B.12, B.14, and B.16). In this sensitivity assessment, SINATRA Pro picks up minute structural variations from short trajectories, given that these trajectories statistically contain the structural dynamics information. It begs the question of if SINATRA Pro also picks up signs of small- and long-timescale transition from a short trajectory, and how long does the trajectory have to be for SINATRA Pro to pick up a signal.

The significance of this problem is that there is no quick way to assess if a conformation is going to transit to another conformation. Existing tools and metrics, such as root-mean-squared deviation (RMSD), and Normal Mode Analysis (NMA), are susceptible to noises and starting structure of the MD simulation. It would be advantageous to design a statistical tool that can assess if a conformational transition will happen soon in an MD simulation and explore any hidden conformations of dynamic protein systems such as kinases beyond the timescale of MD simulations. Advanced statistical tools, combined with the topological data analytical approach, would allow us to assess the statistical and thermodynamic stability of the native states and transition states.

As the first step in this direction, it would be intuitive to the extent its application to detect and describe conformation change in protein dynamics, i.e., frames in the transitions versus frames in native states. I use a short MD trajectory data of transition events in WT and compound mutant Abl-1 Kinase as a testing ground. Fig. 5.4 shows the preliminary results, where SINATRA Pro identifies structural variations mainly in the helix αC and the neighboring structures, which are heavily involved in the DFG flip transition, from the short < 1 ns trajectory.

Another potential statistical approach is to use an unsupervised learning algorithm, such as a variational autoencoder model, to evaluate the dynamic stability of the protein, i.e., whether the protein structures are susceptible to conformational transition or identify transitional or unstable states from a trajectory or an ensemble of structures. A further statistical application can consider

the kinetics in serial MD simulations, requiring a statistical model that incorporates time-series information.

5.3 Gaussian Process Regression Framework for identifying biophysically relevant structures

Existing tools [250, 251] can quickly evaluate or predict biophysical properties, such as binding, folding energy, and fitness, of a protein structure. However, they do not directly solve the inverse problem of inferring which parts of the protein structure contribute most towards changes in the physical properties. The ability to provide a solution to this inverse problem can be helpful as it provides information to infer the structural dynamics relevant to the physical properties and perhaps hints at potential advantageous mutations for protein design optimization.

From a theoretical perspective, the problem turns the statistical problem from Gaussian Process (GP) Classification in SINATRA Pro to GP Regression problem. Since the RATE algorithm also solves the GP Regression problem, converting and modifying the SINATRA Pro program for this application would be trivial. The challenging part of this project would be the data quality of the physical properties, which are subject to experimental and statistical uncertainty, and systematic errors. The in silico sampling of protein structures also depends on the MD simulation setup, such as the force field parameters, initial configuration, and the length of the simulation, which all affect the quality of the structural ensemble, i.e., are they physically accurate and do they contain all the relevant conformations of the transitions of interest.

It would be ideal to start with a protein system of a tractable, small size with straightforward binding mechanics, such as TEM-1 β -lactamase, where R164S mutation induced a change in the structural dynamics in the Ω -loop of Wild Type TEM-1. In this project, the effect of the structures on ligand binding properties can be quantified by reference metrics such as docking score [252] and folding energy [251]. For instance, one can calculate the docking score of each frame in an MD trajectory, then use the GP regression model to infer the structural parts of the protein that are most relevant to the docking score. Since the physical metrics are susceptible to noise and uncertainties, it would be advantageous also to improve the statistical model to incorporate them. One should also set up a baseline to assess the reliability of the generated results.

5.4 Tables and Figures

Table 5.1: Root-mean-squared deviation (RMSD) with respect to manual alignment in manually-curated protein distant homologs dataset MALISAM for structures aligned by EC align (this study), TM align [231], DALI [223] and FAST [230].

Method	Average RMSD	% of alignments with RMSD < 6 Å
EC align (Coarse)	7.96	60.8%
EC align (Coarse+DECT)	8.28	63.1%
TM align [231]	4.94	77.7%
DALI [223]	2.46	90.8%
FAST [230]	4.17	83.8%

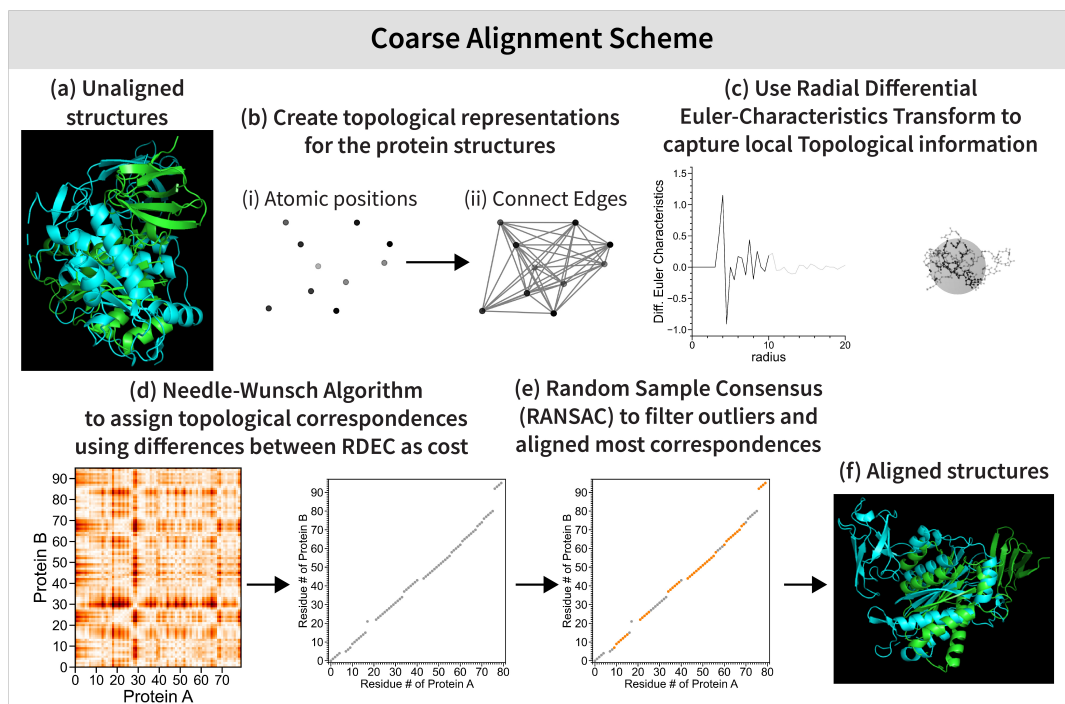


Figure 5.1: Protein structural alignment using directional DEC curves to align directions with similar topological profiles. **(a)** shows the unaligned structures. **(b)** First, the atomic coordinates of the heavy backbone atoms of the protein structures are taken as vertices, edges within a radius cutoff $2 \text{ \AA}$ are connected as edges to create a simplicial complex for each protein structure. **(c)** Then, the global directional topological profile is summarized by calculating the DECT in 500 directions. **(d)** After applying a Gaussian Kernel to smooth out the DEC curves, the matrix containing the Euclidean distances between the DEC curves is taken as the cost matrix to assign correspondences by minimizing the linear cumulative cost. **(e)** I then apply RANSAC [174] to align most directional correspondences (aligned correspondences in orange) from (d). **(f)** shows the final aligned structures using this DECT alignment scheme.

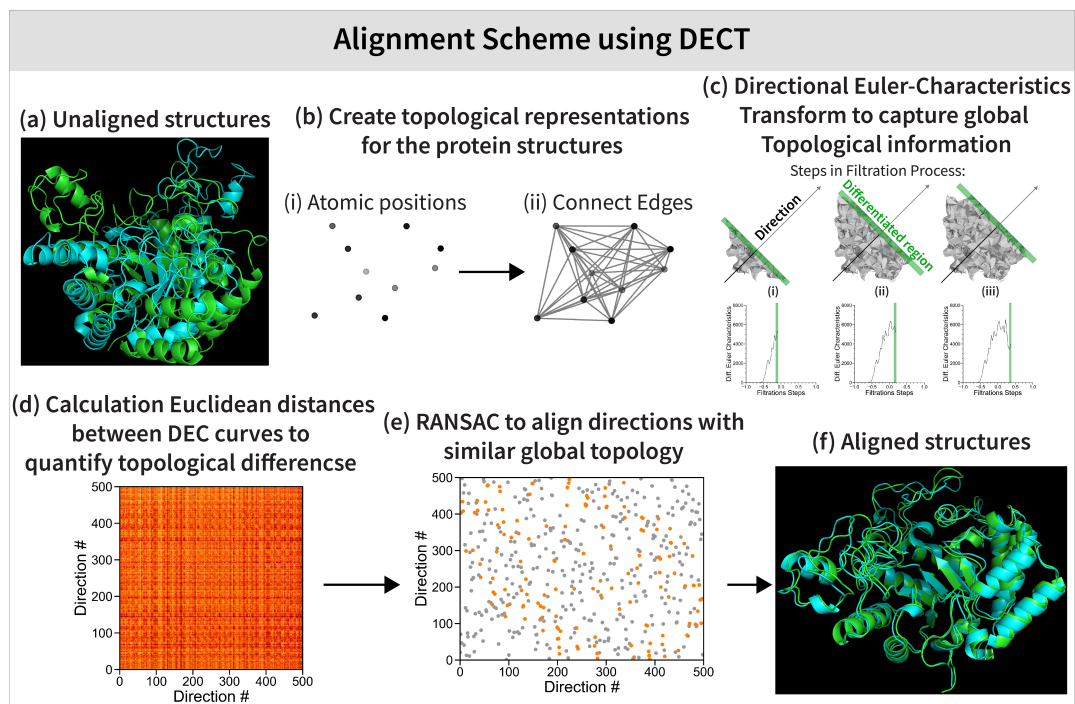


Figure 5.2: Coarse Alignment scheme to remove structural outliers and align topologically relevant structures. **(a)** shows the unaligned structures. **(b)** First, the atomic coordinates of the heavy backbone atoms of the protein structures are taken as vertices, edges within a radius cutoff $2 \text{ \AA}$ are connected as edges to create a simplicial complex for each protein structure. **(c)** Then, the local topology around each residue is summarized by calculating the Radial DECT, taking the C_{α} atom of each residue as the centroid. **(d)** After applying a Gaussian Kernel to smooth out the RDEC curves, the matrix containing the Euclidean distances between the RDEC curves (multiplied by the BLOSUM scores) is taken as the cost matrix to assign correspondences using the Needle-Wunsch Algorithm [249]. **(e)** I then apply RANSAC [174] to align most atomic correspondences (aligned correspondences in orange) and ignore structural outliers (in grey) from (d). **(f)** shows the final aligned structures using this coarse alignment scheme.

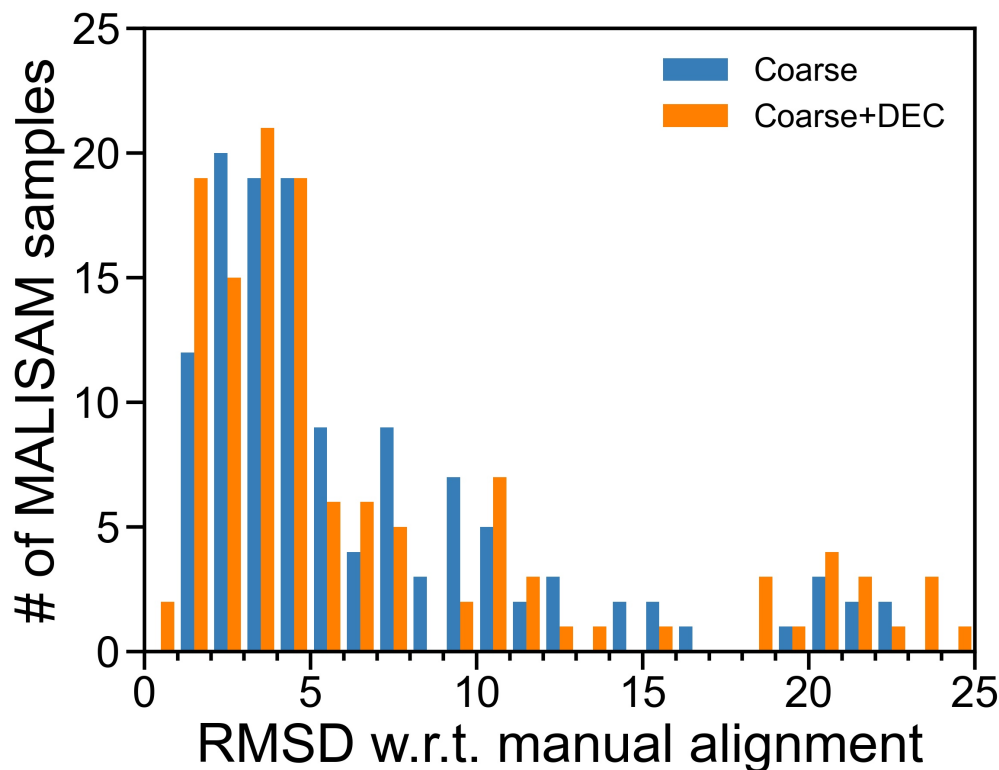


Figure 5.3: EC align recreates manual alignment in manually-curated protein distant homologs dataset MALISAM. The resulting alignments are assessed by the root-mean-squared-deviation (RMSD) with respect to manual alignments. First, structure A in the EC-aligned pair aligns with that in the manually aligned pair. Then RMSD between the EC-aligned structure B and manually aligned structure B is calculated as the discrepancies between EC alignment and manual alignment. In this assessment, 60.8% of the coarse alignment and 63.1% of the Coarse+DEC alignment are within $\text{RMSD} = 6 \text{ \AA}$ with respect to manual alignment in MALISAM. Parameters are optimized by grid search, for coarse alignment, radius cutoff for simplicial construction = $1.5 \text{ \AA}$, gap penalty ratio for NW algorithm = 0.15, maximum radius for RDECT = $12.0 \text{ \AA}$, number of filtrations per Angstrom for RDECT = 200, width of Gaussian kernel applied for smoothing = $1.4 \text{ \AA}$; in RANSAC, number of draws = 6, number of iterations = 50000, distance tolerance = $6.6 \text{ \AA}$. For fine alignment, number of iterations = 5, number of directions for DECT = 500, width of Gaussian kernel applied for smoothing = 0.5, number of filtration per Angstrom in DECT = 200, number of draws in each iteration of RANSAC = 100, tolerance in dot product between aligned directions in RANSAC = 0.9.

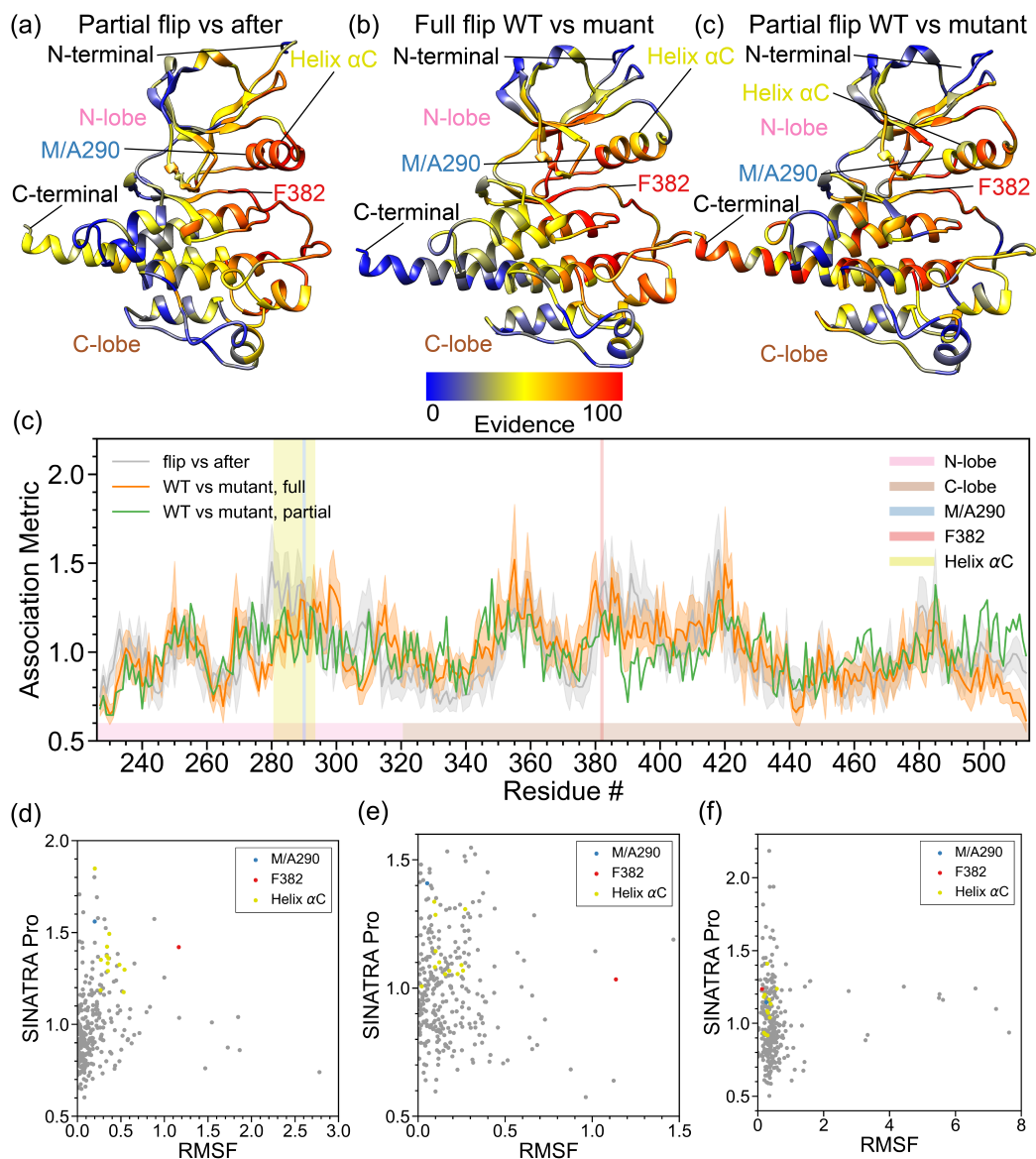


Figure 5.4: SINATRA Pro results on Abl1 Transition trajectory (a) Short MD trajectory (~ 700 ps) of the Wild Type (WT) capturing the DFG-flip transition is split into two parts, the translational state (“Partial flip”, first ~ 140 ps) and the destination native state (“after”, ~ 140 -700). SINATRA Pro is applied to compare 100 structures drawn from the “Partial flip” and “after” trajectories. (b) Short MD trajectory (~ 700 ps) of the WT capturing the DFG flip transition and MD trajectory of the same length of the mutant capturing a failure in transition due to the mutation. SINATRA Pro is applied to compare 100 structures drawn from each of full ~ 700 ps WT and mutant trajectory. (c) The transition partial MD trajectory (~ 140 ps) of the WT capturing the DFG-flip transition and MD trajectory cropped from the first ~ 140 ps of the mutant trajectory. SINATRA Pro compares 50 structures drawn from the partial ~ 140 ps WT and mutant trajectories. (d) shows the association metrics computed for each residue within each analysis. The final row plots the correlation between the SINATRA Pro association metrics and the root mean square fluctuation (RMSF) for all atoms with correspondences in the (e) whole protein, (f) fragment 65-230, and (g) fragment 65-213. Parameters for SINATRA Pro is same as in Chapter 4 $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$.

Chapter 6

Discussion and Conclusion

In this thesis, I present the analytical approaches for optimizing the accuracy of molecular dynamics simulations and extracting structural signatures from the simulated ensembles.

While there are spectrum of *in silico* and *in vitro* methods for sampling physical ensemble of protein structures, the capability of MD simulations in generating such ensembles in a convenient and reliable manner have put MD simulations in the spotlight of molecular biophysics research. While the research community is reliant on the reliability of MD simulations, the underlying physical model i.e. state-of-the-art protein force fields shows significant discrepancies with respect to experimental results. In Chapter 2, we proposed a residue-type specific modifications on backbone torsion potentials of ff99SBws improved the accuracy of all-atom simulation of low-complexity IDPs in terms of their agreement with experimental results [1]. While the refined parameters in the proposed modified force field ff99SBws-STQ can be further improve by incorporating more experimental data on IDPs and by a more statistically robust parameter tuning methods, the proposed force field provide a significantly more accurate description of the structural and dynamical properties of IDPs, which are rich in polar residues and involved in the formation of biomolecular condensates.

In Chapter 3, by combining NMR experiments, and metadynamics MD simulations with the improved IDP force field ff99SBws-STQ, we demonstrate that the liquid-liquid phase separation among intrinsically disordered SYGQ LC and RGG intrinsically disordered regions, and RNA polymerase II, are contributed by weak and distributed molecular interactions among various amino acid types [2]. We demonstrated that the combination of experimental and computational methods provides a comprehensive perspective on the biophysical mechanism underlying the physiological process, e.g.,

LLPS involving IDPs.

Conventional protein analysis approaches, such as RMSF and PCA, cannot statistically separate interclass heterogeneity, such as mutation-induced or binding-induced changes in structural dynamics, from intraclass variations, such as protein dynamics and thermal fluctuations. SINATRA Pro presented a statistically robust approach to identify such interclass heterogeneity between two protein structural ensembles, with mathematical consideration of intraclass variations [3]. This approach provides an unbiased interpretation of the structural data as a hypothesis generation tool that practitioners can use to design more informed experiments for answering downstream scientific questions on the biophysical mechanism of the biomolecular systems. While we have discussed the further applications of SINATRA Pro in detail in Chapter 4 and 5, I believe the idea of the topological analytic approach demonstrated in SINATRA Pro has broad applicability in providing the solution to a range of biological problems in molecular biophysics.

Recent and ongoing advancements in statistical and topological approaches for protein analysis and structural prediction are pushing the boundary of computational protein research. Results obtained from computational studies bear statistical and quantitative robustness rather than weak, qualitative description. I believe this will continue to be the trend and central theme for my future research. The proposal of AlphaFold [222] not only significantly impacted the research community in protein folding but also has substantial implications for disordered proteins [253]. Future efforts in computational protein research will also focus on the analysis of dynamic predictions than just structural prediction, especially on disordered proteins.

Appendix A

Supporting Material for “Residue-level refinement of all-atom protein force field for intrinsically disordered proteins”

Protein	N	Sequence
FUS ₀₋₄₃ *	44	GMASNDYTQQATQSYGAYPTQPGQGYSSQSSQPYGQQSYSGYSQ
TDP43 ₃₁₀₋₃₅₀	41	GMNFGAFSINPAMMAAAQAALQSSWGMMLASQQNQSGPS
RNA Pol II ₁₉₂₇₋₁₉₇₀	44	SPTYSPTSPKGSTYSPTSPPGYSPSTPYSLTSPAISPDDSDDEEN
FUS ₄₁₋₈₄	44	YSQSTDTSGYGQSSYSSYGGQSNTGYGTQSTPQGYGSTGGYGSS
FUS ₇₇₋₁₂₀	44	STGGYGSSQSSQSSYGGQSSYPGYGQPAPSSSTSGSYGSSSQSS
FUS ₁₂₀₋₁₆₃	44	SSYGQPQSGYSQQPSYGGQQQSYGQQQSYNPPQGYGQQNQYNS
FUS ₃₇₋₉₇	61	SYSGYSQSTDTSGYGQSSYSSYGGQSQNTGYGTQSTPQGYGSTGGYGSSQSSQSSYGGQSSY
hnRNPA2 ₁₉₀₋₂₃₃	44	GRGGNFGFGDSRGGGGNFGPGPGSNFRGGSDGYSGRGGFDGYN
hnRNPA2 ₂₆₅₋₃₀₈	44	GNQGGGYGGGYDNYGGGNYGSGNYNDFGNYNQPSNYGPMKSGN

*a starting glycine is included in FUS₀₋₄₃ as residue 0

Table A.4: Root-mean-squared deviation (RMSD) of simulated $\Delta\delta C_\alpha$, $\Delta\delta C_\beta$, $\Delta\delta C_\alpha - \Delta\delta C_\beta$ (in ppm) from NMR ^{13}C chemical shifts of three IDPs for ff99SBws, modified ff99SBws variants (S,T,ST,TQ,STQ) (this study), and ff99SB-disp.

$\Delta\delta C_\alpha - \Delta\delta C_\beta$	FUS ₀₋₄₃	TDP-43 ₃₁₀₋₃₅₀	RNA Pol II ₁₉₂₇₋₁₉₇₀
ff99SBws	0.90 ± 0.11	0.64 ± 0.09	0.68 ± 0.11
ff99SB-disp	0.90 ± 0.11	0.78 ± 0.07	0.74 ± 0.11
ff99SBws-S	0.86 ± 0.10		
ff99SBws-T	0.98 ± 0.13		
ff99SBws-ST	0.64 ± 0.07	0.71 ± 0.08	0.74 ± 0.08
ff99SBws-TQ	0.62 ± 0.08	0.66 ± 0.06	0.71 ± 0.10
ff99SBws-STQ	0.54 ± 0.07	0.51 ± 0.04	0.74 ± 0.08
$\Delta\delta C_\alpha$	FUS ₀₋₄₃	TDP-43 ₃₁₀₋₃₅₀	RNA Pol II ₁₉₂₇₋₁₉₇₀
ff99SBws	0.58 ± 0.07	0.43 ± 0.06	0.67 ± 0.11
ff99SB-disp	0.59 ± 0.07	0.54 ± 0.05	0.62 ± 0.09
ff99SBws-S	0.51 ± 0.06		
ff99SBws-T	0.61 ± 0.08		
ff99SBws-ST	0.41 ± 0.04	0.41 ± 0.04	0.62 ± 0.08
ff99SBws-TQ	0.36 ± 0.05	0.40 ± 0.04	0.64 ± 0.09
ff99SBws-STQ	0.37 ± 0.04	0.36 ± 0.03	0.62 ± 0.08
$\Delta\delta C_\beta$	FUS ₀₋₄₃	TDP-43 ₃₁₀₋₃₅₀	RNA Pol II ₁₉₂₇₋₁₉₇₀
ff99SBws	0.43 ± 0.05	0.42 ± 0.05	0.51 ± 0.06
ff99SB-disp	0.39 ± 0.05	0.39 ± 0.06	0.44 ± 0.11
ff99SBws-S	0.45 ± 0.06		
ff99SBws-T	0.44 ± 0.05		
ff99SBws-ST	0.37 ± 0.04	0.44 ± 0.05	0.47 ± 0.10
ff99SBws-TQ	0.40 ± 0.05	0.41 ± 0.06	0.42 ± 0.11
ff99SBws-STQ	0.34 ± 0.04	0.34 ± 0.06	0.47 ± 0.10

Protein sequences from NMR experiments

FUS Low Complexity domain FUS_{1–163} (BMRB 26672)[68]

MASNDYTQQA TQSYGAYPTQ PGQGYSSQSS QPYGQSYSG YSQSTDTSY GQSSYSSYGQ
SQNTGYGTQS TPQGYGSTGG YGSSQSSQSS YGQSSYPGY GQPAPSSTS GSYGSSSQSS
SYGQPQSGSY SQQPSYGGQQ QSYGQQSYN PPQGYGQQN YNS

(163 residues)

TDP43 Wild Type Low Complexity C-terminal domain TDP43_{267–414} (BMRB 26823)[15]

GHMNRQLERS GRFGGNPGGF GNQGGFGNSR GGGAGLGNNQ GSNMGGGMNF GAFSINPAMM
AAAQAALQSS WGMMLASQ QNQSGPSGNN QNQGNMQREP NQAFGSGNNS YSGSNSGAII
GWGSASNAGS GSGFNGGFGS SMDSKSSGWG M

(151 residues)

RNA Pol II C-terminal Domain 27-52 RNA Pol II_{1770–1970} (BMRB 27063) [66]

GHMSPNYTPT SPNYSPTSPS YSPTSPSYSP TSPSYSPSSP RYTPQSPTYT PSSPSYSPSS
PSYSPASPKY TPTSPSYSPS SPEYTPSPK YSPTSPKYSP TSPKYSPTSP TYSPTTPKYS
PTSPTYSPS PVTPTSPKY SPTSPTYSP SPKYSPTSPT YSPTSPKGST YSPTSPGYSP
TSPTYSLTSP AISPDDSDEE N

(201 residues)

hnRNPA2 low complexity domain 190-341 hnRNPA2_{190–341} (BMRB 27123) [51]

GHMGRGGNFG FGDSRGGGNN FGPGPGSNFR GGSDGYGSGR GFGDGYNGYG GGPGGGNFGG
SPGYGGGRGG YGGGGPGYGN QGGGYGGGYD NYGGGNYGSG NYNDFGNYNQ QPSNYGPMKS
GNFGGSRNMG GPYGGGNYGP GSGGGGGYG GRSRY

(155 residues)

The residues in italics correspond to TEV-cleavage and cloning “scar” residues not present in the native sequences.

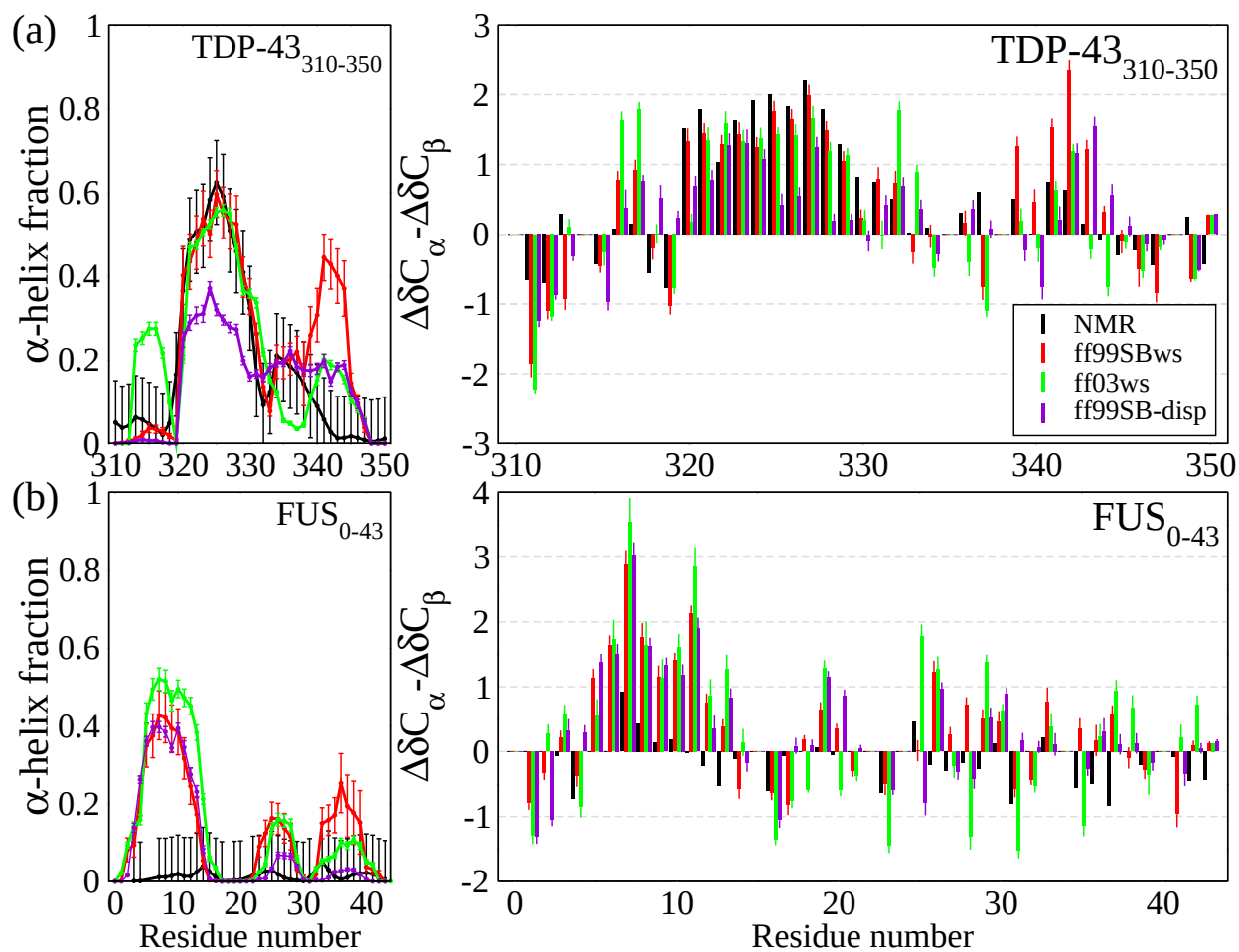


Figure A.1: Comparison of α -helix propensities and NMR ^{13}C chemical shifts between experiments and simulations. **(a)** α -helix propensities and **(b)** Chemical shifts $\Delta\delta\text{C}_\alpha - \Delta\delta\text{C}_\beta$ of TDP-43₃₁₀₋₃₅₀ and FUS₀₋₄₃ using all-atom force fields ff99SBws, ff03ws, and ff99SB-disp, compared to experimental results. Chemical shifts are calculated from protein structure using SPARTA+ algorithm. Secondary structures are assigned by DSSP. Experimental values of secondary structure propensities are calculated from NMR chemical shifts using the $\delta 2\text{D}$ program.

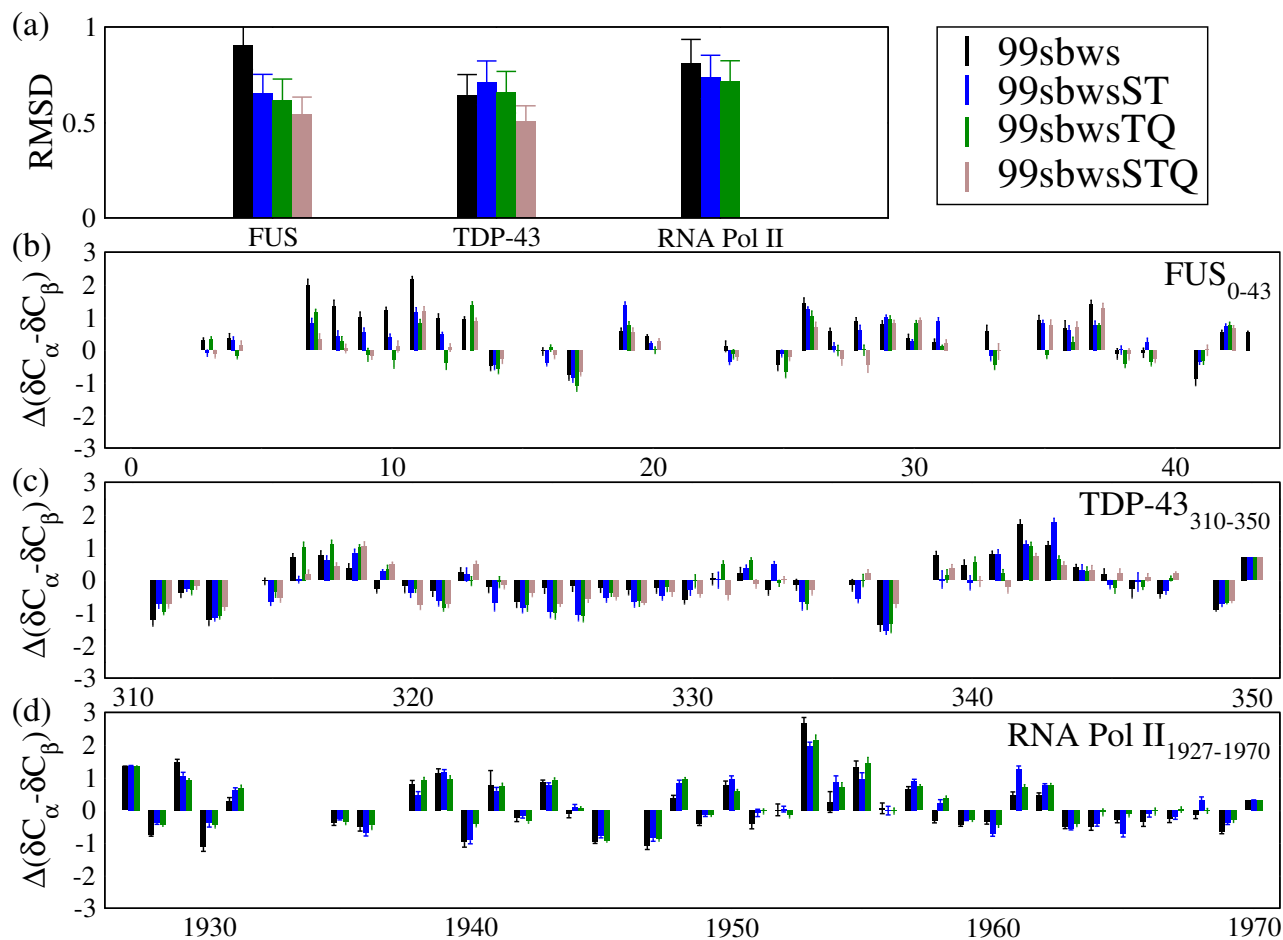


Figure A.2: Comparison of NMR ^{13}C chemical shifts between experiments and simulations. (a) Comparison of Root-mean-squared deviation (RMSD) of simulated $\delta C_\alpha - \delta C_\beta$ (in ppm) from NMR ^{13}C chemical shifts of simulated sequences and (b-d) chemical shift deviation from the experiment $\Delta(\delta C_\alpha - \delta C_\beta)$ with ff99SBws and modified force fields (see legend).

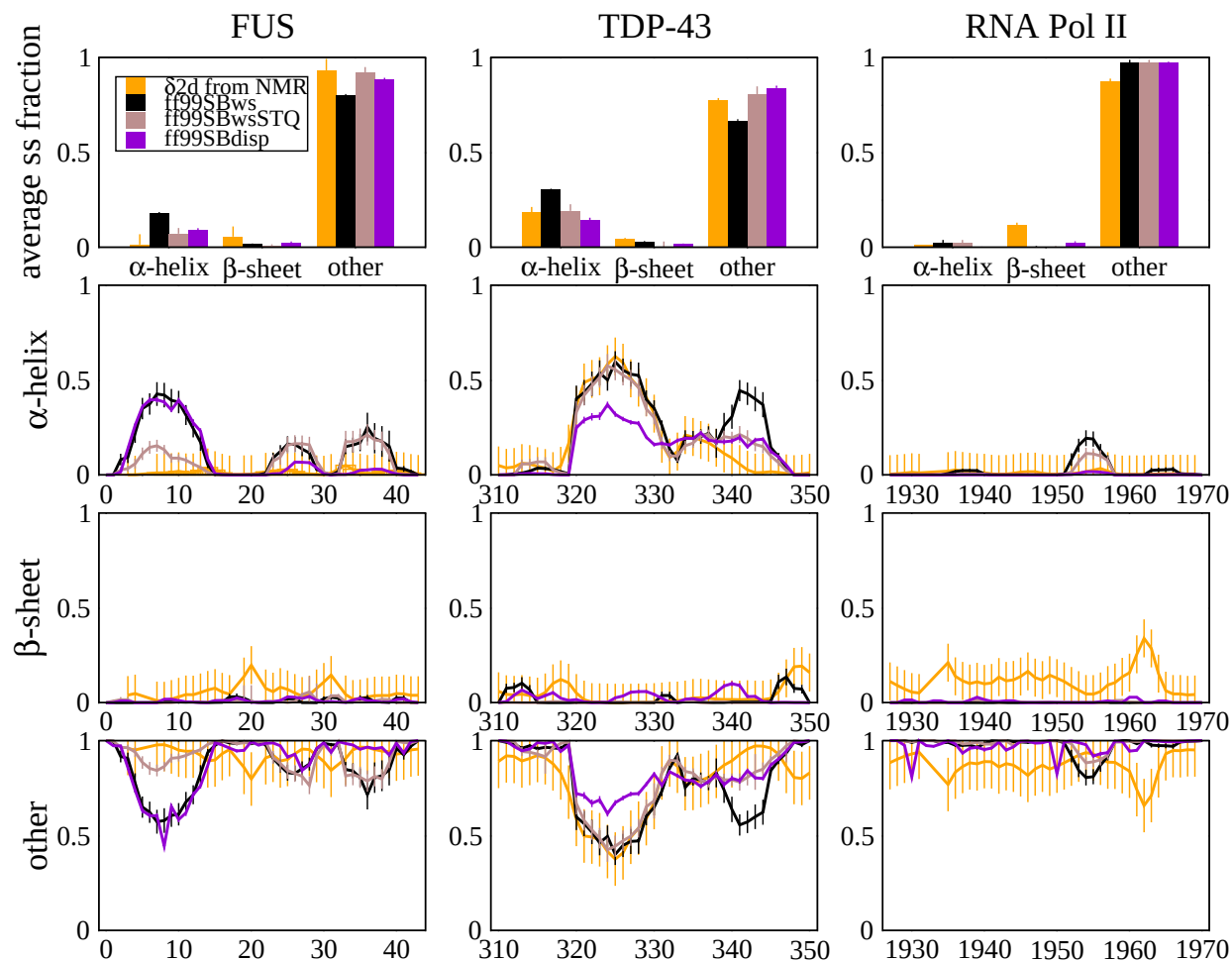


Figure A.3: Secondary structure propensities of FUS₀₋₄₃, TDP-43₃₁₀₋₃₅₀, and RNA Pol II₁₉₂₇₋₁₉₇₀ with ff99SBws, ff99SBws-STQ (this study) and ff99SB-disp, compared to δ 2D secondary structure propensities calculated from experimental NMR chemical shifts.

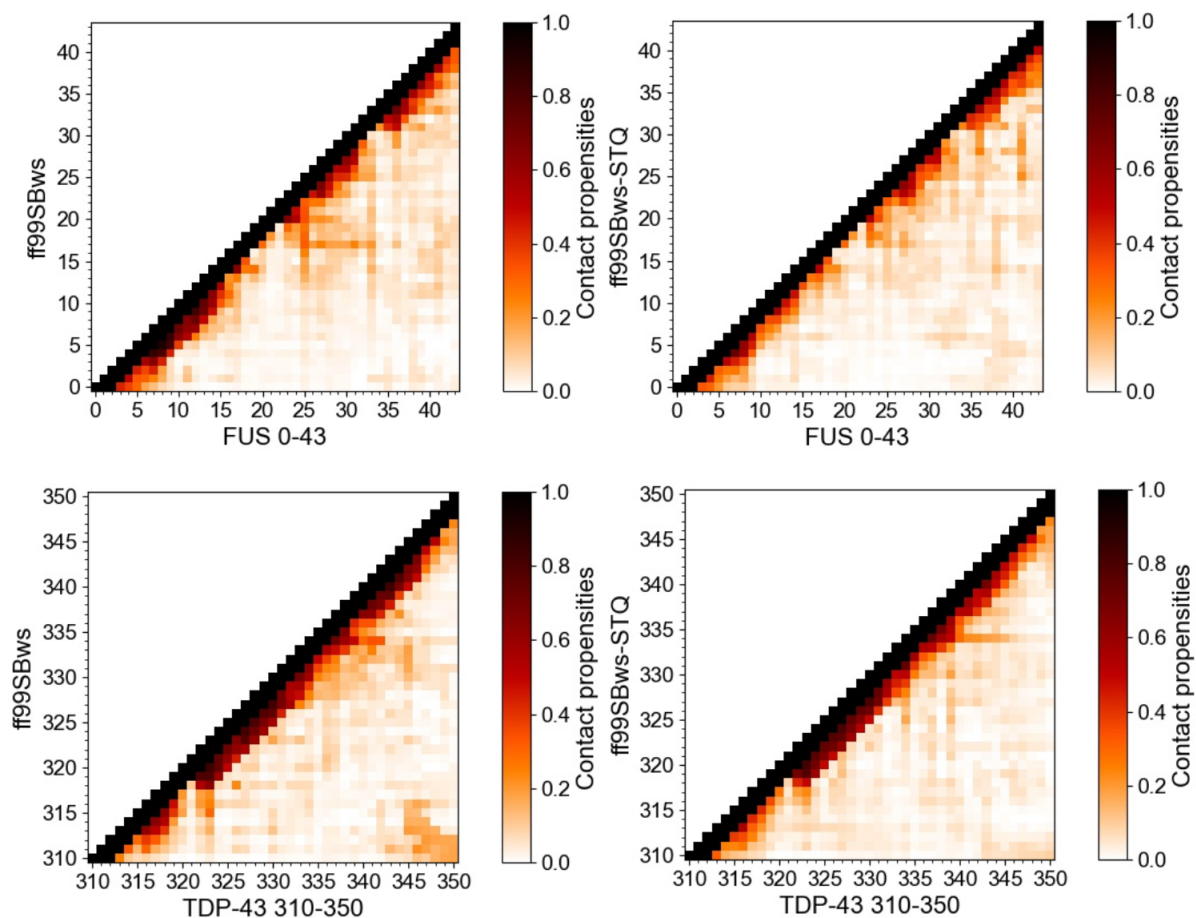


Figure A.4: Intrachain residue-to-residue contact propensities in FUS_{0-43} and $\text{TDP-43}_{310-350}$ with ff99SBws and ff99SBws-STQ (this study). Two residues are considered in contact when any heavy atoms from each of the residue are within 6 Angstrom in simulation. Contact propensities are the proportion of frames in simulation where two residues are in contact, with contact propensity = 1 being always in contact.

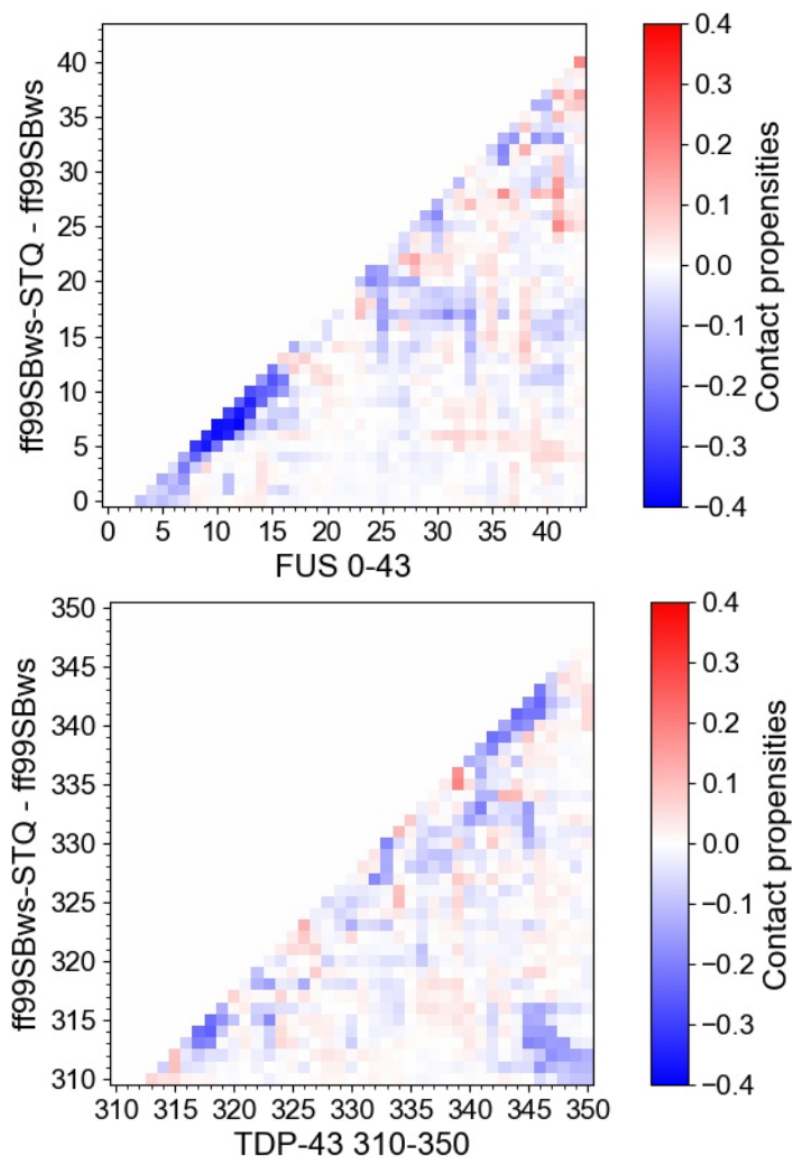


Figure A.5: Change in Intrachain residue-to-residue contact propensities in FUS₀₋₄₃ and TDP-43₃₁₀₋₃₅₀ from ff99SBws to ff99SBws-STQ. (this study) Two residues are considered in contact when any heavy atoms from each of the residue are within 6 Angstrom in simulation. Contact propensities are the proportion of frames in simulation where two residues are in contact, with contact propensity = 1 being always in contact.

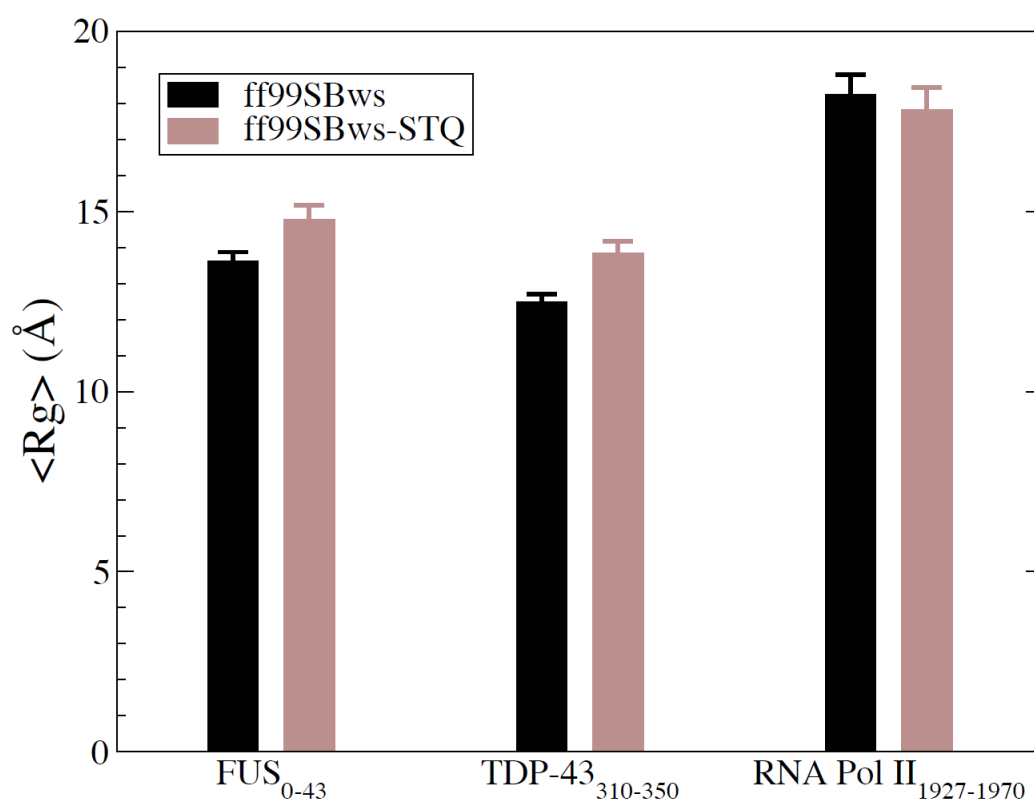


Figure A.6: Average radius of gyration (R_g) in FUS₀₋₄₃, TDP-43₃₁₀₋₃₅₀ and RNA Pol II₁₉₂₇₋₁₉₇₀ with ff99SBws and ff99SBws-STQ (this study).

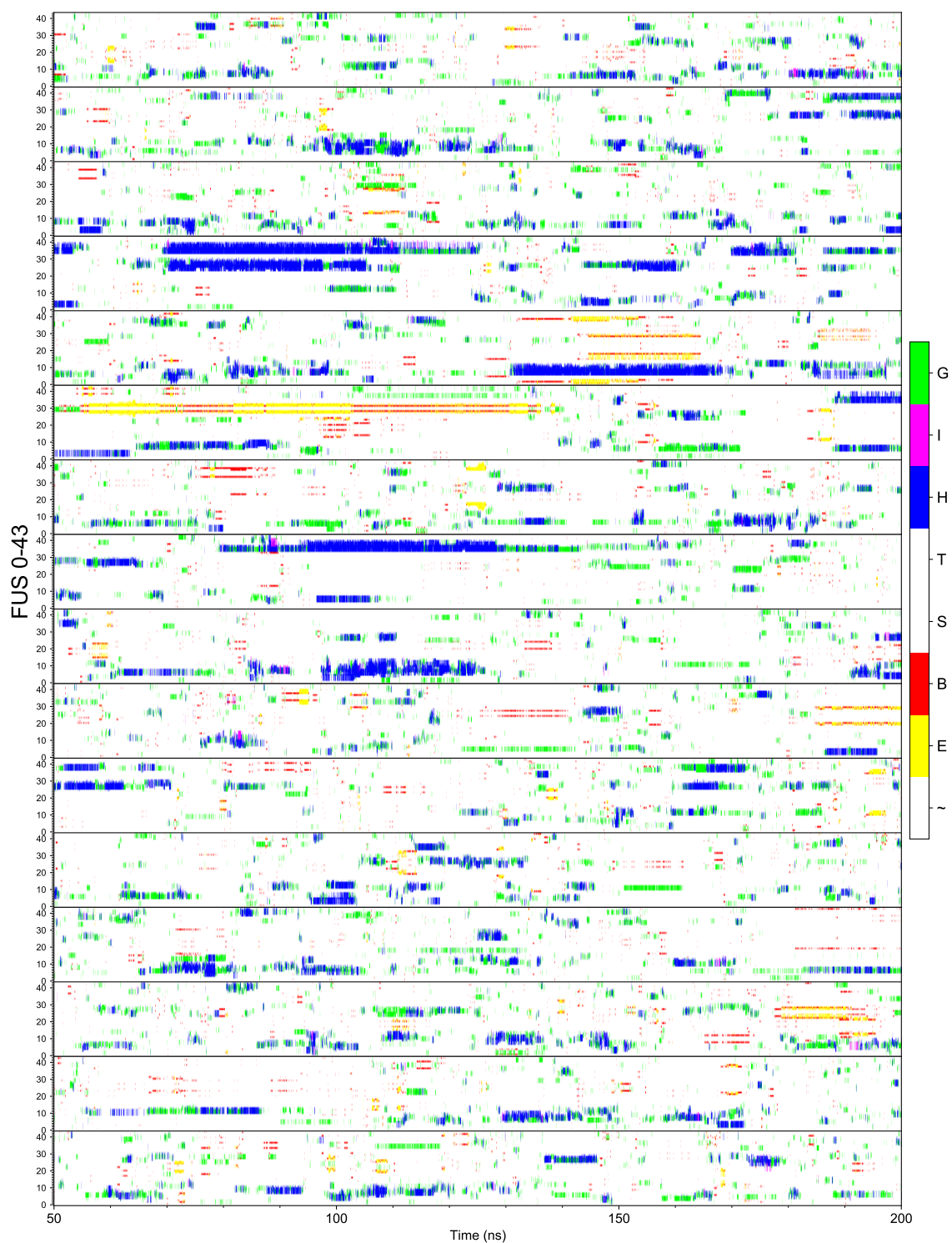


Figure A.7: DSSP secondary structure as a function of simulation time for each demultiplexed replica moving through the temperature space of FUS₀₋₄₃ with ff99SBws-STQ (this study).

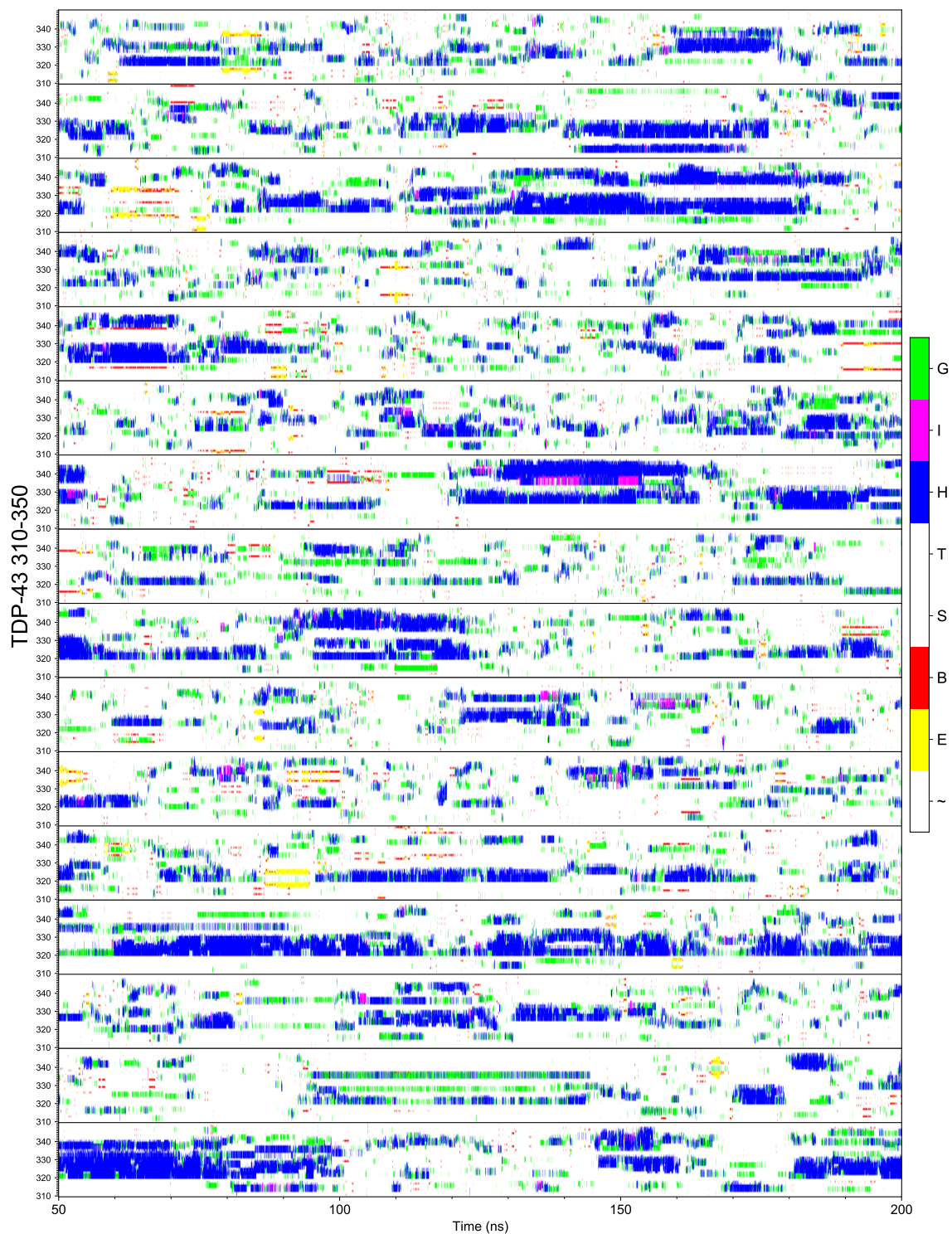


Figure A.8: DSSP secondary structure as a function of simulation time for each demultiplexed replica moving through the temperature space of TDP-43_{310–350} with ff99SBws-STQ (this study).

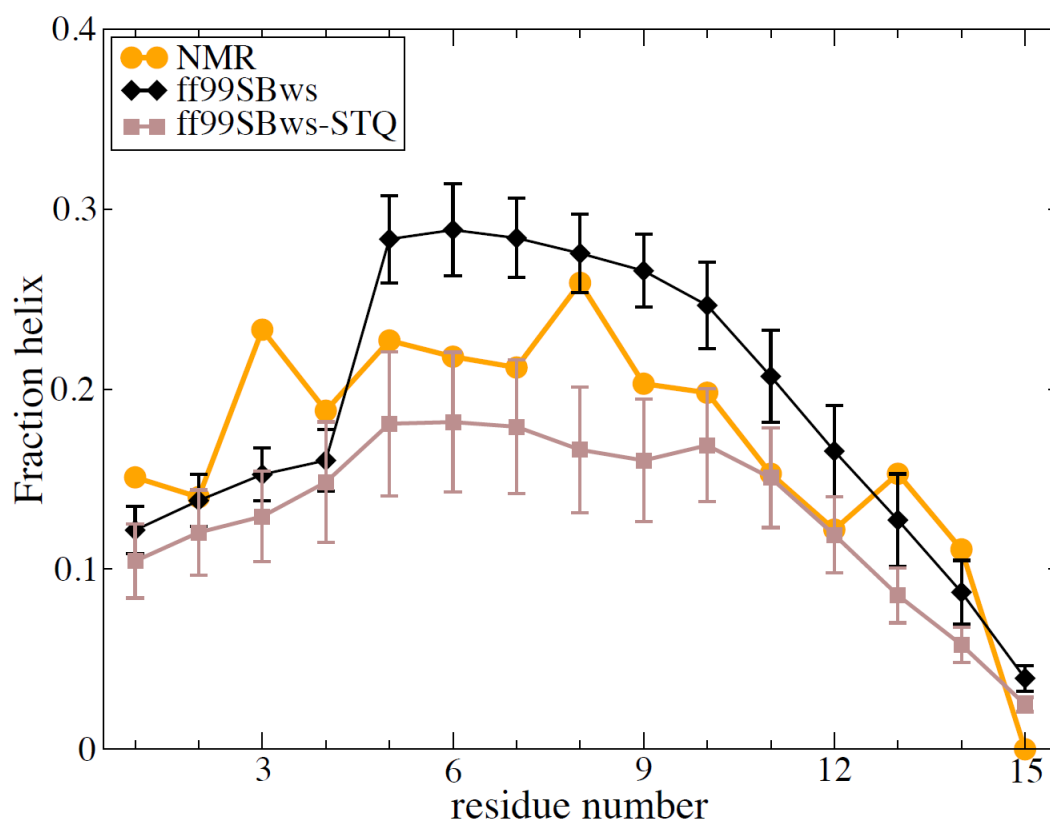


Figure A.9: Fraction Helix of (AAQAA)₃ with ff99SBws and ff99SBws-STQ (this study) compared to NMR experimental data [254].

Appendix B

Supplementary material for “SINATRA Pro: A Topological Data Analytic Approach for Discovering Biophysical Signatures in Protein Dynamics”

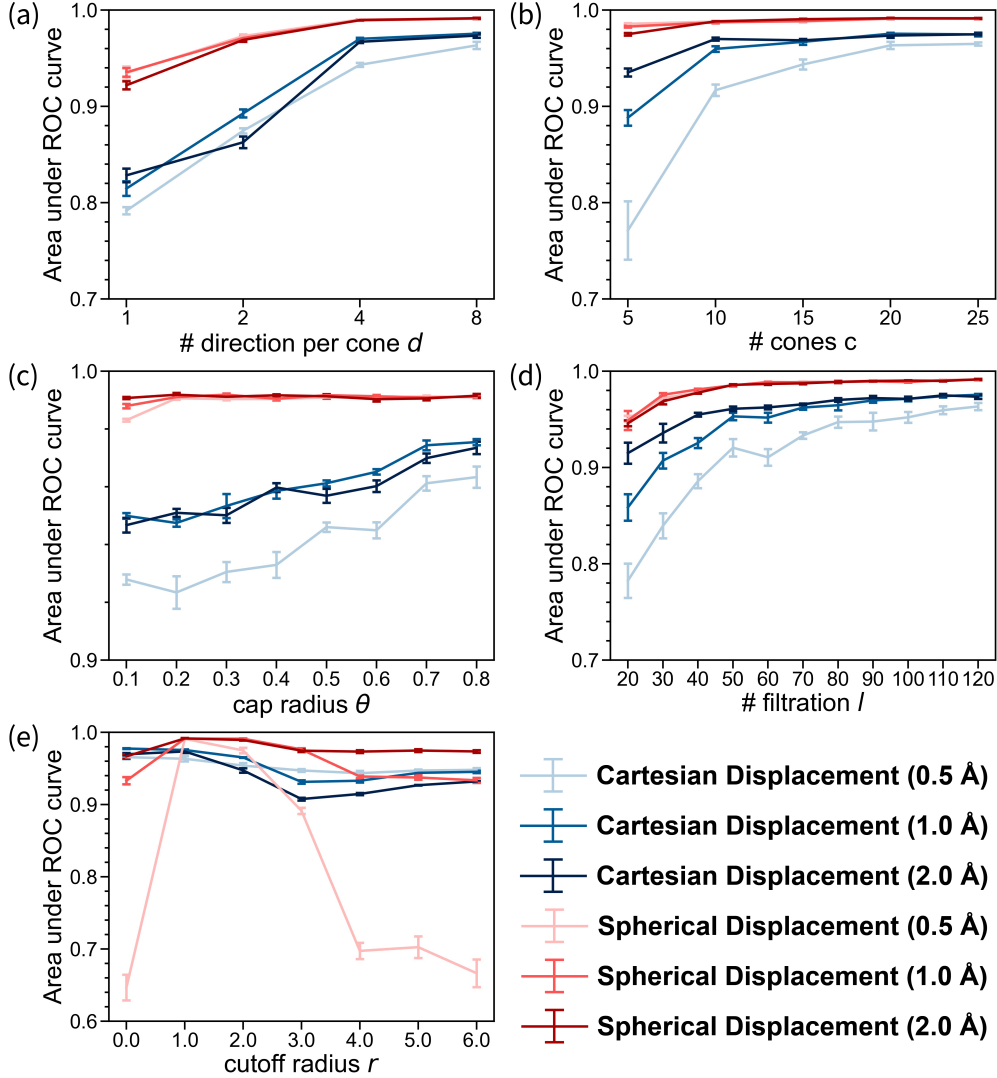


Figure B.1: Power and sensitivity analysis assessing the robustness of SINATRA Pro to different free parameter settings in controlled molecular dynamic (MD) simulations. To generate data for these simulations, we consider two phenotypic classes using real structural data of wild-type β -lactamase (TEM). In the first phenotypic class, structural protein data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). In the second phenotypic class, proteins are drawn from 0.5 ns intervals later relative to the first group (e.g., $t_{\text{MD}} = [0.5, 1.5, 2.5, 3.5, \dots, 99.5] \text{ ns} + \delta$) to introduce physical thermal noise, and then we displace the atomic positions of each atom in the Ω -loop region by (1) a constant cartesian vector of (light blue) 0.5 ångströms (Å), (blue) 1.0 Å, and (dark blue) 2.0 Å, or (2) by a spherically uniform random vector of (pink) 0.5 Å, (red) 1.0 Å, and (dark red) 2.0 Å. Altogether, we have a dataset of $N = 1000$ proteins per simulation scenario: 100 ns interval $\times$ 5 different choices $\delta = \{0.0, 0.1, 0.2, 0.3, 0.4\}$ ns $\times$ 2 phenotypic classes (original wild-type versus perturbed). The area under the curve (AUC) details the ability of SINATRA Pro to identify “true class defining” atoms located within the Ω -loop region as a function of changing the different free parameters used in the algorithm. Here, we assess the robustness of the algorithm to (a) d number of directions per cone, (b) c number of cones, (c) θ cap radius used to generate directions within each cone, (d) l number of sublevel sets (or filtration steps) used to compute the topological summary statistics, and (e) the radius cutoff r in Å used to construct the simplicial complex. While varying each parameter, the other parameters are fixed at $\{r = 1.0 \text{ Å}, c = 20, d = 8, \theta = 0.80, l = 120\}$. Guidelines for how to choose the free parameters are given in Table 4.1 in Chapter 4.

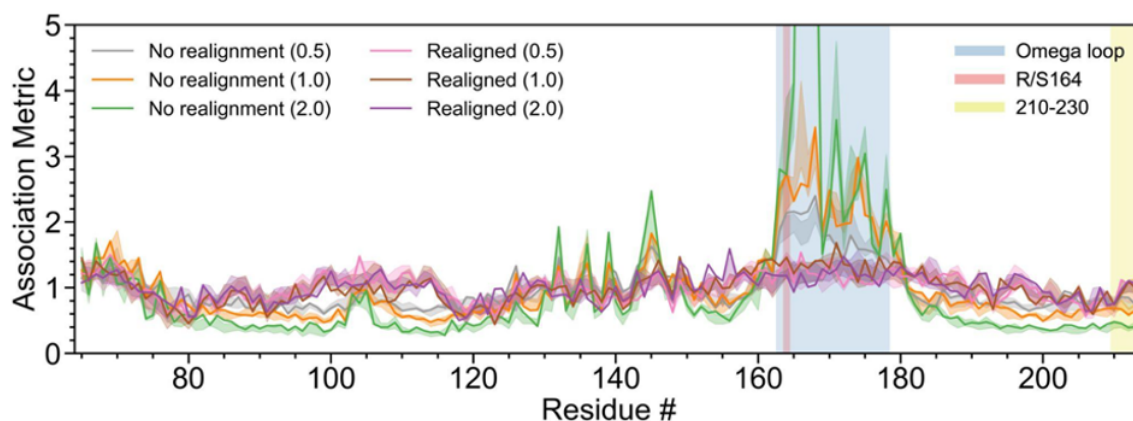


Figure B.2: Effect of realignment after displacement in the controlled experiments aimed at assessing the ability of SINATRA Pro to detect artificial changes in the Ω -loop of β -lactamase (blue region). In Chapter 4, we conduct a controlled simulation study where consider two phenotypic classes using real structural data of wild-type β -lactamase (TEM). In the first phenotypic class, structural protein data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). In the second phenotypic class, proteins are drawn from 0.5 ns intervals later relative to the first group (e.g., $t_{\text{MD}} = [0.5, 1.5, 2.5, 3.5, \dots, 99.5] \text{ ns} + \delta$) to introduce physical thermal noise. Here, we displace the atomic positions of each atom in the Ω -loop region by a constant Cartesian vector of 0.5, 1.0, and 2.0 ångströms (Å), respectively. The purpose of this figure is to show why we do not realign the proteins after structural perturbation has occurred. Realigning the structures after introducing a perturbation poses a slightly different and notably less controlled simulation study. For example, in this constant displacement case, realigning the structures will shift the whole structure against the perturbing vector and result in an unintentional displacement on the opposite side of the structure (as depicted by the pink, brown, and purple lines). A true positive in this case is not as well-defined as when we keep the unperturbed structure in place and define the perturbed structure as the ground truth for the positive signal (as shown by the grey, orange, and green lines).

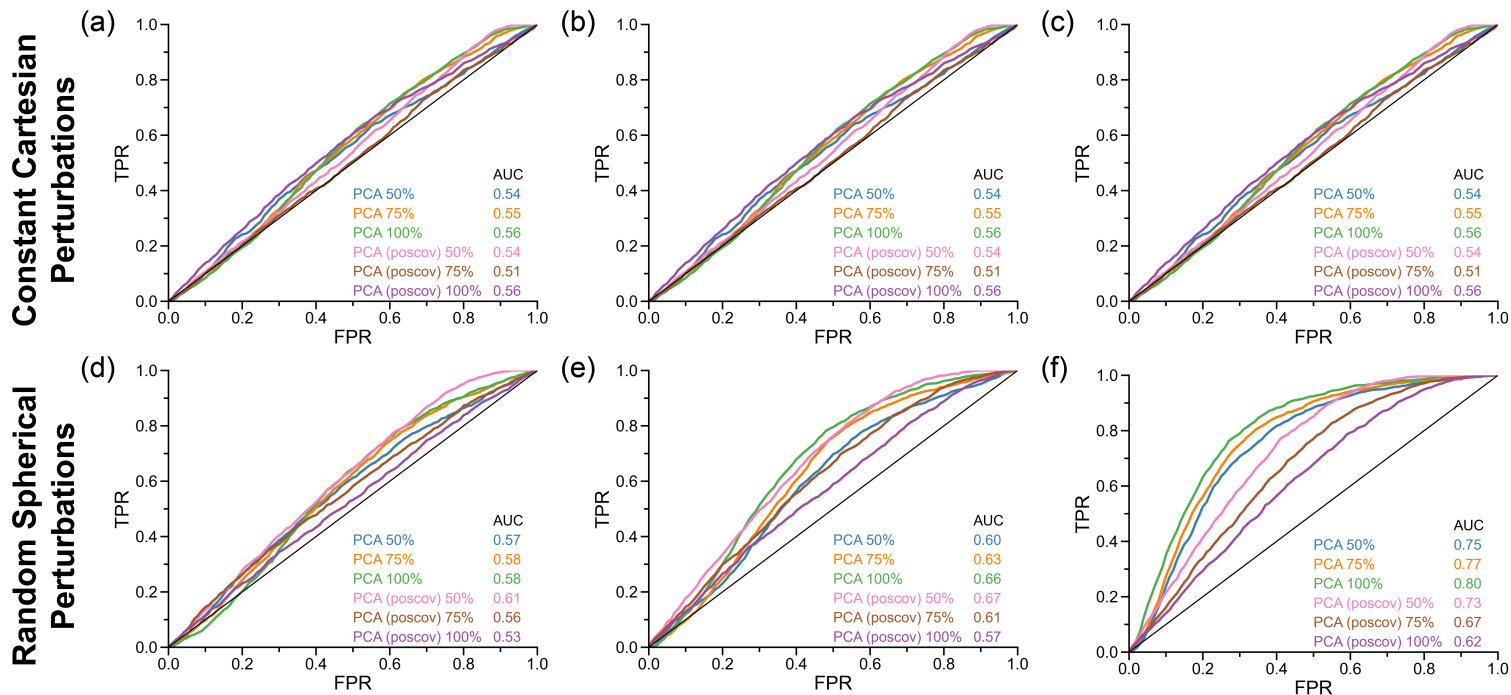


Figure B.3: Receiver operating characteristic (ROC) assessing the differentiating power of PCA under different model parameter configurations in controlled molecular dynamic (MD) simulations. Here, in addition to the Cartesian-based PCA approach described in Chapter 4, we also perform an additional PCA strategy on the covariance matrix between atomic positions. This is done by first taking the atomic positions of all frames, centering their mean to be zero, and normalizing them to have unit variance equal to one. Next, the position covariance matrix is generated between the two datasets using the function $\mathbb{V}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})^T]$ which takes on positive values if two variables are correlated and negative if two variables are anti-correlated. We then run PCA on the covariance matrix and choose the number of principal components that explain at least some percentage of its cumulative variance. Note that the output from PCA produces vectors that have dimensionality equal to the total number of atoms in the protein structures, and these can be interpreted as a measure of how explanatory each atomic position is in determining the variation between two sets of data (e.g., class *A* and *B*, respectively). In this analysis, the principal components are taken as feature vectors to generate the ROC curves, where the least correlated (or anti-correlated) variables with the class labels are considered to be the differentiating features. Above, we are ranking atoms according to the least absolute correlation. In the legend, the approach used to in Chapter 4 is simply listed as PCA, while the method run on the positional covariance matrix is indicated by `poscov`. Each method is evaluated using a different number of principal components based on the cumulative variance explained: 50%, 75%, and 100%. We then use area under the curves (AUC) to summarize this performance. Both strategies show similar performance across all simulation scenarios in the constant perturbations, while the original PCA approach that explains 100% of the variation over the Cartesian (x, y, z) -coordinates for the atoms is best in the simulations with random spherical perturbations. The latter is what is displayed in Chapter 4.

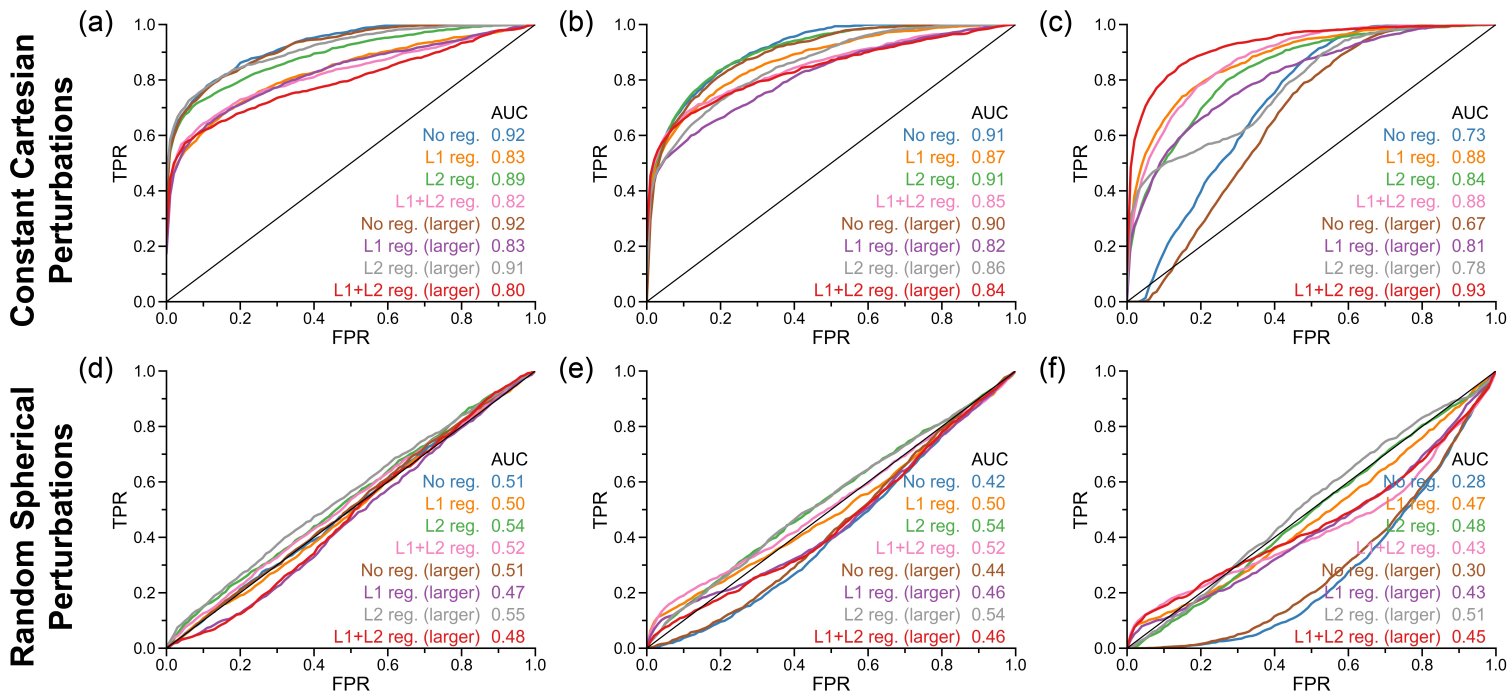


Figure B.4: Receiver operating characteristic (ROC) assessing the differentiating power of the Neural Network under different model parameter and architecture configurations in controlled molecular dynamic (MD) simulations. Here, in addition to the Neural Network described in Chapter 4 (depicted in blue), we also perform an additional search over different architectures and model training procedures. For the former, we try deepening the architecture with Rectified Linear Unit (ReLU) nonlinear activation functions [163] to the following: (1) an input layer of Cartesian coordinates of all of the atoms; (2) a hidden layer with $H = 2048$ neurons; (3) a second hidden layer with $H = 2048$ neurons; (4) a third hidden layer with $H = 512$ neurons; (5) a third hidden layer with $H = 128$ neurons; and (6) an outer layer with a single node which uses a sigmoid link function for protein classification. For the latter modification, we try regularizing the network weights using a combination of L_1 , L_2 , and $L_1 + L_2$ penalties. These correspond to the “Least Absolute Shrinkage and Selection Operator” or LASSO solution [255], Ridge Regression [256], and the Elastic Net [257], respectively. Once again, batch normalization was implemented between each layer and a normalized saliency map to rank the importance of each atom [164]. We assess power by taking the sum of the saliency map values corresponding to each atomic position which is summarized by the area under the curves (AUC).

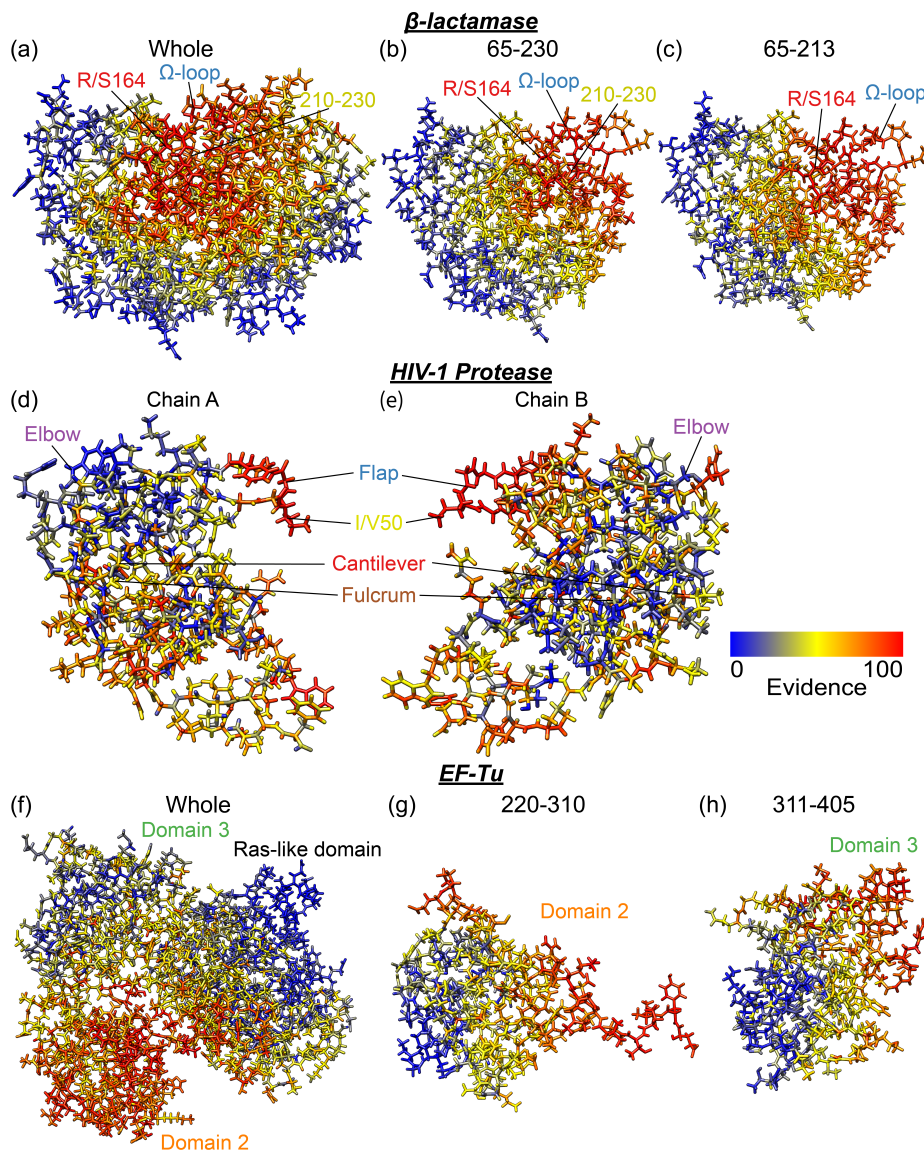


Figure B.5: Atomic-level results for detecting biophysical signatures in (top row) TEM β -lactamase, (middle row) HIV-1 protease bound to Amprenavir, and (bottom row) GTP-bound EF-Tu. In these analyses, we compare the molecular dynamics (MD) trajectories of alternative states for each protein to the corresponding trajectories of (i) Arg164Ser TEM β -lactamase, (ii) Ile50Val HIV-1 protease bound to Amprenavir, and (iii) GDP-bound EF-Tu, respectively. We analyze datasets based on different fragments of each protein. Specifically, in the case of TEM β -lactamase, we analyze (a) the whole protein structure, (b) residues 65-230, and (c) residues 65-213; in HIV-1 protease, we analyze (d) chain A and (e) chain B; and, in EF-Tu, we analyze (f) the whole protein structure, (g) residues 220-310, and (h) residues 311-405. Here, consistency between fragments within a protein type shows the robustness of SINATRA Pro to identify the same signal even when it does not have access to the full structure. The heatmaps highlight the atomic evidence potential on a scale from [0 – 100]. A maximum of 100 represents the threshold at which the first atom of the protein is reconstructed, while 0 denotes the threshold when the last atom is reconstructed. Annotated are regions of interest (ROIs) according to literature sources that have previously suggested some level of structural association for each chemical change of interest, including: (i) the Ω -loop (residues 163-178) in TEM; (ii) the flap region (residues 47-55) in HIV-1 protease; and (iii) Domain 2 (residues 208-308) in EF-Tu.

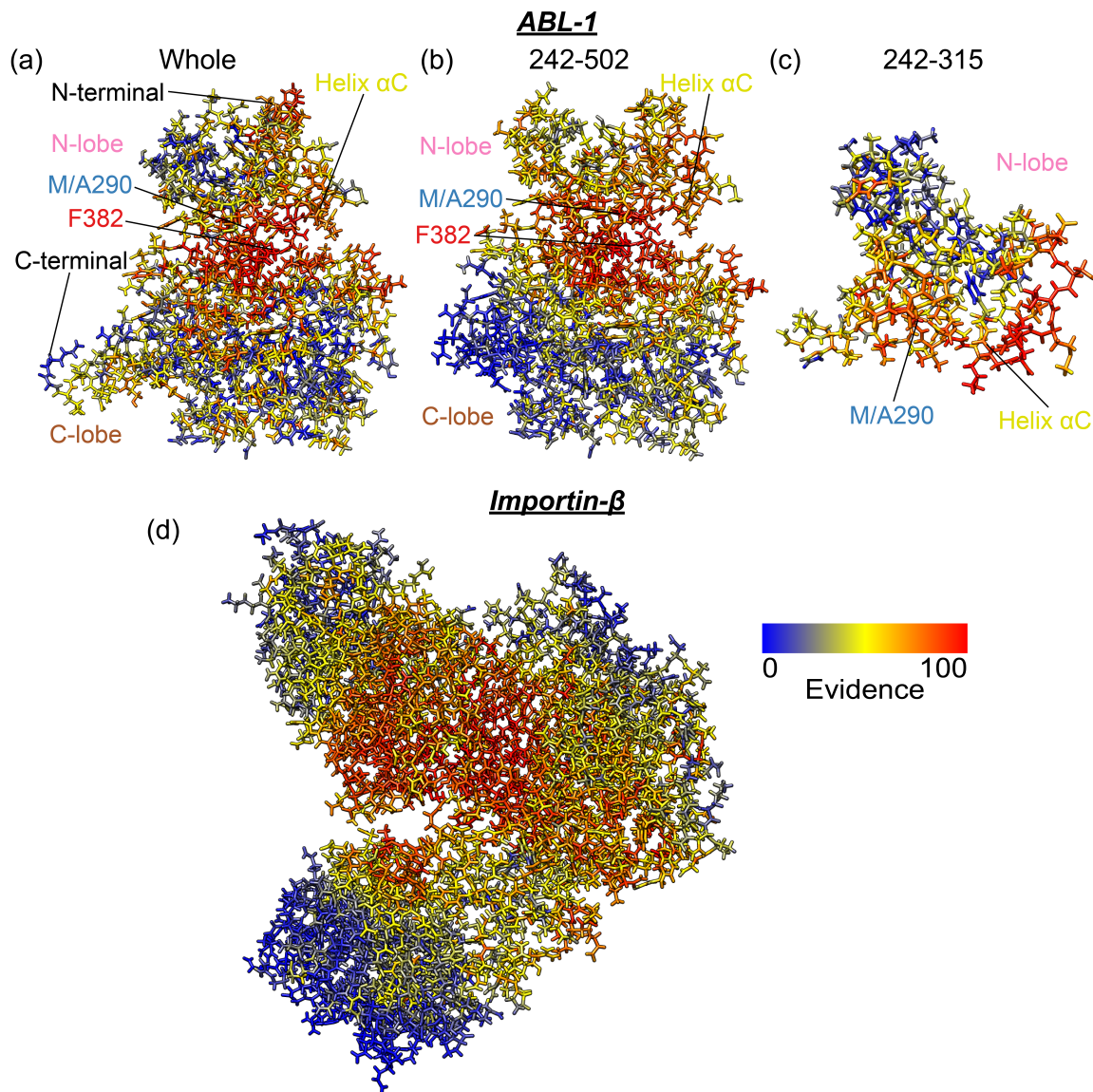


Figure B.6: Atomic-level results for detecting biophysical signatures in (top row) Abl1 and (bottom row) IBB-bound Importin- β . In these analyses, we compare the molecular dynamics (MD) trajectories of the alternative states for each protein to the corresponding trajectories of (i) Met290Val Abl1 and (ii) unbound Importin- β , respectively. We analyze Abl1 based on different fragments the protein. Specifically, we analyze (a) the whole protein structure, (b) residues 242-502, and (c) residues 242-315. Here, consistency between fragments within a protein type shows the robustness of SINATRA Pro to identify the same signal even when it does not have access to the full structure. The heatmaps highlight the atomic evidence potential on a scale from [0 – 100]. A maximum of 100 represents the threshold at which the first atom of the protein is reconstructed, while 0 denotes the threshold when the last atom is reconstructed. Annotated are regions of interest (ROIs) according to literature sources that have previously suggested some level of structural association for each chemical change of interest, including the DFG motif in Abl1. Note that, in the context of Importin- β , the superhelix includes the entire structure and so we do not include any additional annotations.

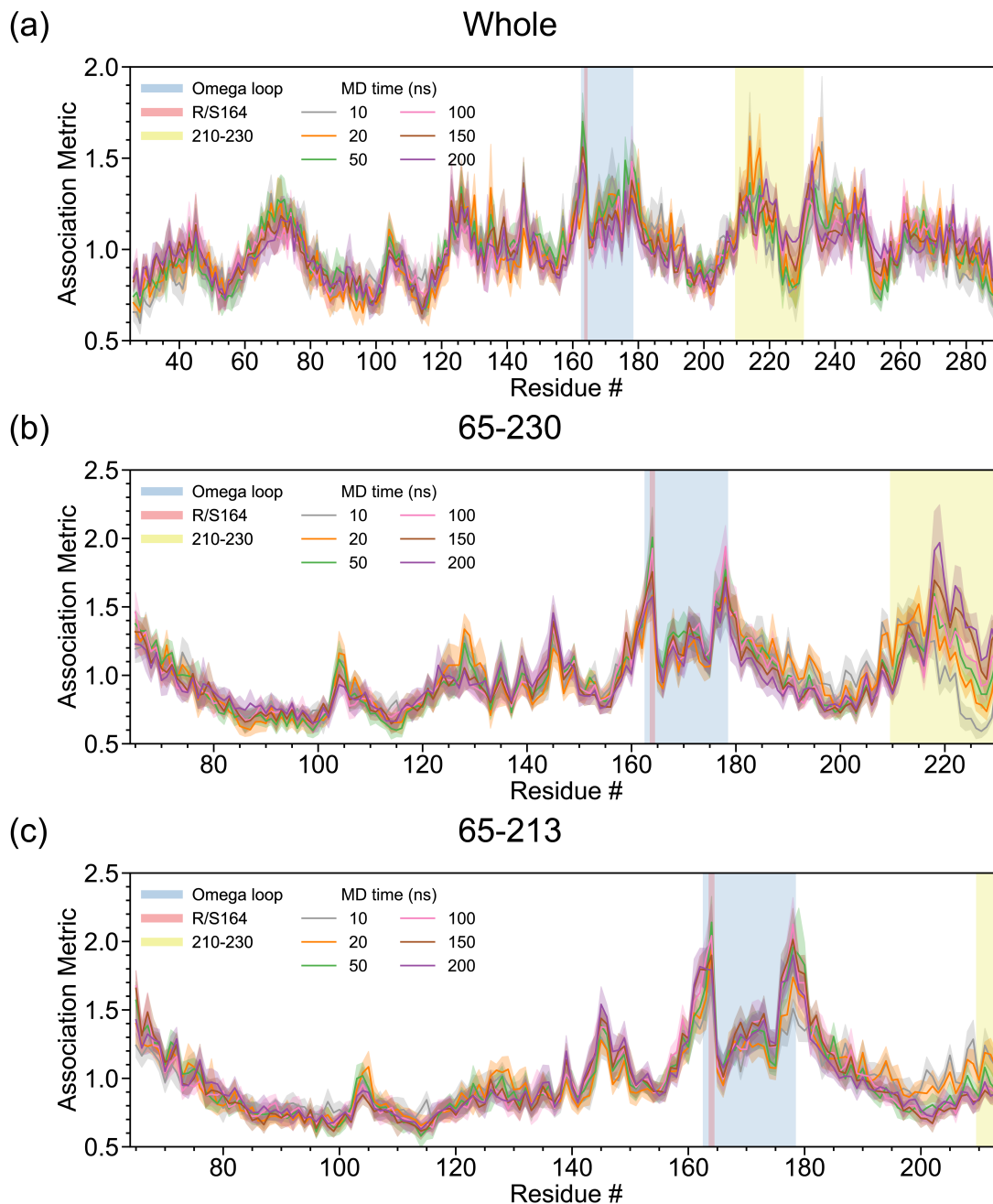


Figure B.7: Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in the Ω -loop of TEM β -lactamase induced by the Arg164Ser mutation using SINATRA Pro. In this analysis, we compare the molecular dynamics (MD) trajectories of wild-type TEM β -lactamase versus the Arg164Ser mutant [180, 181]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 10 ns (grey), 20 ns (orange), 50 ns (green), 100 ns (pink), 150 ns (brown), and 200 ns (purple) MD trajectory. As an example of how data are sampled, in the 150 ns simulation case, we have $t_{\text{MD}} = [0, 1.5, 3, \dots, 148.5] \text{ ns} + \delta$, where $\delta = \{0.0, 0.15, 0.3, \dots, 1.35\} \text{ ns}$ is a time offset parameter. Panels (a)-(c) show the mean association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods) with the (a) whole protein, (b) fragment 65-230, and (c) fragment 65-213. The overlap of lines shows the robustness of SINATRA Pro to identify the same signal regardless of trajectory length.

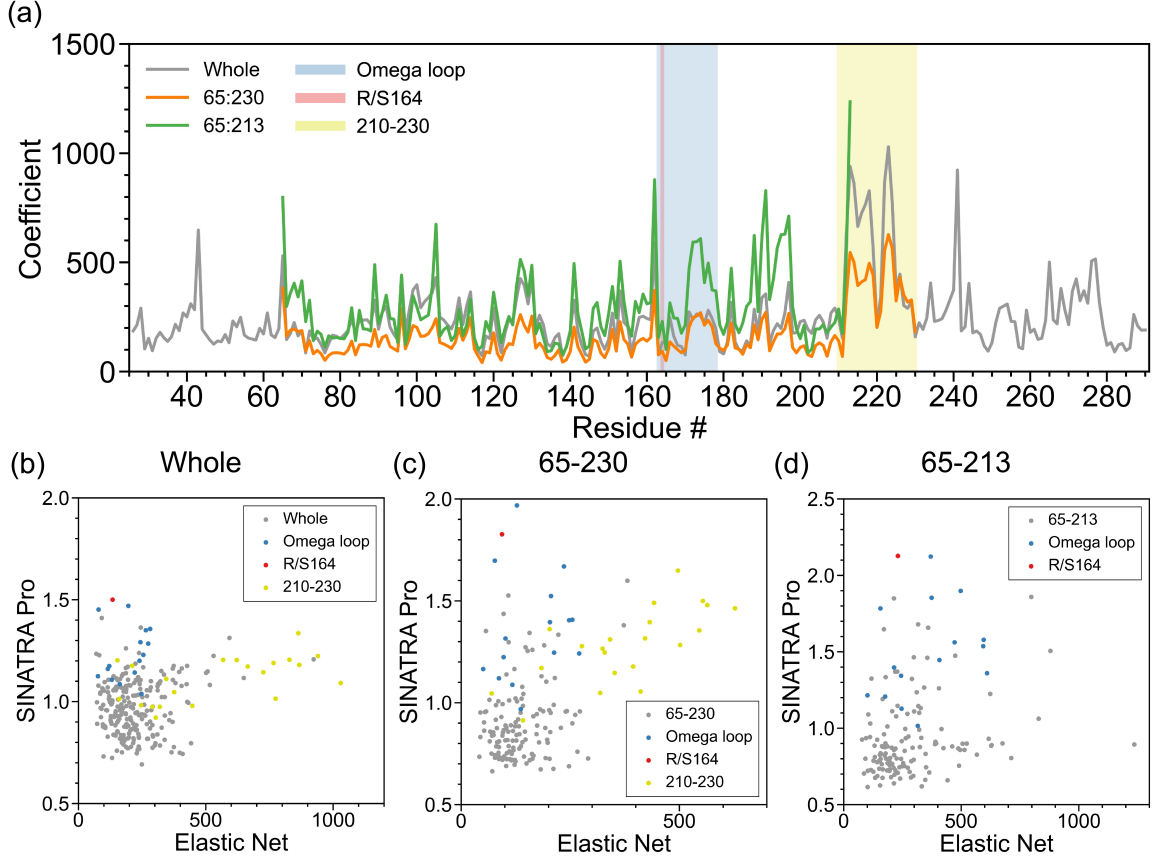


Figure B.8: Real data analyses aimed at detecting structural changes in the Ω -loop of TEM β -lactamase induced by the Arg164Ser mutation using atomic-level regularization with Elastic Net classification. In this analysis, we compare the molecular dynamics (MD) trajectories of wild-type TEM β -lactamase versus the Arg164Ser mutant [180, 181]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). To generate these results, we first concatenate the (x, y, z) -coordinates of all atoms within each protein and treat them as features in a data frame. Next, we use Elastic Net regularization [257] to assign sparse regression coefficients to each coordinate of every atom (where the penalization term is chosen via cross-validation). Panel (a) shows the mean absolute coefficient of all atoms within each residue computed over each fragment-based analysis (see Material and Methods in Chapter 4). The final row plots the correlation between the SINATRA Pro association metrics and the Elastic Net coefficients for all atoms with correspondences in the (b) whole protein, (c) fragment 65-230, and (d) fragment 65-213.

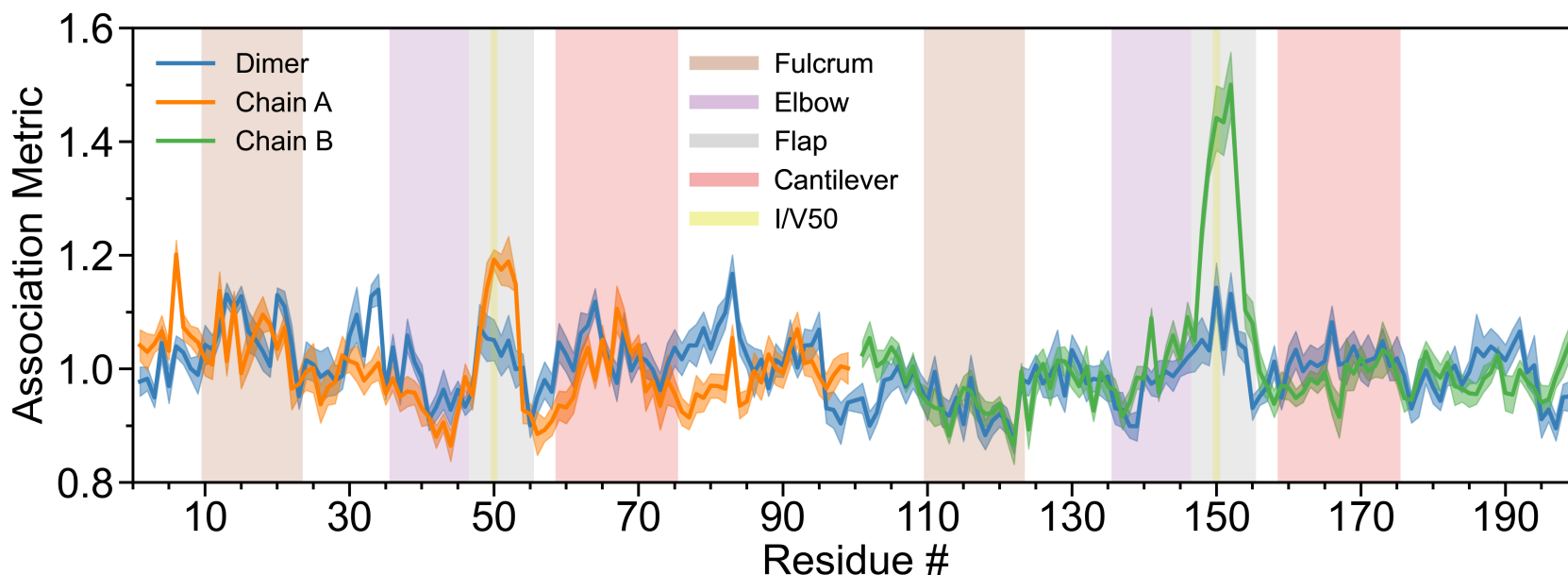


Figure B.9: SINATRA Pro analysis on the dimeric form of HIV-1 protease to detect structural change in the flap region driven by a Ile50Val mutation. In this analysis, we compare the molecular dynamic (MD) trajectories of wild-type HIV-1 protease versus Ile50Val mutants (i.e., within residues 47-55). For both phenotypic classes, structural data are drawn from from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns}$ $\times$ 2 phenotypic classes (wild-type versus mutant). This figure depicts results after applying SINATRA Pro using parameters $\{r = 6.0 \text{ \AA}, c = 20, d = 8, \theta = 0.80, l = 120\}$ chosen via a grid search. We compare these results to the analyses with chains A and B presented in Chapter 4. Here, the signal in the region of interest (i.e., the flap of each chain) persisted in the dimeric form, but is overshadowed by the noise in the other parts of the protein since the relative orientation between the two monomers causes each chain to be misaligned with itself. Highlighted are residues for regions of the protein corresponding to the fulcrum (brown), elbow (purple), flap (blue), cantilever (red), and I/V50 (yellow) [188, 190, 191].

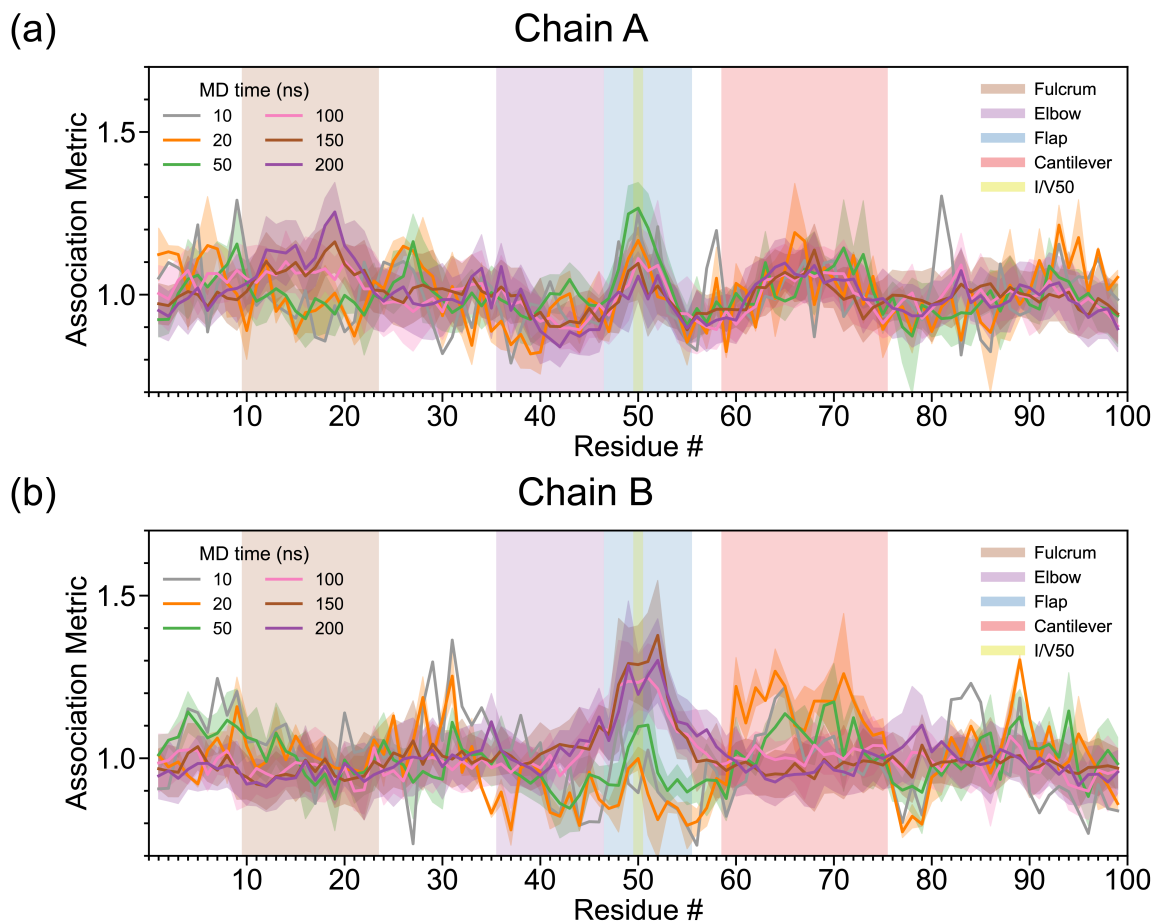


Figure B.10: Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in the flap region of HIV-1 protease driven by the Ile50Val mutation using SINATRA Pro. In this analysis, we compare the molecular dynamics (MD) trajectories of wild-type HIV-1 protease versus the Ile50Val mutant (i.e., within residues 47-55) [188, 190, 191]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 10 ns (grey), 20 ns (orange), 50 ns (green), 100 ns (pink), 150 ns (brown), and 200 ns (purple) MD trajectory. As an example of how data are sampled, in the 150 ns simulation case, we have $t_{\text{MD}} = [0, 1.5, 3, \dots, 148.5] \text{ ns} + \delta$, where $\delta = \{0.0, 0.15, 0.3, \dots, 1.35\} \text{ ns}$ is a time offset parameter. Panels (a)-(c) show the mean association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods) with (a) chain A and (b) chain B, respectively. The overlap of lines shows the robustness of SINATRA Pro to identify the same signal regardless of trajectory length.

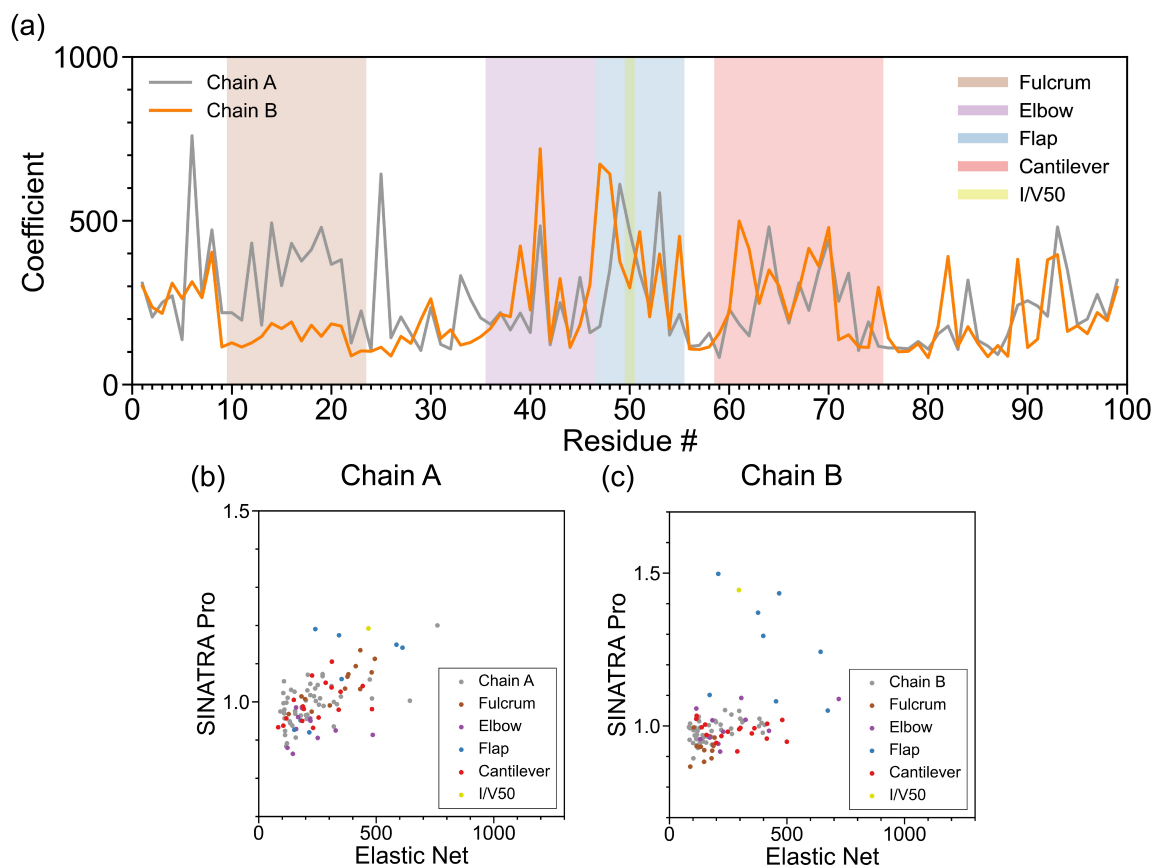


Figure B.11: Real data analyses aimed at detecting structural changes in the flap region of HIV-1 protease driven by the Ile50Val mutation using atomic-level regularization with Elastic Net classification. In this analysis, we compare the molecular dynamics (MD) trajectories of wild-type HIV-1 protease versus the Ile50Val mutant (i.e., within residues 47-55) [188, 190, 191]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). To generate these results, we first concatenate the (x, y, z) -coordinates of all atoms within each protein and treat them as features in a data frame. Next, we use Elastic Net regularization [257] to assign sparse regression coefficients to each coordinate of every atom (where the penalization term is chosen via cross-validation). Panel (a) shows the mean absolute coefficient of all atoms within each residue computed over each fragment-based analysis (see Material and Methods in Chapter 4). The final row plots the correlation between the SINATRA Pro association metrics and the Elastic Net coefficients for all atoms with correspondences in (b) chain A and (c) chain B, respectively.

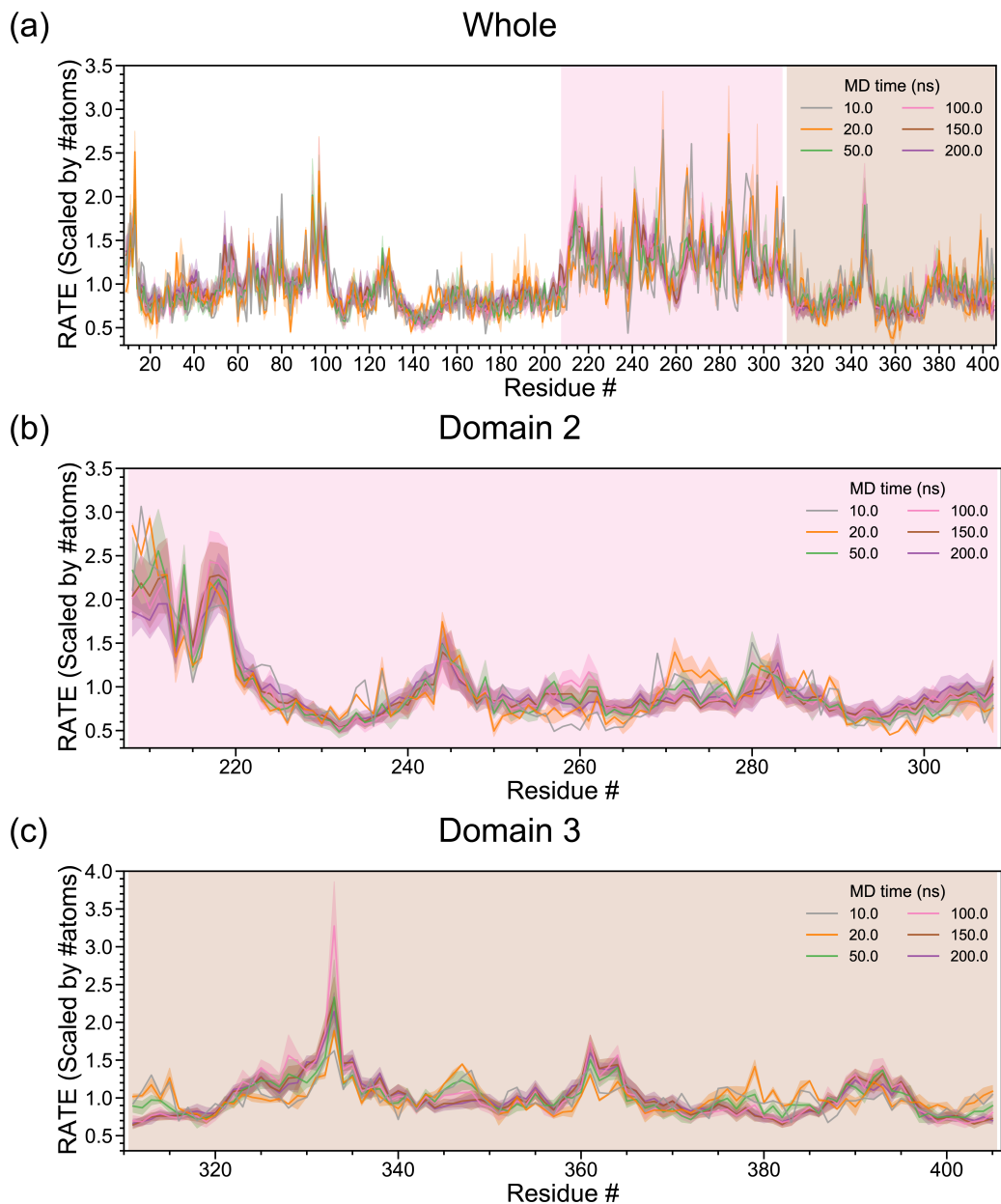


Figure B.12: Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in Domain 2 of the elongation factor EF-Tu upon guanosine triphosphate (GTP) hydrolysis using SINATRA Pro. In this analysis, we compare the the molecular dynamics (MD) trajectories of GTP-bound EF-Tu versus GDP-bound EF-Tu [196, 213, 214]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 10 ns (grey), 20 ns (orange), 50 ns (green), 100 ns (pink), 150 ns (brown), and 200 ns (purple) MD trajectory. As an example of how data are sampled, in the 150 ns simulation case, we have $t_{\text{MD}} = [0, 1.5, 3, \dots, 148.5] \text{ ns} + \delta$, where $\delta = \{0.0, 0.15, 0.3, \dots, 1.35\} \text{ ns}$ is a time offset parameter. Panels (a)-(c) show the mean association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods) with the (a) whole protein, (b) fragment 208-308, and (c) fragment 311-405. The overlap of lines shows the robustness of SINATRA Pro to identify the same signal regardless of trajectory length.

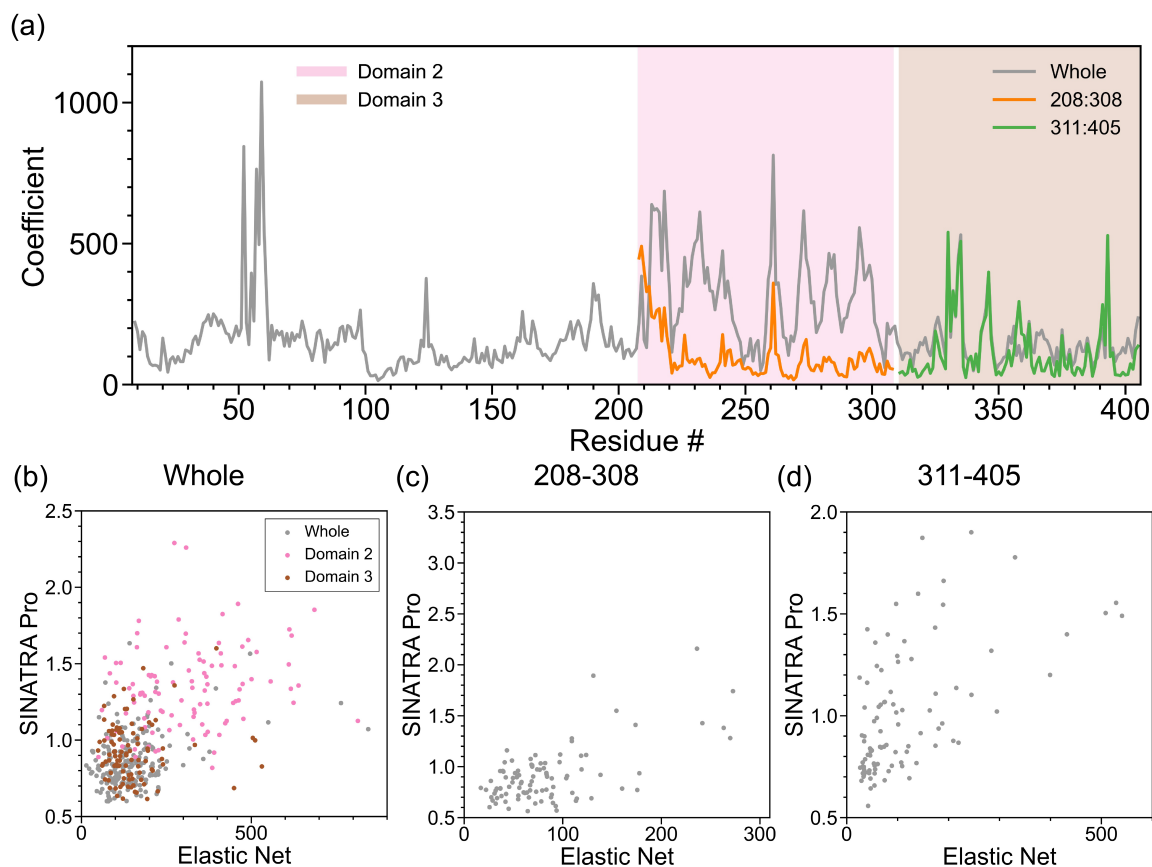


Figure B.13: Real data analyses aimed at detecting structural changes in Domain 2 of the elongation factor EF-Tu upon guanosine triphosphate (GTP) hydrolysis using atomic-level regularization with Elastic Net classification. In this analysis, we compare the molecular dynamics (MD) trajectories of GTP-bound EF-Tu versus GDP-bound EF-Tu [196, 213, 214]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). To generate these results, we first concatenate the (x, y, z) -coordinates of all atoms within each protein and treat them as features in a data frame. Next, we use Elastic Net regularization [257] to assign sparse regression coefficients to each coordinate of every atom (where the penalization term is chosen via cross-validation). Panel (a) shows the mean absolute coefficient of all atoms within each residue computed over each fragment-based analysis (see Material and Methods in Chapter 4). The final row plots the correlation between the SINATRA Pro association metrics and the Elastic Net coefficients for all atoms with correspondences in the (b) whole protein, (c) fragment 220-310, and (d) fragment 311-405.

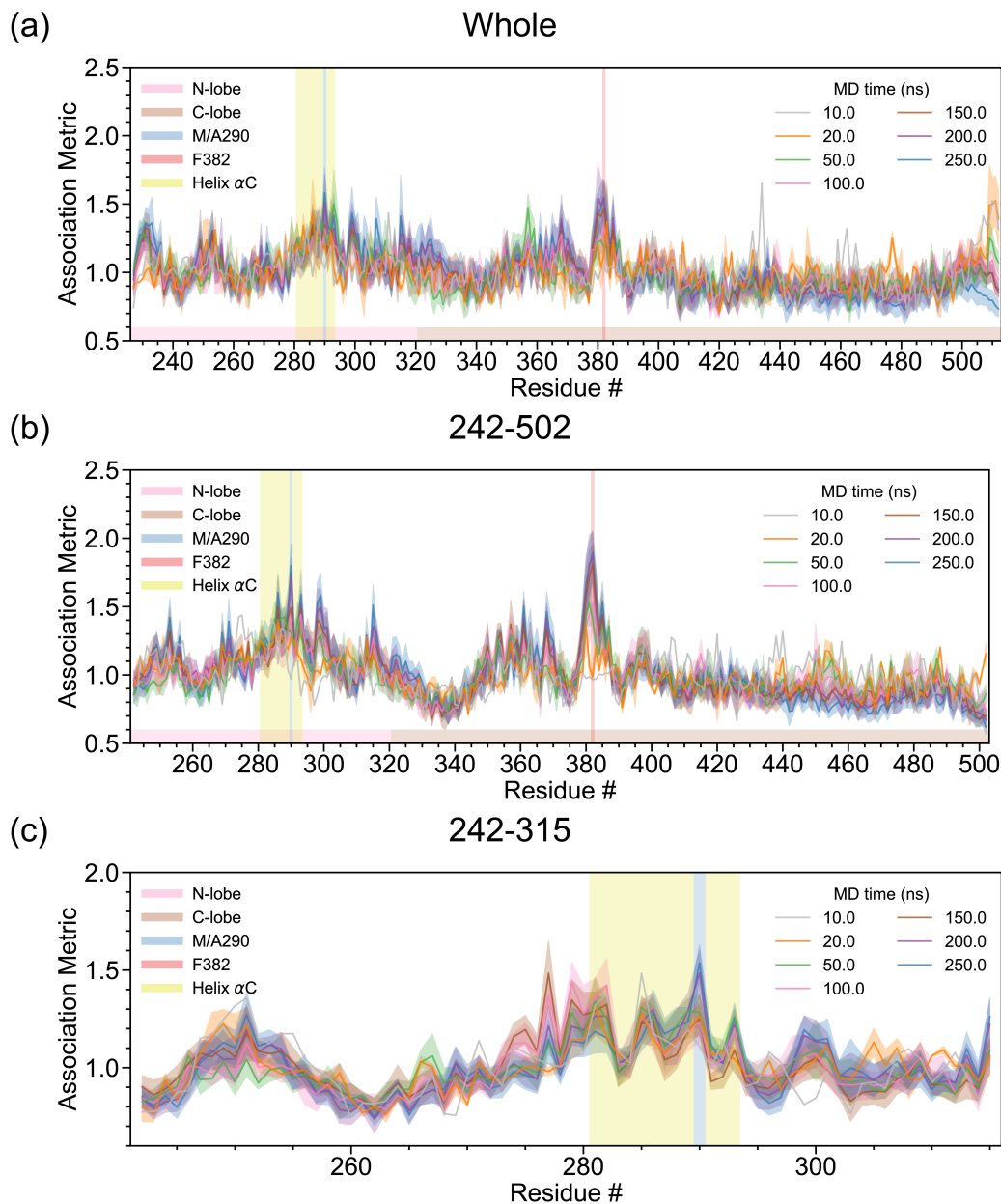


Figure B.14: Sensitivity analyses on different lengths of MD simulations aimed at detecting consistent structural changes in the N-terminal pocket of the Abl1 Tyrosine protein kinase due to the Met290Val mutation in the α C helix using SINATRA Pro. In this analysis, we compare the molecular dynamics (MD) trajectories of wild-type Abl1 kinase domain versus the Met290Val mutant [128, 202, 215–217]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 10 ns (grey), 20 ns (orange), 50 ns (green), 100 ns (pink), 150 ns (brown), and 200 ns (purple) MD trajectory. As an example of how data are sampled, in the 150 ns simulation case, we have $t_{\text{MD}} = [0, 1.5, 3, \dots, 148.5] \text{ ns} + \delta$, where $\delta = \{0.0, 0.15, 0.3, \dots, 1.35\} \text{ ns}$ is a time offset parameter. Panels (a)-(c) show the mean association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods) with the (a) whole protein, (b) fragment 242-502, and (c) fragment 242-315. The overlap of lines shows the robustness of SINATRA Pro to identify the same signal regardless of trajectory length.

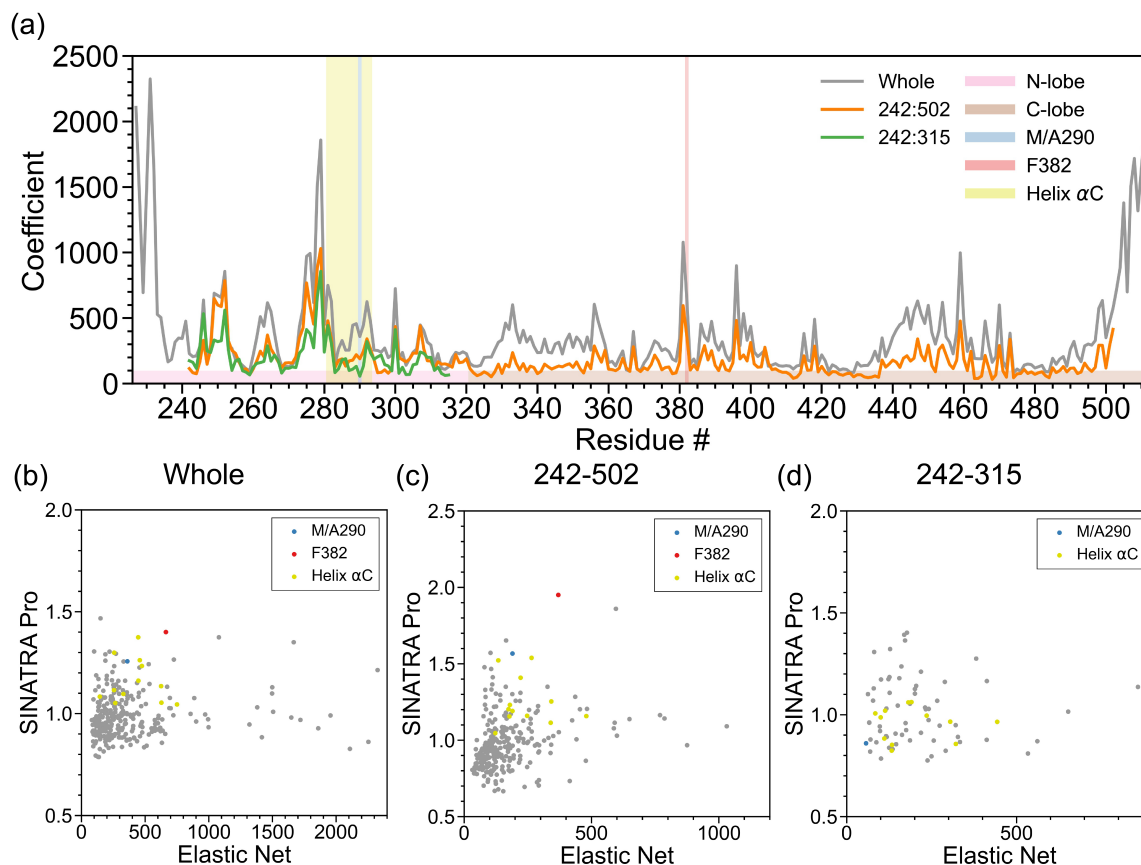


Figure B.15: Real data analyses aimed at detecting structural changes in the N-terminal pocket of the Abl1 Tyrosine protein kinase due to the Met290Val mutation in the α C helix using atomic-level regularization with Elastic Net classification. In this analysis, we compare the molecular dynamics (MD) trajectories of wild-type Abl1 kinase domain versus the Met290Val mutant. [128, 202, 215–217]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 150 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \times 1.5 \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 3000$ proteins in the study: 150 ns long interval $\times$ 15 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 1.4\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). To generate these results, we first concatenate the (x, y, z) -coordinates of all atoms within each protein and treat them as features in a data frame. Next, we use Elastic Net regularization [257] to assign sparse regression coefficients to each coordinate of every atom (where the penalization term is chosen via cross-validation). Panel (a) shows the mean absolute coefficient of all atoms within each residue computed over each fragment-based analysis (see Material and Methods in Chapter 4). The final row plots the correlation between the SINATRA Pro association metrics and the Elastic Net coefficients for all atoms with correspondences in the (b) whole protein, (c) fragment 242-502, and (d) fragment 242-315.

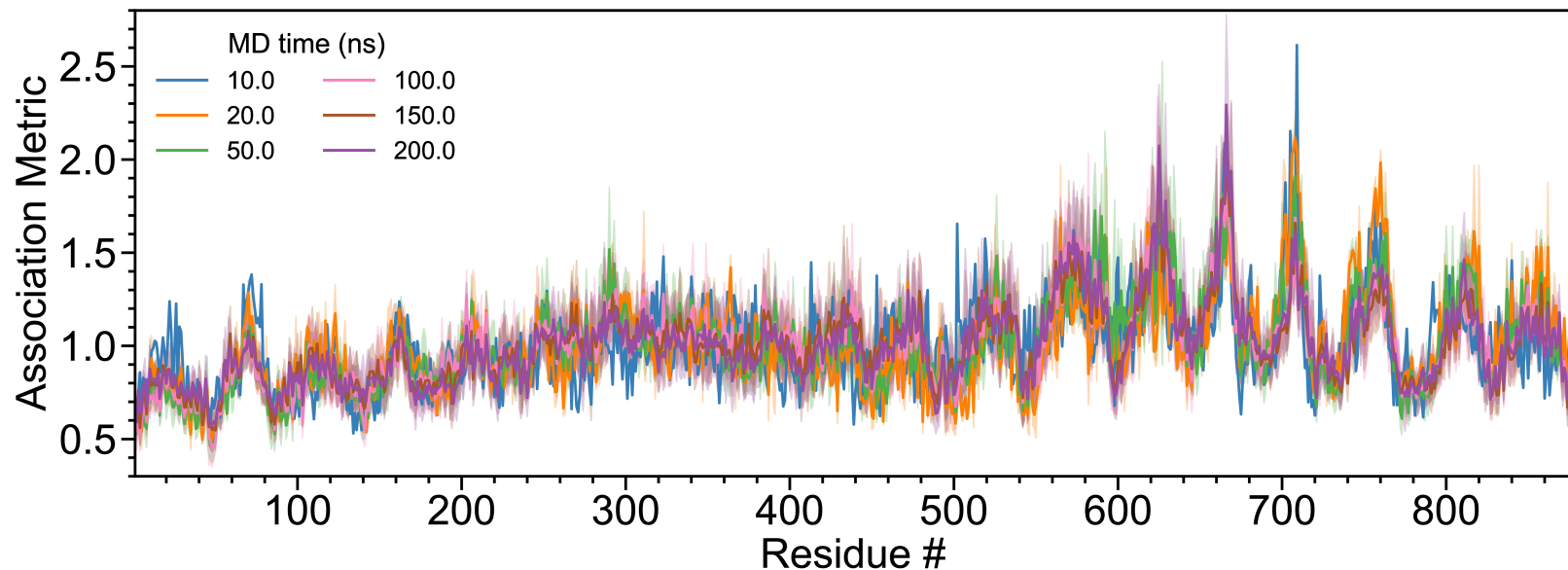


Figure B.16: Sensitivity analyses on different lengths of MD simulations aimed at detecting uncoiling of the superhelix in Importin- β upon release of an IBB peptide using SINATRA Pro. In this analysis, we compare the the molecular dynamics (MD) trajectories of IBB-bound Importin- β versus unbound Importin- β [204–206]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 10 ns (grey), 20 ns (orange), 50 ns (green), 100 ns (pink), 150 ns (brown), and 200 ns (purple) MD trajectory. As an example of how data are sampled, in the 150 ns simulation case, we have $t_{\text{MD}} = [0, 1.5, 3, \dots, 148.5] \text{ ns} + \delta$, where $\delta = \{0.0, 0.15, 0.3, \dots, 1.35\} \text{ ns}$ is a time offset parameter. Panels (a)-(c) show the mean association metrics (and their corresponding standard errors) computed for each residue within each analysis (see Material and Methods). The overlap of lines shows the robustness of SINATRA Pro to identify the same signal regardless of trajectory length.

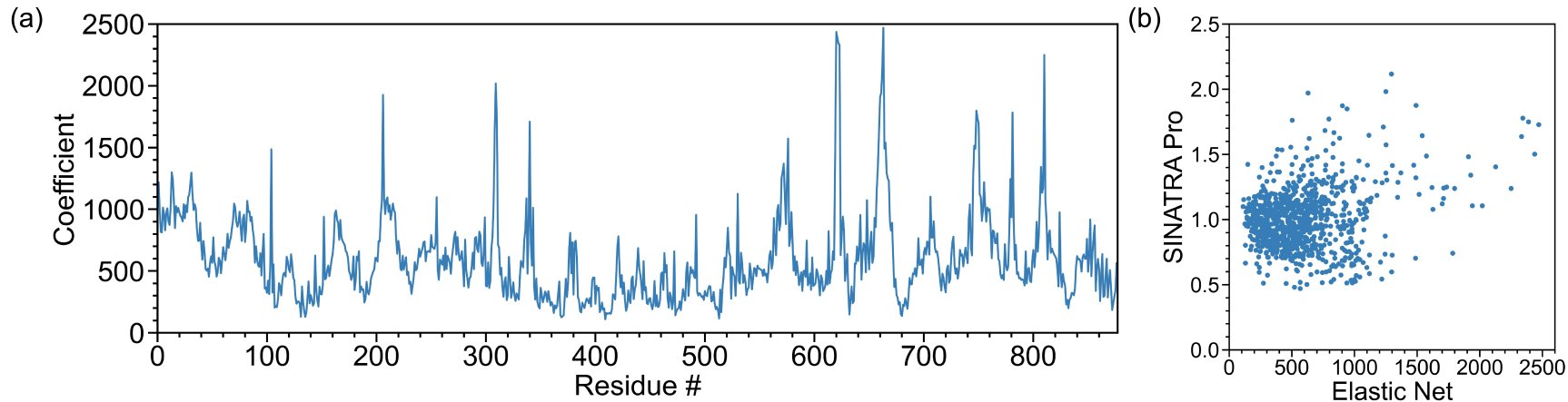


Figure B.17: Real data analyses at detecting uncoiling of the superhelix in Importin- β upon release of IBB using atomic-level regularization with Elastic Net classification. In this analysis, we compare the molecular dynamics (MD) trajectories of IBB-bound Importin- β versus unbound Importin- β [204–206]. For both phenotypic classes, structural data are drawn from equally spaced intervals over a 100 ns MD trajectory (e.g., $t_{\text{MD}} = [0, 1, 2, 3, \dots, 99] \text{ ns} + \delta$, where δ is a time offset parameter). Altogether, we have a final dataset of $N = 2000$ proteins in the study: 100 ns long interval $\times$ 10 different choices $\delta = \{0.0, 0.1, 0.2, \dots, 0.9\} \text{ ns} \times 2$ phenotypic classes (wild-type versus mutant). To generate these results, we first concatenate the (x, y, z) -coordinates of all atoms within each protein and treat them as features in a data frame. Next, we use Elastic Net regularization [257] to assign sparse regression coefficients to each coordinate of every atom (where the penalization term is chosen via cross-validation). Panel (a) shows the mean absolute coefficient of all atoms within each residue computed over each fragment-based analysis (see Material and Methods in Chapter 4). In panel (b), we plot the correlation between the SINATRA Pro association metrics and the Elastic Net coefficients for all atoms with correspondences.

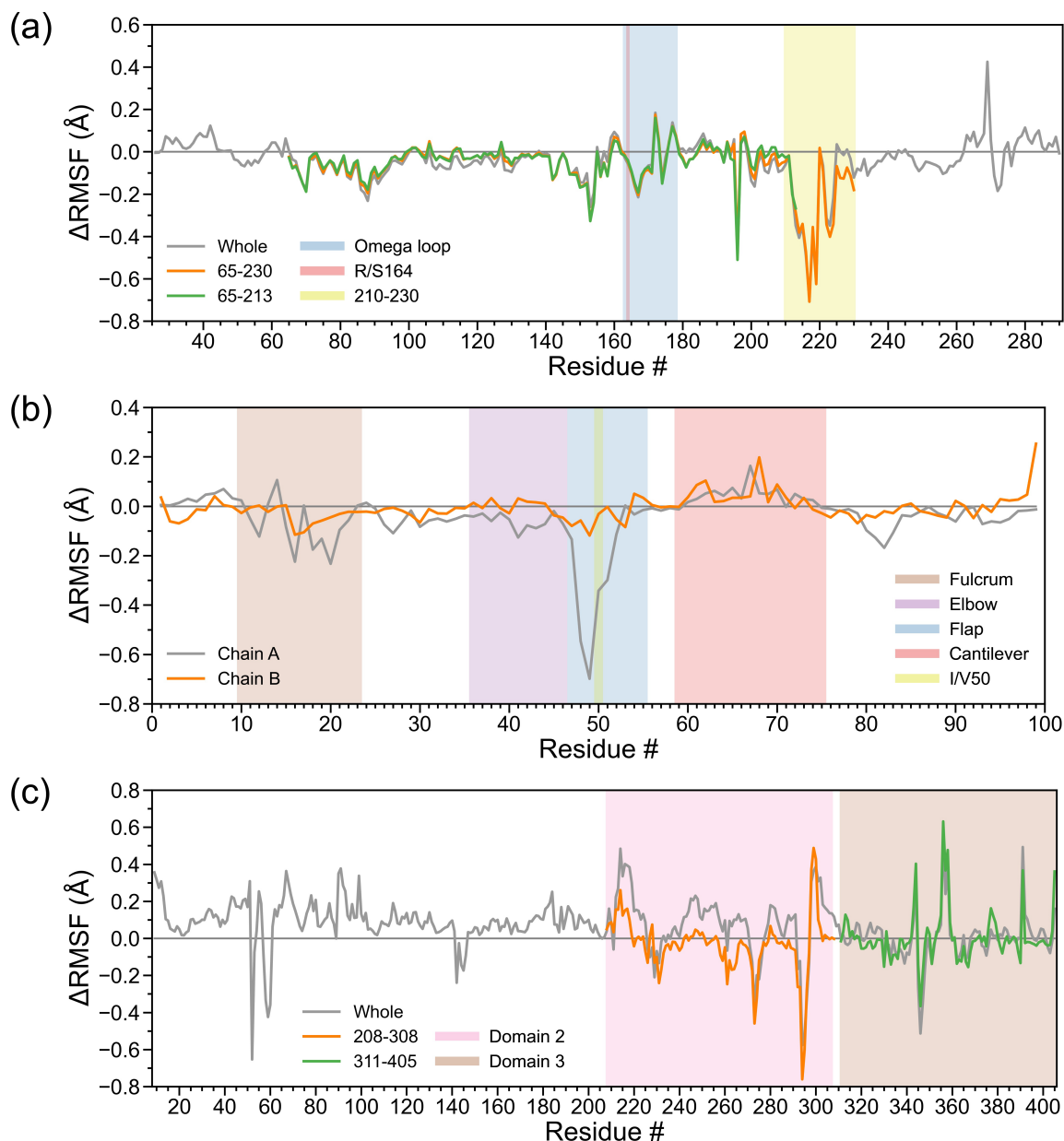


Figure B.18: Residue-level results from running root mean square fluctuation (RMSF) analysis on TEM β -lactamase, HIV-1 protease, and GTP-bound EF-Tu. In these analyses, we compare the molecular dynamics (MD) trajectories of the alternative state for each protein to the corresponding trajectories of (a) Arg164Ser mutant β -lactamase, (b) Ile50Val mutant HIV-1 protease, and (c) GDP-bound EF-Tu, respectively. We analyze datasets based on different fragments of each protein. Specifically, (a) in TEM β -lactamase, we analyze the whole protein structure, residues 65-230, and residues 65-213; (b) in HIV-1 protease, we analyze chain A and chain B; and, (c) in EF-Tu, we analyze the whole protein structure, residues 220-310, and residues 311-405. The y-axis denotes the absolute difference (or Δ -change) in RMSF between wild-type and mutants classes.

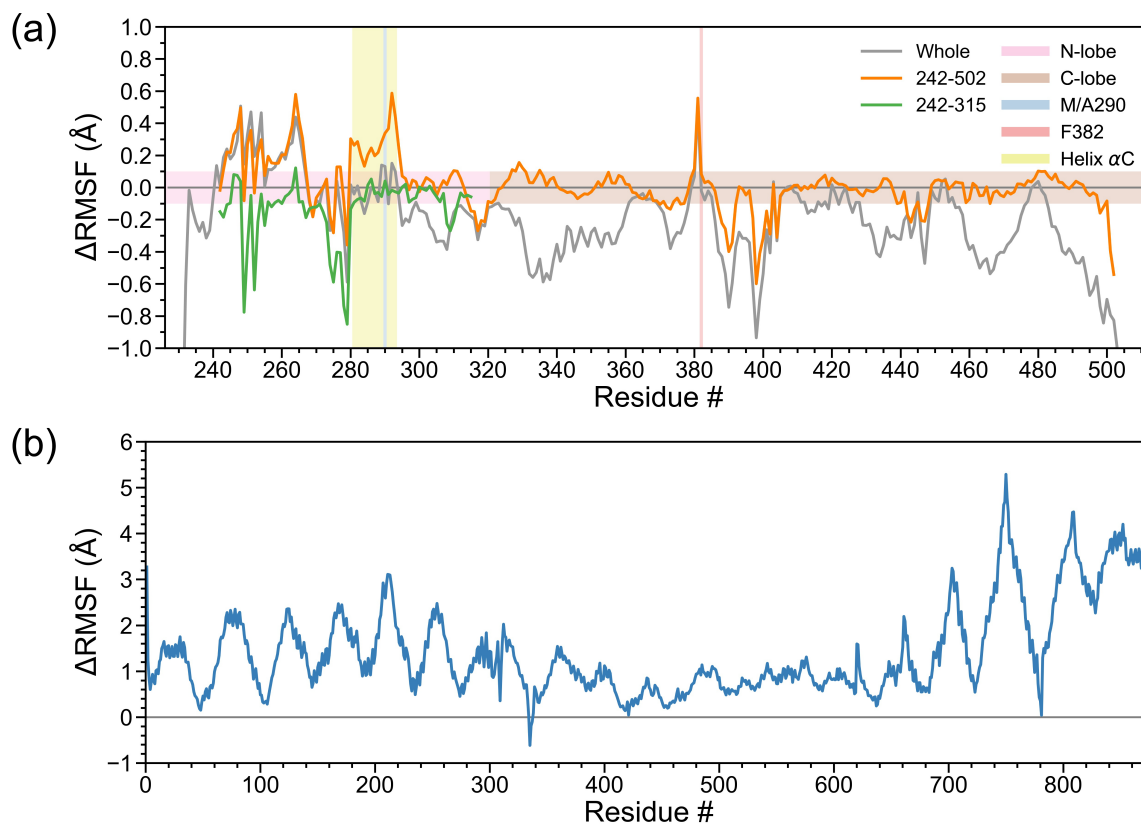


Figure B.19: Residue-level results from running root mean square fluctuation (RMSF) analysis on Abl1 and IBB-bound Importin- β . In these analyses, we compare the molecular dynamics (MD) trajectories of the alternative state for each protein to the corresponding trajectories of **(a)** Met290Val Abl1 and **(b)** unbound Importin- β , respectively. We analyze Abl1 based on different fragments the protein. Specifically, we analyze the whole protein structure, residues 242-502, and residues 242-315. Note that, in the context of Importin- β , the superhelix includes the entire structure and so we do include any additional sub-fragment analyses. The y -axis on denotes the absolute difference (or Δ -change) in RMSF between wild-type and mutants classes.

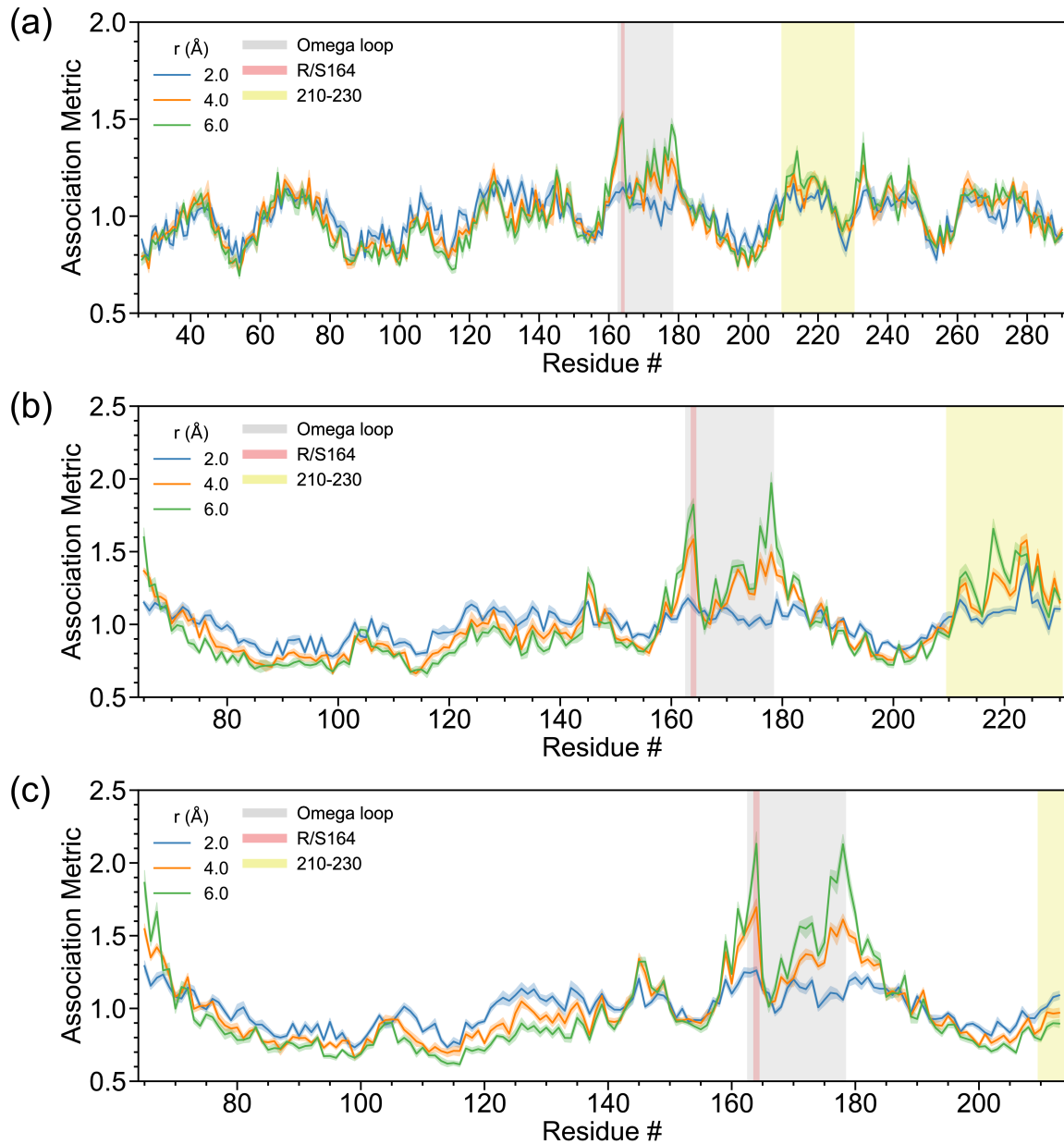


Figure B.20: Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cutoffs r values used to construct the simplicial complexes for TEM β -lactamase. Recall that we use the atomic positions for each protein to create mesh representations of their 3D structures (see Fig. 4.1 in Chapter 4). First, we draw an edge between any two atoms if their Euclidean distance smaller than some value r , namely $\text{dist}[(x_1, y_1, z_1), (x_2, y_2, z_2)] < r$. Next, we fill in all the triangles (or faces) formed by these connected edges. We treat the resulting triangulated mesh as an simplicial complex with which we can perform topological data analysis. Here, we consider the construction of mesh representations for each protein while setting $r = \{2.0, 4.0, 6.0\}$ ångströms (Å). Other SINATRA parameters were fixed: $c = 20$ cones, $d = 8$ directions per cone, $\theta = 0.80$ cap radius used to generate directions in a cone, and $l = 120$ sublevel sets per filtration. In each plot, we show the association metrics (and their corresponding standard errors) computed for each residue while analyzing (a) the whole protein, (b) fragment 65-230, and (c) fragment 65-213. Note that an overlap in signal shows the robustness of SINATRA Pro to this input parameter value.

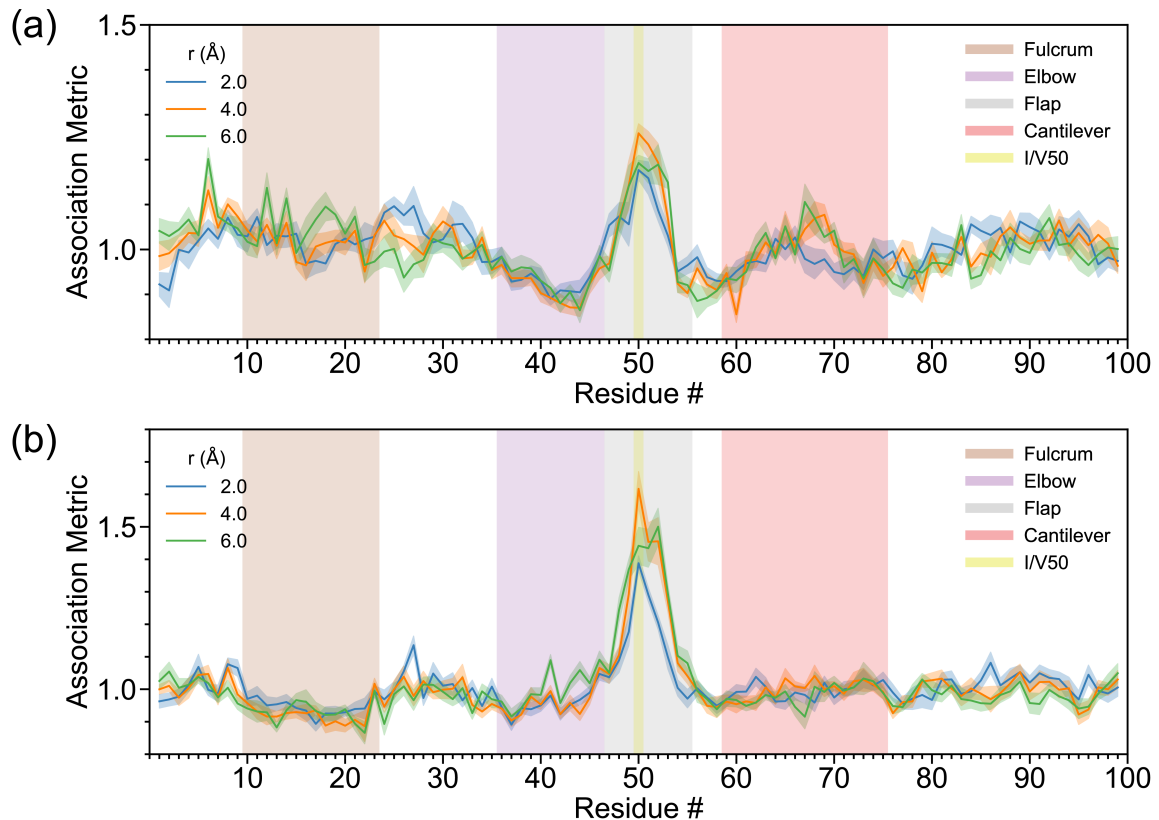


Figure B.21: Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cutoffs r values used to construct the simplicial complexes for HIV-1 protease. Recall that we use the atomic positions for each protein to create mesh representations of their 3D structures (see Fig. 4.1 in Chapter 4). First, we draw an edge between any two atoms if their Euclidean distance smaller than some value r , namely $\text{dist}[(x_1, y_1, z_1), (x_2, y_2, z_2)] < r$. Next, we fill in all the triangles (or faces) formed by these connected edges. We treat the resulting triangulated mesh as an simplicial complex with which we can perform topological data analysis. Here, we consider the construction of mesh representations for each protein while setting $r = \{2.0, 4.0, 6.0\}$ ångströms (Å). Other SINATRA parameters were fixed: $c = 20$ cones, $d = 8$ directions per cone, $\theta = 0.80$ cap radius used to generate directions in a cone, and $l = 120$ sublevel sets per filtration. In each plot, we show the association metrics (and their corresponding standard errors) computed for each residue while analyzing (a) chain A and (b) chain B. Note that an overlap in signal shows the robustness of SINATRA Pro to this input parameter value.

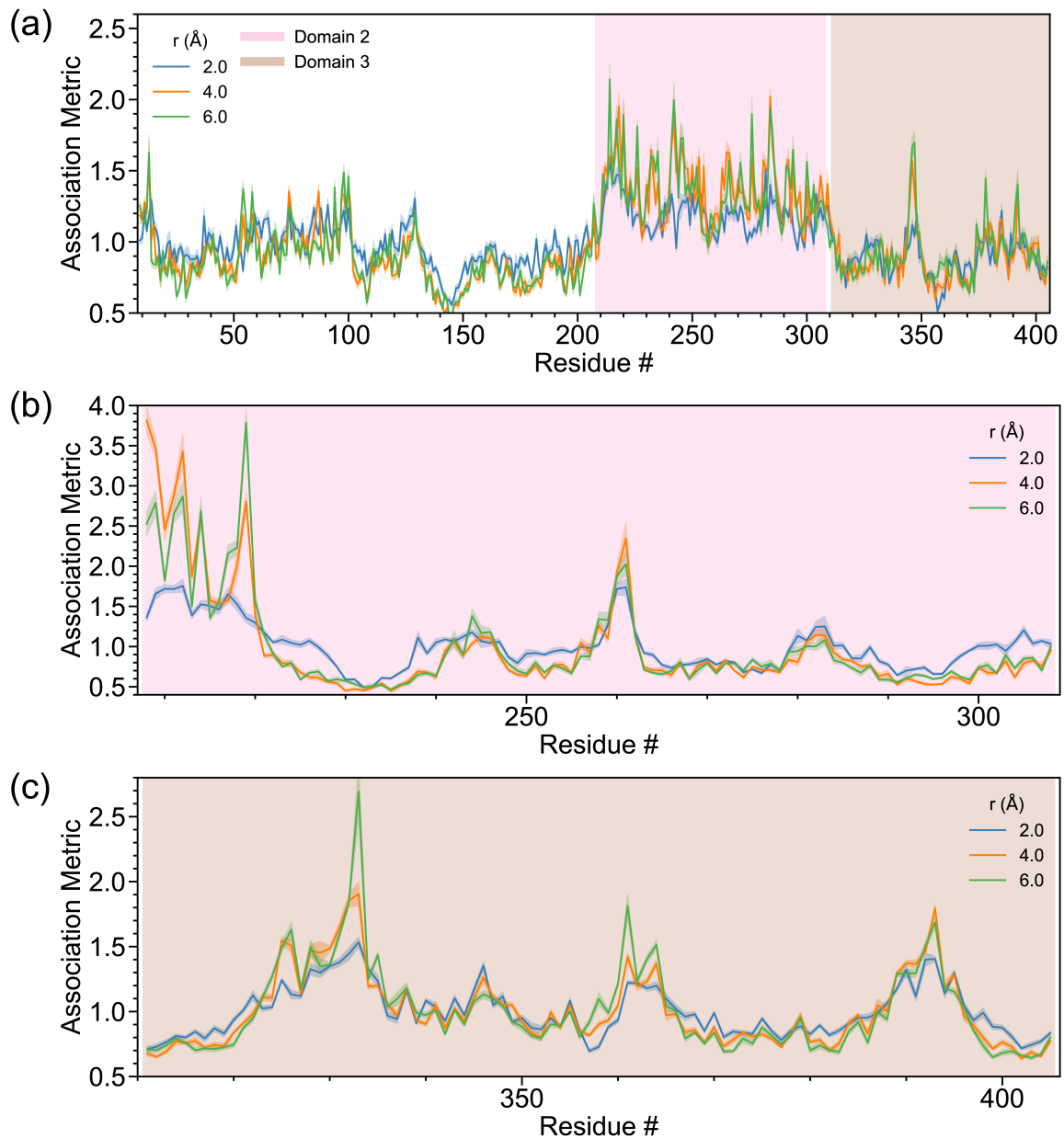


Figure B.22: Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cutoffs r values used to construct the simplicial complexes for EF-Tu. Recall that we use the atomic positions for each protein to create mesh representations of their 3D structures (see Fig. 4.1 in Chapter 4). First, we draw an edge between any two atoms if their Euclidean distance smaller than some value r , namely $\text{dist}[(x_1, y_1, z_1), (x_2, y_2, z_2)] < r$. Next, we fill in all the triangles (or faces) formed by these connected edges. We treat the resulting triangulated mesh as an simplicial complex with which we can perform topological data analysis. Here, we consider the construction of mesh representations for each protein while setting $r = \{2.0, 4.0, 6.0\}$ ångströms (Å). Other SINATRA parameters were fixed: $c = 20$ cones, $d = 8$ directions per cone, $\theta = 0.80$ cap radius used to generate directions in a cone, and $l = 120$ sublevel sets per filtration. In each plot, we show the association metrics (and their corresponding standard errors) computed for each residue while analyzing (a) whole protein, (b) fragment 208-308, and (c) fragment 311-405. Note that an overlap in signal shows the robustness of SINATRA Pro to this input parameter value.

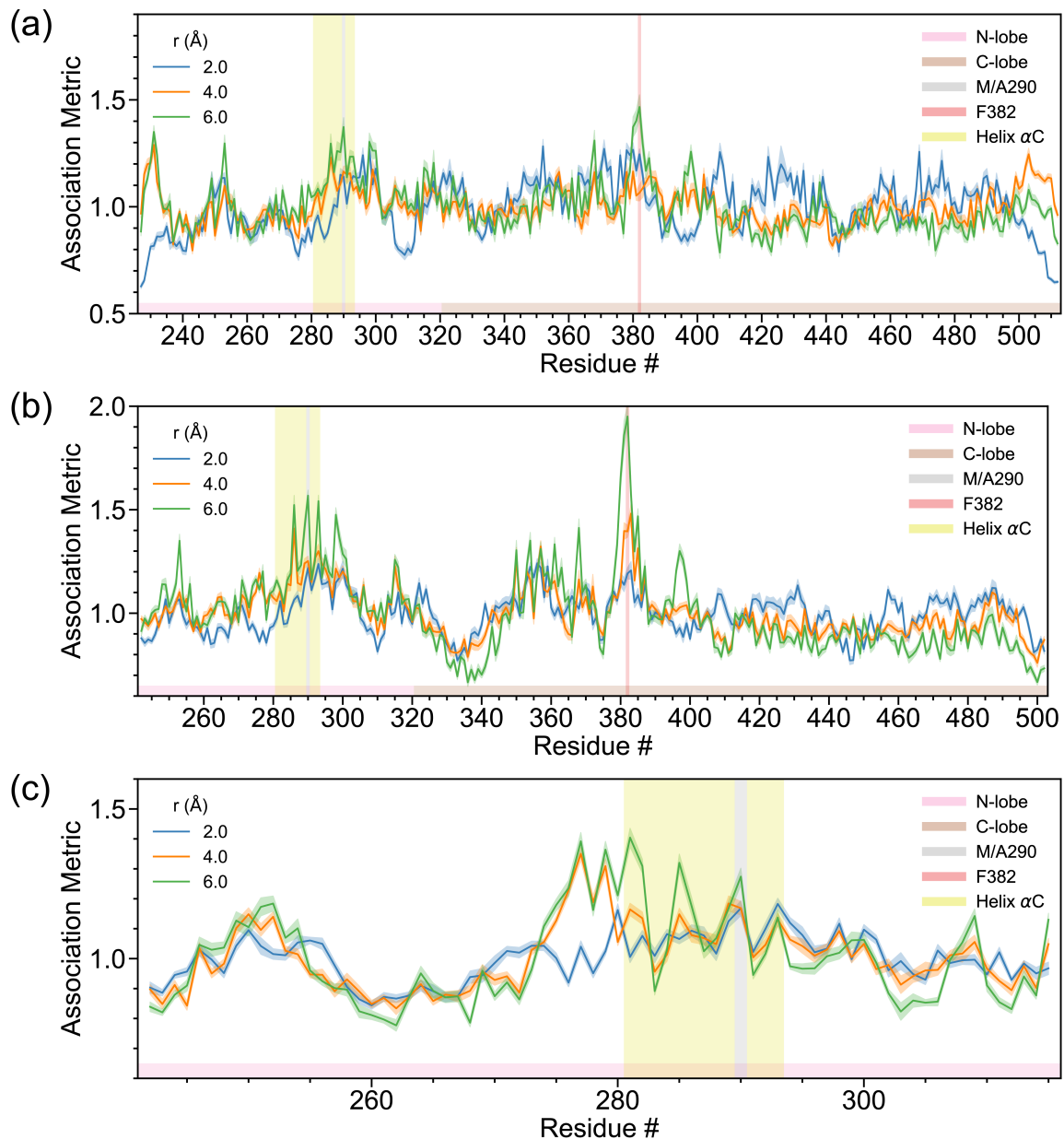


Figure B.23: Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cutoffs r values used to construct the simplicial complexes for Abl1. Recall that we use the atomic positions for each protein to create mesh representations of their 3D structures (see Fig. 4.1 in Chapter 4). First, we draw an edge between any two atoms if their Euclidean distance smaller than some value r , namely $\text{dist}[(x_1, y_1, z_1), (x_2, y_2, z_2)] < r$. Next, we fill in all the triangles (or faces) formed by these connected edges. We treat the resulting triangulated mesh as an simplicial complex with which we can perform topological data analysis. Here, we consider the construction of mesh representations for each protein while setting $r = \{2.0, 4.0, 6.0\}$ ångströms (Å). Other SINATRA parameters were fixed: $c = 20$ cones, $d = 8$ directions per cone, $\theta = 0.80$ cap radius used to generate directions in a cone, and $l = 120$ sublevel sets per filtration. In each plot, we show the association metrics (and their corresponding standard errors) computed for each residue while analyzing (a) the whole protein, (b) fragment 242-502, and (c) fragment 242-315. Note that an overlap in signal shows the robustness of SINATRA Pro to this input parameter value.

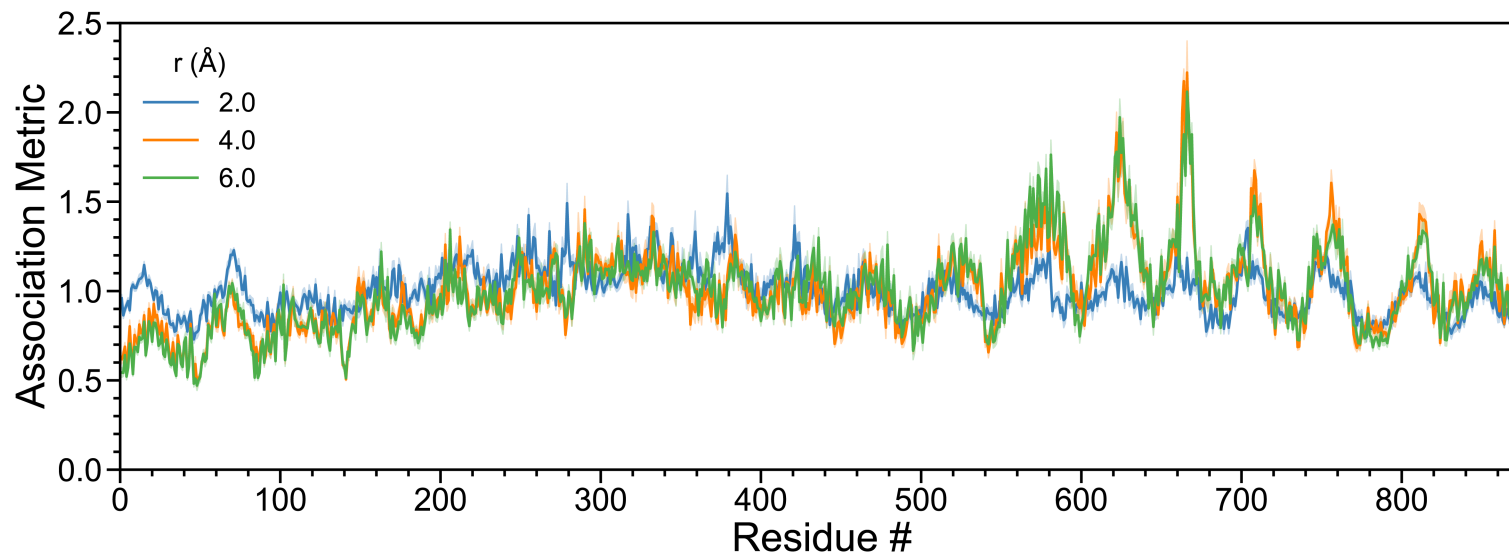


Figure B.24: Sensitivity analysis assessing the robustness of SINATRA Pro to different radius cutoffs r values used to construct the simplicial complexes for Importin- β . Recall that we use the atomic positions for each protein to create mesh representations of their 3D structures (see Fig. 4.1 in Chapter 4). First, we draw an edge between any two atoms if their Euclidean distance smaller than some value r , namely $\text{dist}[(x_1, y_1, z_1), (x_2, y_2, z_2)] < r$. Next, we fill in all the triangles (or faces) formed by these connected edges. We treat the resulting triangulated mesh as an simplicial complex with which we can perform topological data analysis. Here, we consider the construction of mesh representations for each protein while setting $r = \{2.0, 4.0, 6.0\}$ ångströms (Å). Other SINATRA parameters were fixed: $c = 20$ cones, $d = 8$ directions per cone, $\theta = 0.80$ cap radius used to generate directions in a cone, and $l = 120$ sublevel sets per filtration. Note that an overlap in signal shows the robustness of SINATRA Pro to this input parameter value.

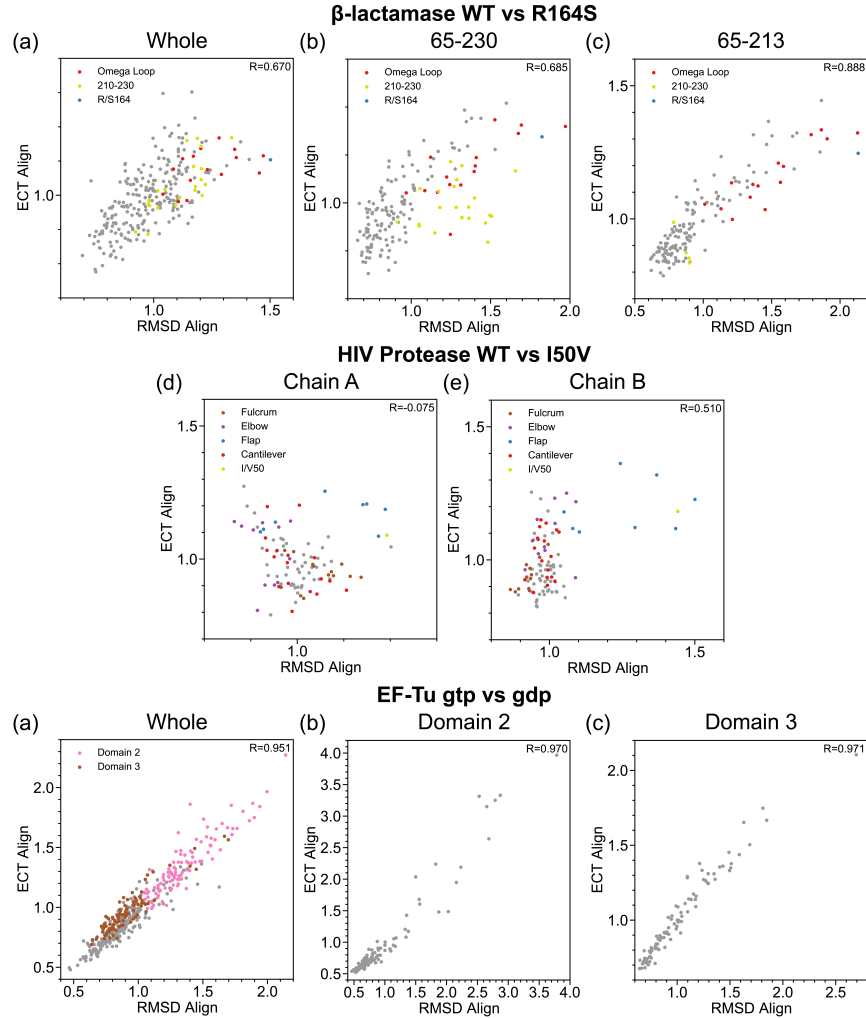


Figure B.25: Real data analysis results demonstrating the idea of running SINATRA Pro with sequence-independent (or correspondence free) structural alignment based on topological summary statistics. Here, we perform sequence-independent structural alignment where we *implicitly* normalize the 3D protein structures by rotationally aligning their topological summary statistics. To carry out this alignment procedure, we first take each pair of protein structures and superimpose the center of mass of the backbone alpha-carbons (C_α) atoms to the same origin. Next, we compute topological summary statistics over the mesh representation of each structure in $m = 500$ spherically uniformly distributed directions (see the Material and Methods in Chapter 4). We take the squared Euclidean distance between any two directions to be the cost needed to align structures via their topological summaries; and we determine the “optimal” direction alignment by finding the rotation that minimizes the cumulative cost of aligning all directional pairs between proteins. We use the random sample consensus (RANSAC) method to determine the rotational matrix that aligns the angle between any two directions to be within an error threshold of 0.9 [?]. More specifically, we require that the dot product between two directions has to be larger than 0.9 to be considered aligned in RANSAC. The figures above compare the correlation (R) between the association metrics from SINATRA Pro calculated from the structures aligned using Euler characteristic (EC) transform and structures pre-aligned using RMSD for β -lactamase: (a) the whole protein, (b) fragment 65-230, and (c) fragment 65-213; monomers of HIV-1 protease: (d) chain A and (e) chain B; and fragments of the elongation factor EF-Tu: (f) the whole protein, (g) fragment 220-310 (Domain 2), and (h) fragment 311-405 (Domain 3). Correlations near one symbolize high agreement between the two alignment schemes.

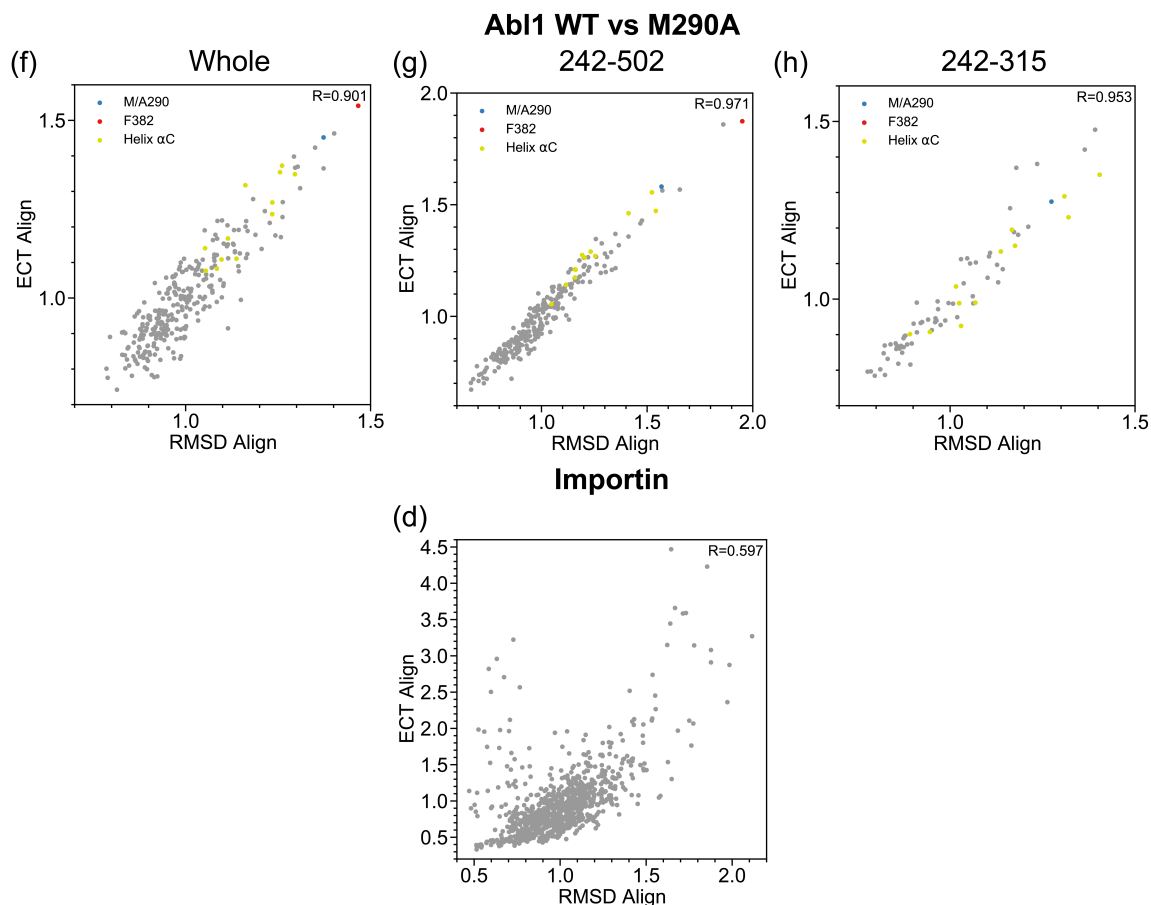


Figure B.26: Real data analysis results demonstrating the idea of running SINATRA Pro with sequence-independent (or correspondence free) structural alignment based on topological summary statistics. Here, we perform sequence-independent structural alignment where we *implicitly* normalize the 3D protein structures by rotationally aligning their topological summary statistics. To carry out this alignment procedure, we first take each pair of protein structures and superimpose the center of mass of the backbone alpha-carbons (C_α) atoms to the same origin. Next, we compute topological summary statistics over the mesh representation of each structure in $m = 500$ spherically uniformly distributed directions (see the Material and Methods in Chapter 4). We take the squared Euclidean distance between any two directions to be the cost needed to align structures via their topological summaries; and we determine the “optimal” direction alignment by finding the rotation that minimizes the cumulative cost of aligning all directional pairs between proteins. We use the random sample consensus (RANSAC) method to determine the rotational matrix that aligns the angle between any two directions to be within an error threshold of 0.9 [?]. More specifically, we require that the dot product between two directions has to be larger than 0.9 to be considered aligned in RANSAC. The figures above compare the correlation (R) between the association metrics from SINATRA Pro calculated from the structures aligned using ECT and structures pre-aligned using RMSD for fragments of Abl1 Tyrosine protein kinase: (a) the whole protein, (b) fragment 242-502, and (c) fragment 242-315; and (d) unbound Importin- β . Correlations near one symbolize high agreement between the two alignment schemes.

Total Proteins $N = 50$	Number of Cones $c = 15$			
	Directions per Cone $d = 4$		Directions per Cone $d = 8$	
	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$
(1) Read in PDB Structures	45.0 ± 0.6	45.0 ± 0.5	44.6 ± 0.0	45.0 ± 0.5
(2) Construct Meshes/Simplicial Complexes	79.0 ± 1.2	78.1 ± 0.2	78.1 ± 0.6	78.3 ± 0.4
(3) Compute Diff. Euler Characteristics	74.1 ± 0.4	74.5 ± 0.2	86.0 ± 0.7	86.6 ± 0.4
(4) Compute Atomic Variable Importance	85.0 ± 11.5	128.3 ± 3.1	126.1 ± 4.5	461.5 ± 10.8
(5) Reconstruct PDB Structures/Enrichments	22.7 ± 0.4	22.7 ± 0.1	23.0 ± 0.2	23.4 ± 0.4
Total Runtime:	305.9 ± 11.6	348.6 ± 3.1	357.8 ± 4.6	694.7 ± 0.9
Total Proteins $N = 50$	Number of Cones $c = 20$			
	Directions per Cone $d = 4$		Directions per Cone $d = 8$	
	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$
(1) Read in PDB Structures	45.1 ± 0.6	45.5 ± 0.7	45.2 ± 0.7	44.9 ± 0.5
(2) Construct Meshes/Simplicial Complexes	85.1 ± 13.1	78.7 ± 0.8	81.4 ± 5.8	77.8 ± 0.5
(3) Compute Diff. Euler Characteristics	79.4 ± 0.3	80.3 ± 1.3	93.2 ± 0.5	93.7 ± 0.4
(4) Compute Atomic Variable Importance	82.3 ± 0.3	192.3 ± 3.2	195.6 ± 3.4	991.2 ± 7.4
(5) Reconstruct PDB Structures/Enrichments	22.8 ± 0.2	23.1 ± 0.5	23.4 ± 0.4	23.3 ± 0.3
Total Runtime:	314.8 ± 13.2	420.0 ± 3.7	438.7 ± 6.8	1230.9 ± 7.5

Table B.1: Empirical runtimes for running the SINATRA algorithm as a function of its free parameters and inputs. Each entry represents the time (in seconds) it takes to run each step of the SINATRA Pro algorithm based on: (i) the total number of proteins analyzed $N = 50$, (ii) the number of cones of directions $c = \{15, 20\}$, (iii) the number of directions within each cone $d = \{4, 8\}$, and (iv) the number of sublevel sets (i.e., filtration steps) used to compute the Euler characteristic (EC) along a given direction $l = \{25, 50\}$. We simulate 10 different datasets for each combination of parameter values. Values appearing after the $\pm$ symbol are the standard deviations of these estimated times across the different runs. Each analysis was performed using simulated protein structures with ~ 2700 atoms and all runtimes were computed using a central processing unit (CPU) with 8 cores and 128 gigabytes (GB) of RAM.

Total Proteins $N = 100$	Number of Cones $c = 15$			
	Directions per Cone $d = 4$		Directions per Cone $d = 8$	
	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$
(1) Read in PDB Structures	52.2 ± 0.3	50.6 ± 0.5	50.4 ± 0.1	50.4 ± 0.1
(2) Construct Meshes/Simplicial Complexes	159.7 ± 2.1	158.7 ± 0.6	157.8 ± 1.1	157.6 ± 0.8
(3) Compute Diff. Euler Characteristics	151.8 ± 0.3	152.2 ± 0.7	178.1 ± 0.9	178.4 ± 0.6
(4) Compute Atomic Variable Importance	89.5 ± 0.7	141.2 ± 1.5	141.8 ± 4.8	489.5 ± 3.4
(5) Reconstruct PDB Structures/Enrichments	46.3 ± 0.3	46.5 ± 0.2	46.8 ± 0.1	47.2 ± 0.3
Total Runtime:	499.5 ± 2.3	549.2 ± 1.8	574.9 ± 5.0	923.1 ± 3.6
Total Proteins $N = 100$	Number of Cones $c = 20$			
	Directions per Cone $d = 4$		Directions per Cone $d = 8$	
	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$
(1) Read in PDB Structures	50.4 ± 0.1	50.3 ± 0.1	50.3 ± 0.1	50.4 ± 0.2
(2) Construct Meshes/Simplicial Complexes	158.1 ± 0.4	157.8 ± 1.1	157.9 ± 0.5	157.0 ± 0.9
(3) Compute Diff. Euler Characteristics	162.7 ± 0.6	162.5 ± 0.4	193.4 ± 0.8	193.6 ± 0.6
(4) Compute Atomic Variable Importance	98.1 ± 0.5	211.3 ± 2.1	209.4 ± 1.7	1022.8 ± 17.4
(5) Reconstruct PDB Structures/Enrichments	46.8 ± 0.2	46.7 ± 0.3	47.4 ± 0.2	47.6 ± 0.8
Total Runtime:	516.1 ± 0.9	628.6 ± 2.4	658.4 ± 1.9	1471.4 ± 17.5

Empirical runtimes for running the SINATRA algorithm as a function of its free parameters and inputs. (cont.) Each entry represents the time (in seconds) it takes to run each step of the SINATRA Pro algorithm based on: (i) the total number of proteins analyzed $N = 100$, (ii) the number of cones of directions $c = \{15, 20\}$, (iii) the number of directions within each cone $d = \{4, 8\}$, and (iv) the number of sublevel sets (i.e., filtration steps) used to compute the Euler characteristic (EC) along a given direction $l = \{25, 50\}$. We simulate 10 different datasets for each combination of parameter values. Values appearing after the $\pm$ symbol are the standard deviations of these estimated times across the different runs. Each analysis was performed using simulated protein structures with ~ 2700 atoms and all runtimes were computed using a central processing unit (CPU) with 8 cores and 128 gigabytes (GB) of RAM.

Total Proteins $N = 200$	Number of Cones $c = 15$			
	Directions per Cone $d = 4$		Directions per Cone $d = 8$	
	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$
(1) Read in PDB Structures	63.5 ± 1.3	63.3 ± 1.6	64.3 ± 2.8	63.1 ± 1.1
(2) Construct Meshes/Simplicial Complexes	305.3 ± 1.4	304.6 ± 2.2	305.4 ± 0.9	304.2 ± 1.1
(3) Compute Diff. Euler Characteristics	296.1 ± 0.8	297.1 ± 0.4	347.8 ± 1.7	348.0 ± 1.2
(4) Compute Atomic Variable Importance	111.9 ± 0.8	158.8 ± 0.7	161.8 ± 4.9	493.3 ± 1.1
(5) Reconstruct PDB Structures/Enrichments	89.1 ± 0.5	89.8 ± 1.2	90.7 ± 0.5	90.9 ± 0.3
Total Runtime:	865.9 ± 2.2	913.5 ± 3.1	970.1 ± 6.0	1299.5 ± 2.3
Total Proteins $N = 200$	Number of Cones $c = 20$			
	Directions per Cone $d = 4$		Directions per Cone $d = 8$	
	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$	Sublevel Sets $l = 25$	Sublevel Sets $l = 50$
(1) Read in PDB Structures	63.7 ± 1.2	63.5 ± 1.5	63.9 ± 1.1	64.0 ± 1.1
(2) Construct Meshes/Simplicial Complexes	310.5 ± 11.8	304.5 ± 0.6	304.8 ± 0.9	304.7 ± 1.1
(3) Compute Diff. Euler Characteristics	320.0 ± 0.5	319.6 ± 0.4	377.0 ± 1.4	377.6 ± 1.5
(4) Compute Atomic Variable Importance	119.9 ± 0.5	226.4 ± 0.5	227.5 ± 2.3	1001.7 ± 6.9
(5) Reconstruct PDB Structures/Enrichments	91.2 ± 0.7	90.6 ± 0.7	92.8 ± 1.1	92.6 ± 1.2
Total Runtime:	905.3 ± 11.9	1004.6 ± 1.9	1066.1 ± 3.3	1840.6 ± 7.3

Empirical runtimes for running the SINATRA algorithm as a function of its free parameters and inputs. (cont.) Each entry represents the time (in seconds) it takes to run each step of the SINATRA Pro algorithm based on: (i) the total number of proteins analyzed $N = 200$, (ii) the number of cones of directions $c = \{15, 20\}$, (iii) the number of directions within each cone $d = \{4, 8\}$, and (iv) the number of sublevel sets (i.e., filtration steps) used to compute the Euler characteristic (EC) along a given direction $l = \{25, 50\}$. We simulate 10 different datasets for each combination of parameter values. Values appearing after the $\pm$ symbol are the standard deviations of these estimated times across the different runs. Each analysis was performed using simulated protein structures with ~ 2700 atoms and all runtimes were computed using a central processing unit (CPU) with 8 cores and 128 gigabytes (GB) of RAM.

Bibliography

- [1] W. S. Tang, N. L. Fawzi, and J. Mittal, Refining all-atom protein force fields for polar-rich, prion-like, low-complexity intrinsically disordered proteins, *The Journal of Physical Chemistry B* **124**, 9505 (2020).
- [2] A. C. Murthy, W. S. Tang, N. Jovic, A. M. Janke, D. H. Seo, T. M. Perdikari, J. Mittal, and N. L. Fawzi, Molecular interactions contributing to FUS SYGQ LC-RGG phase separation and co-partitioning with RNA polymerase II heptads, *Nature Structural & Molecular Biology* **28**, 923 (2021).
- [3] W. S. Tang, G. M. da Silva, H. Kirveslahti, E. Skeens, B. Feng, T. Sudijono, K. K. Yang, S. Mukherjee, B. Rubenstein, and L. Crawford, A topological data analytic approach for discovering biophysical signatures in protein dynamics, *bioRxiv* (2021).
- [4] R. Van Der Lee et al., Classification of intrinsically disordered regions and proteins, *Chemical reviews* **114**, 6589 (2014).
- [5] A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner, and C. J. Brown, Intrinsic protein disorder in complete genomes, *Genome Informatics* **11**, 161 (2000).
- [6] P. Tompa, Intrinsically unstructured proteins, *Trends in biochemical sciences* **27**, 527 (2002).
- [7] S. Chavali, A. Gunnarsson, and M. M. Babu, Intrinsically disordered proteins adaptively reorganize cellular matter during stress, *Trends in biochemical sciences* **42**, 410 (2017).
- [8] P. E. Wright and H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, *Nature reviews Molecular cell biology* **16**, 18 (2015).

- [9] J. W. Hofmann, W. W. Seeley, and E. J. Huang, RNA binding proteins and the pathogenesis of frontotemporal lobar degeneration, *Annual Review of Pathology: Mechanisms of Disease* **14**, 469 (2019).
- [10] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, Germline P granules are liquid droplets that localize by controlled dissolution/condensation, *Science* **324**, 1729 (2009).
- [11] Y. Shin and C. P. Brangwynne, Liquid phase condensation in cell physiology and disease, *Science* **357**, eaaf4382 (2017).
- [12] Y. R. Li, O. D. King, J. Shorter, and A. D. Gitler, Stress granules as crucibles of ALS pathogenesis, *J Cell Biol* **201**, 361 (2013).
- [13] A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag, and J. P. Taylor, Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization, *Cell* **163**, 123 (2015).
- [14] J. Mittal, T. H. Yoo, G. Georgiou, and T. M. Truskett, Structural ensemble of an intrinsically disordered polypeptide, *The Journal of Physical Chemistry B* **117**, 118 (2013).
- [15] A. E. Conicella, G. H. Zerbe, J. Mittal, and N. L. Fawzi, ALS mutations disrupt phase separation mediated by α -helical structure in the TDP-43 low complexity C-terminal domain, *Structure* **24**, 1537 (2016).
- [16] D. T. Murray, M. Kato, Y. Lin, K. R. Thurber, I. Hung, S. L. McKnight, and R. Tycko, Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains, *Cell* **171**, 615 (2017).
- [17] M. P. Hughes, M. R. Sawaya, D. R. Boyer, L. Goldschmidt, J. A. Rodriguez, D. Cascio, L. Chong, T. Gonen, and D. S. Eisenberg, Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks, *Science* **359**, 698 (2018).
- [18] T. Chouard, Breaking the protein rules: if dogma dictates that proteins need a structure to function, then why do so many of them live in a state of disorder?, *Nature* **469**, 151 (2011).

- [19] D. H. Brookes and T. Head-Gordon, Experimental inferential structure determination of ensembles for intrinsically disordered proteins, *Journal of the American Chemical Society* **138**, 4530 (2016).
- [20] T. Löhner, A. Jussupow, and C. Camilloni, Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide, *The Journal of chemical physics* **146**, 165102 (2017).
- [21] S. Bottaro, T. Bengtson, and K. Lindorff-Larsen, Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach, in *Structural Bioinformatics*, pages 219–240, Springer, 2020.
- [22] J. Kofinger, L. S. Stelzl, K. Reuter, C. Allande, K. Reichel, and G. Hummer, Efficient ensemble refinement by reweighting, *Journal of chemical theory and computation* **15**, 3390 (2019).
- [23] G.-N. W. Gomes, M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, and C. C. Gradinaru, Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET, *Journal of the American Chemical Society* **142**, 15697 (2020).
- [24] K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, and D. E. Shaw, Structure and dynamics of an unfolded protein examined by molecular dynamics simulation, *Journal of the American Chemical Society* **134**, 3787 (2012).
- [25] A. C. Murthy, G. L. Dignon, Y. Kan, G. H. Zerbe, S. H. Parekh, J. Mittal, and N. L. Fawzi, Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain, *Nature structural & molecular biology* **26**, 637 (2019).
- [26] B. S. Schuster et al., Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior, *Proceedings of the National Academy of Sciences* **117**, 11421 (2020).
- [27] R. B. Best, N. V. Buchete, and G. Hummer, Are current molecular dynamics force fields too helical?, *Biophysical journal* **95**, L07 (2008).

- [28] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, Challenges in protein-folding simulations, *Nature physics* **6**, 751 (2010).
- [29] R. B. Best and G. Hummer, Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides, *J. Phy. Chem. B* **113**, 9004 (2009).
- [30] J. Mittal and R. B. Best, Tackling force-field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water, *Biophysical journal* **99**, L26 (2010).
- [31] R. B. Best and J. Mittal, Balance between α and β structures in ab initio protein folding, *The Journal of Physical Chemistry B* **114**, 8790 (2010).
- [32] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, How fast-folding proteins fold, *Science* **334**, 517 (2011).
- [33] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, How robust are protein folding simulations with respect to force field parameterization?, *Biophysical journal* **100**, L47 (2011).
- [34] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Improved side-chain torsion potentials for the amber ff99sb protein force field, *Proteins* **78**, 1950 (2010).
- [35] C.-Y. Zhou, F. Jiang, and Y.-D. Wu, Residue-specific force field based on protein coil library. rsff2: modification of amber ff99sb, *The journal of physical chemistry B* **119**, 1035 (2014).
- [36] P. Robustelli, S. Piana, and D. E. Shaw, Developing a molecular dynamics force field for both folded and disordered protein states, *Proceedings of the National Academy of Sciences* **115**, E4758 (2018).
- [37] S. Piana, P. Robustelli, D. Tan, S. Chen, and D. E. Shaw, Development of a force field for the simulation of single-chain proteins and protein-protein complexes, *Journal of chemical theory and computation* **16**, 2494 (2020).
- [38] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ_1 and χ_2 dihedral angles, *Journal of Chemical Theory and Computation* **8**, 3257 (2012).

- [39] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb, *Journal of chemical theory and computation* **11**, 3696 (2015).
- [40] R. B. Best, Computational and theoretical advances in studies of intrinsically disordered proteins, *Current opinion in structural biology* **42**, 147 (2017).
- [41] Z. A. Levine and J.-E. Shea, Simulations of disordered proteins and systems with conformational heterogeneity, *Current opinion in structural biology* **43**, 95 (2017).
- [42] J. Huang and A. D. MacKerell Jr, Force field development and simulations of intrinsically disordered proteins, *Current opinion in structural biology* **48**, 40 (2018).
- [43] P. Das, S. Matysiak, and J. Mittal, Looking at the disordered proteins through the computational microscope, *ACS central science* **4**, 534 (2018).
- [44] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment, *Journal of chemical theory and computation* **11**, 5513 (2015).
- [45] R. B. Best and J. Mittal, Microscopic events in β -hairpin folding from alternative unfolded ensembles, *Proceedings of the National Academy of Sciences* **108**, 11087 (2011).
- [46] J. L. Abascal and C. Vega, A general purpose model for the condensed phases of water: TIP4P/2005, *The Journal of chemical physics* **123**, 234505 (2005).
- [47] R. B. Best and J. Mittal, Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse, *The Journal of Physical Chemistry B* **114**, 14916 (2010).
- [48] P. S. Nerenberg and T. Head-Gordon, Optimizing protein- solvent force fields to reproduce intrinsic conformational preferences of model peptides, *Journal of Chemical Theory and Computation* **7**, 1220 (2011).
- [49] R. B. Best, W. Zheng, and J. Mittal, Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association, *J Chem Theory Comput* **10**, 5113 (2014).

- [50] A. E. Conicella, G. L. Dignon, G. H. Zerbe, H. B. Schmidt, M. Alexandra, Y. C. Kim, R. Rohatgi, Y. M. Ayala, J. Mittal, and N. L. Fawzi, TDP-43 α -helical structure tunes liquid-liquid phase separation and function, *Proceedings of the National Academy of Sciences* **117**, 5883 (2020).
- [51] V. H. Ryan, G. L. Dignon, G. H. Zerbe, C. V. Chabata, R. Silva, A. E. Conicella, J. Amaya, K. A. Burke, J. Mittal, and N. L. Fawzi, Mechanistic view of hnRNPA2 low-complexity domain structure, interactions, and phase separation altered by mutation and arginine methylation, *Molecular cell* **69**, 465 (2018).
- [52] G. H. Zerbe, R. B. Best, and J. Mittal, Sequence-and temperature-dependent properties of unfolded and disordered proteins from atomistic simulations, *The Journal of Physical Chemistry B* **119**, 14622 (2015).
- [53] Z. Monahan et al., Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity, *The EMBO journal* **36**, 2951 (2017).
- [54] P. Mier et al., Disentangling the complexity of low complexity proteins, *Briefings in Bioinformatics* **21**, 458 (2020).
- [55] S. Alberti, R. Halfmann, O. King, A. Kapila, and S. Lindquist, A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, *Cell* **137**, 146 (2009).
- [56] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation, *J. Chem. Theory Comput.* **4**, 435 (2008).
- [57] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, PLUMED 2: New feathers for an old bird, *Computer Physics Communications* **185**, 604 (2014).
- [58] Y. Sugita and Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, *Chemical physics letters* **314**, 141 (1999).
- [59] M. Bonomi and M. Parrinello, Enhanced sampling in the well-tempered ensemble, *Phys. Rev. Lett.* **104**, 190601 (2010).

- [60] M. Deighan, M. Bonomi, and J. Pfendner, Efficient simulation of explicitly solvated proteins in the well-tempered ensemble, *Journal of chemical theory and computation* **8**, 2189 (2012).
- [61] A. Laio and M. Parrinello, Escaping free-energy minima, *Proceedings of the National Academy of Sciences* **99**, 12562 (2002).
- [62] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, A smooth particle mesh ewald method, *J. Chem. Phys.* **103**, 8577 (1995).
- [63] Y. Shen and A. Bax, SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network, *Journal of biomolecular NMR* **48**, 13 (2010).
- [64] K. Wolfgang and S. Christian, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**, 2577 (1983).
- [65] C. Camilloni, A. De Simone, W. F. Vranken, and M. Vendruscolo, Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts, *Biochemistry* **51**, 2224 (2012).
- [66] A. M. Janke, D. H. Seo, V. Rahmanian, A. E. Conicella, K. L. Mathews, K. A. Burke, J. Mittal, and N. L. Fawzi, Lysines in the RNA Polymerase II C-terminal domain contribute to TAF15 fibril recruitment, *Biochemistry* **57**, 2549 (2018).
- [67] M. Kjaergaard, S. Brander, and F. M. Poulsen, Random coil chemical shift for intrinsically disordered proteins: effects of temperature and ph, *Journal of biomolecular NMR* **49**, 139 (2011).
- [68] K. A. Burke, A. M. Janke, C. L. Rhine, and N. L. Fawzi, Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA Polymerase II, *Mol. Cell.* **60**, 231 (2015).
- [69] R. B. Best, D. de Sancho, and J. Mittal, Residue-specific α -helix propensities from molecular simulation, *Biophys J* **102**, 1462 (2012).

- [70] C. P. Brangwynne, T. J. Mitchison, and A. A. Hyman, Active liquid-like behavior of nucleoli determines their size and shape in *Xenopus laevis* oocytes, *Proceedings of the National Academy of Sciences* **108**, 4334 (2011).
- [71] M. Feric, N. Vaidya, T. S. Harmon, D. M. Mitrea, L. Zhu, T. M. Richardson, R. W. Kriwacki, R. V. Pappu, and C. P. Brangwynne, Coexisting liquid phases underlie nucleolar subcompartments, *Cell* **165**, 1686 (2016).
- [72] A. R. Strom, A. V. Emelyanov, M. Mir, D. V. Fyodorov, X. Darzacq, and G. H. Karpen, Phase separation drives heterochromatin domain formation, *Nature* **547**, 241 (2017).
- [73] S. Chong, C. Dugast-Darzacq, Z. Liu, P. Dong, G. M. Dailey, C. Cattoglio, A. Heckert, S. Banala, L. Lavis, X. Darzacq, and R. Tjian, Imaging dynamic and selective low-complexity domain interactions that control gene transcription, *Science* **361** (2018).
- [74] A. Boija et al., Transcription factors activate genes through the phase-separation capacity of their activation domains, *Cell* **175**, 1842 (2018).
- [75] W. K. Cho, J. H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, Mediator and RNA polymerase II clusters associate in transcription-dependent condensates, *Science* **361**, 412 (2018).
- [76] B. R. Sabari et al., Coactivator condensation at super-enhancers links phase separation and gene control, *Science* **361**, eaar3958 (2018).
- [77] F. Svetoni, P. Frisone, and M. P. Paronetto, Role of FET proteins in neurodegenerative disorders, 2016.
- [78] J. I. Hoell, E. Larsson, S. Runge, J. D. Nusbaum, S. Duggimpudi, T. A. Farazi, M. Hafner, A. Borkhardt, C. Sander, and T. Tuschl, RNA targets of wild-type and mutant FET family proteins, *Nature Structural and Molecular Biology* **18**, 1428 (2011).
- [79] K. Kapeli, G. A. Pratt, A. Q. Vu, K. R. Hutt, F. J. Martinez, B. Sundararaman, R. Batra, P. Freese, N. J. Lambert, S. C. Huelga, S. J. Chun, T. Y. Liang, J. Chang, J. P. Donohue, L. Shiue, J. Zhang, H. Zhu, F. Cambi, E. Kasarskis, S. Hoon, M. Ares, C. B. Burge, J. Ravits, F. Rigo, and G. W. Yeo, Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses, *Nature Communications* **7** (2016).

- [80] F. E. Loughlin, P. J. Lukavsky, T. Kazeeva, S. Reber, E. M. Hock, M. Colombo, C. Von Schroetter, P. Pauli, A. Cléry, O. Mühlemann, M. Polymenidou, M. D. Ruepp, and F. H. Allain, The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity, *Molecular Cell* **73**, 490 (2019).
- [81] J. C. Schwartz, X. Wang, E. R. Podell, and T. R. Cech, RNA seeds higher-order assembly of FUS protein, *Cell Reports* **5**, 918 (2013).
- [82] T. A. Shelkovernikova, H. K. Robinson, J. A. Southcombe, N. Ninkina, and V. L. Buchman, Multistep process of FUS aggregation in the cell cytoplasm involves RNA-dependent and RNA-independent mechanisms, *Human molecular genetics* **23**, 5211 (2014).
- [83] M. Kato et al., Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels, *Cell* **149**, 753 (2012).
- [84] T. Murakami, S. Qamar, J. Q. Lin, G. S. K. Schierle, E. Rees, A. Miyashita, A. R. Costa, R. B. Dodd, F. T. S. Chan, C. H. Michel, D. Kronenberg-Versteeg, Y. Li, S. P. Yang, Y. Wakutani, W. Meadows, R. R. Ferry, L. Dong, G. G. Tartaglia, G. Favrin, W. L. Lin, D. W. Dickson, M. Zhen, D. Ron, G. Schmitt-Ulms, P. E. Fraser, N. A. Shneider, C. Holt, M. Vendruscolo, C. F. Kaminski, and P. St George-Hyslop, ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function, *Neuron* **88**, 678 (2015).
- [85] Z. Sun, Z. Diaz, X. Fang, M. P. Hart, A. Chesi, J. Shorter, and A. D. Gitler, Molecular determinants and genetic modifiers of aggregation and toxicity for the ALS disease protein fus/tls, *PLoS Biology* **9**, e1000614 (2011).
- [86] A. Patel et al., A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation, *Cell* **162**, 1066 (2015).
- [87] E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, and T. Mittag, Valence and patterning of aromatic residues determine the phase behavior of prion-like domains, *Science* **367**, 694 (2020).
- [88] T. Yoshizawa, R. Ali, J. Jiou, H. Y. J. Fung, K. A. Burke, S. J. Kim, Y. Lin, W. B. Peeples, D. Saltzberg, M. Soniat, J. M. Baumhardt, R. Oldenbourg, A. Sali, N. L. Fawzi, M. K.

- Rosen, and Y. M. Chook, Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites, *Cell* **173**, 693 (2018).
- [89] J. Kang, L. Lim, Y. Lu, and J. Song, A unified mechanism for LLPS of ALS/FTLD-causing FUS as well as its modulation by ATP and oligonucleic acids, *PLOS Biology* **17**, e3000327 (2019).
- [90] S. Qamar, G. Wang, S. J. Randle, F. S. Ruggeri, J. A. Varela, J. Q. Lin, E. C. Phillips, A. Miyashita, D. Williams, F. Ströhl, W. Meadows, R. Ferry, V. J. Dardov, G. G. Tartaglia, L. A. Farrer, G. S. Kaminski Schierle, C. F. Kaminski, C. E. Holt, P. E. Fraser, G. Schmitt-Ulms, D. Klenerman, T. Knowles, M. Vendruscolo, and P. St George-Hyslop, FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation- π interactions., *Cell* **173**, 720 (2018).
- [91] E. Bogaert, S. Boeynaems, M. Kato, L. Guo, T. R. Caulfield, J. Steyaert, W. Scheveneels, N. Wilmans, W. Haeck, N. Hersmus, J. Schymkowitz, F. Rousseau, J. Shorter, P. Callaerts, W. Robberecht, P. Van Damme, and L. Van Den Bosch, Molecular dissection of FUS points at synergistic effect of low-complexity domains in toxicity, *Cell Reports* **24**, 529 (2018).
- [92] M. Hofweber, S. Hutten, B. Bourgeois, E. Spreitzer, A. Niedner-Boblenz, M. Schifferer, M. D. Ruepp, M. Simons, D. Niessing, T. Madl, and D. Dormann, Phase separation of FUS is suppressed by its nuclear import receptor and arginine methylation, *Cell* **173**, 706 (2018).
- [93] J. Wang et al., A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins, *Cell* **174**, 688 (2018).
- [94] J. P. Brady, P. J. Farber, A. Sekhar, Y.-H. Lin, R. Huang, A. Bah, T. J. Nott, H. S. Chan, A. J. Baldwin, J. D. Forman-Kay, and L. E. Kay, Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation., *Proceedings of the National Academy of Sciences of the United States of America* **114**, E8194 (2017).
- [95] A. Bremer, M. Farag, W. M. Borchers, I. Peran, E. W. Martin, R. V. Pappu, and T. Mittag, Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains, *bioRxiv* , 2021.01.01.425046 (2021).

- [96] J. C. Schwartz, C. C. Ebmeier, E. R. Podell, J. Heimiller, D. J. Taatjes, and T. R. Cech, FUS binds the CTD of RNA Polymerase II and regulates its phosphorylation at Ser2, *Genes and Development* **26**, 2690 (2012).
- [97] I. Kwon, M. Kato, S. Xiang, L. Wu, P. Theodoropoulos, H. Mirzaei, T. Han, S. Xie, J. L. Corden, and S. L. McKnight, Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains, *Cell* **155**, 1049 (2013).
- [98] H. Zinszner, R. Albalat, and D. Ron, A novel effector domain from the RNA-binding protein TLS or EWS is required for oncogenic transformation by CHOP, *Genes and Development* **8**, 2513 (1994).
- [99] H. Ichikawa, K. Shimizu, Y. Hayashi, and M. Ohki, An RNA-binding protein gene, TLS/FUS, is fused to ERG in human myeloid leukemia with t(16;21) chromosomal translocation, *Cancer research* **54**, 2865 (1994).
- [100] M.-T. Wei, Y.-C. Chang, S. F. Shimobayashi, Y. Shin, and C. P. Brangwynne, Nucleated transcriptional condensates amplify gene expression, *bioRxiv* , 737387 (2019).
- [101] F. Lu, B. Portz, and D. S. Gilmour, The C-terminal domain of RNA Polymerase II is a multivalent targeting sequence that supports drosophila development with only consensus heptads, *Molecular Cell* **73**, 1232 (2019).
- [102] M. Boehning, C. Dugast-Darzacq, M. Rankovic, A. S. Hansen, T. Yu, H. Marie-Nelly, D. T. McSwiggen, G. Kokic, G. M. Dailey, P. Cramer, X. Darzacq, and M. Zweckstetter, RNA polymerase II clustering through carboxy-terminal domain phase separation, *Nature Structural and Molecular Biology* **25**, 833 (2018).
- [103] Y. Luo and B. Roux, Simulation of osmotic pressure in concentrated aqueous salt solutions, *Journal of Physical Chemistry Letters* **1**, 183 (2010).
- [104] A. Barducci, G. Bussi, and M. Parrinello, Well-tempered metadynamics: A smoothly converging and tunable free-energy method, *Phys. Rev. Lett.* **100**, 020603 (2008).
- [105] G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, and J. Mittal, Relation between single-molecule properties and phase behavior of intrinsically disordered proteins, *Proceedings of the National Academy of Sciences* **115**, 9929 (2018).

- [106] B. R. Brooks, C. L. r. Brooks, A. D. J. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, CHARMM: the biomolecular simulation program, *Journal of computational chemistry* **30**, 1545 (2009).
- [107] P. Tiwary and M. Parrinello, A time-independent free energy estimator for metadynamics, *The Journal of Physical Chemistry B* **119**, 736 (2015).
- [108] R. M. Vernon, P. A. Chong, B. Tsang, T. H. Kim, A. Bah, P. Farber, H. Lin, and J. D. Forman-Kay, Pi-Pi contacts are an overlooked protein feature relevant to phase separation, *eLife* **7** (2018).
- [109] N. L. Fawzi, A. H. Phillips, J. Z. Ruscio, M. Doucleff, D. E. Wemmer, and T. Head-Gordon, Structure and dynamics of the $\alpha\beta_{21-30}$ peptide from the interplay of NMR experiments and molecular simulations, *Journal of the American Chemical Society* **130**, 6145 (2008).
- [110] M. Kjaergaard, A. Nørholm, R. Hendus-Altenburger, S. F. Pedersen, F. M. Poulsen, and B. B. Kragelund, Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α -helices or loss of polyproline II?, *Protein Science* **19**, 1555 (2010).
- [111] R. Wuttke, H. Hofmann, D. Nettels, M. B. Borgia, J. Mittal, R. B. Best, and B. Schuler, Temperature-dependent solvation modulates the dimensions of disordered proteins, *Proceedings of the National Academy of Sciences* **111**, 5213 (2014).
- [112] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowitz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin, Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles, *Molecular Cell* **57**, 936 (2015).
- [113] T. H. Kim, B. Tsang, R. M. Vernon, N. Sonenberg, L. E. Kay, and J. D. Forman-Kay, Phospho-dependent phase separation of FMRP and CAPRIN1 recapitulates regulation of translation and deadenylation., *Science* **365**, 825 (2019).

- [114] L. E. Wong, T. H. Kim, D. R. Muhandiram, J. D. Forman-Kay, and L. E. Kay, NMR Experiments for Studies of Dilute and Condensed Protein Phases: Application to the Phase-Separating Protein CAPRIN1, *Journal of the American Chemical Society* **142**, 2471 (2020).
- [115] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, Sequence determinants of protein phase behavior from a coarse-grained model, *PLoS Computational Biology* **14**, e1005941 (2018).
- [116] S. Das, Y. H. Lin, R. M. Vernon, J. D. Forman-Kay, and H. S. Chan, Comparative roles of charge, π , and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins, *Proceedings of the National Academy of Sciences of the United States of America* **117**, 28795 (2020).
- [117] M. Paloni, R. Bailly, L. Ciandrini, and A. Barducci, Unraveling molecular interactions in liquid-liquid phase separation of disordered proteins by atomistic simulations, *Journal of Physical Chemistry B* **124**, 9009 (2020).
- [118] A. Vitalis and R. V. Pappu, ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions, *Journal of Computational Chemistry* **30**, 673 (2009).
- [119] G. H. Zerze, W. Zheng, R. B. Best, and J. Mittal, Evolution of all-atom protein force fields to improve local and global properties, *The journal of physical chemistry letters* (2019).
- [120] J.-E. Shea, R. B. Best, and J. Mittal, Physics-based computational and theoretical approaches to intrinsically disordered proteins., *Current Opinion in Structural Biology* **67** (2021).
- [121] A. Turupcu, J. Tirado-Rives, and W. L. Jorgensen, Explicit representation of cation- π interactions in force fields with $1/r^4$ nonbonded terms, *Journal of Chemical Theory and Computation* **16**, 7184 (2020).
- [122] T. Ukmar-Godec, S. Hutten, M. P. Grieshop, N. Rezaei-Ghaleh, M.-S. Cima-Omori, J. Biernat, E. Mandelkow, J. Söding, D. Dormann, and M. Zweckstetter, Lysine/RNA-interactions drive and regulate biomolecular condensation, *Nature Communications* **10**, 2909 (2019).

- [123] S. Ambadipudi, J. Biernat, D. Riedel, E. Mandelkow, and M. Zweckstetter, Liquid-liquid phase separation of the microtubule-binding repeats of the Alzheimer-related protein Tau, *Nature Communications* **8**, 275 (2017).
- [124] S. Ambadipudi, J. G. Reddy, J. Biernat, E. Mandelkow, and M. Zweckstetter, Residue-specific identification of liquid phase separation hot spots of the Alzheimer's disease-related protein Tau, *Chemical Science* **10**, 6503 (2019).
- [125] Y. E. Guo, J. C. Manteiga, J. E. Henninger, B. R. Sabari, A. Dall'Agnese, N. M. Hannett, J. H. Spille, L. K. Afeyan, A. V. Zamudio, K. Shrinivas, B. J. Abraham, A. Boija, T. M. Decker, J. K. Rimel, C. B. Fant, T. I. Lee, I. I. Cisse, P. A. Sharp, D. J. Taatjes, and R. A. Young, Pol II phosphorylation regulates a switch between transcriptional and splicing condensates, 2019.
- [126] H. Lu, D. Yu, A. S. Hansen, S. Ganguly, R. Liu, A. Heckert, X. Darzacq, and Q. Zhou, Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II, *Nature* **558**, 318 (2018).
- [127] C. A. Orengo, A. E. Todd, and J. M. Thornton, From protein structure to function, *Current Opinion in Structural Biology* **9**, 374 (1999).
- [128] Y. Shan, M. A. Seeliger, M. P. Eastwood, F. Frank, H. Xu, M. Ø. Jensen, R. O. Dror, J. Kuriyan, and D. E. Shaw, A conserved protonation-dependent switch controls drug binding in the Abl kinase, *Proceedings of the National Academy of Sciences* **106**, 139 (2009).
- [129] C. Wilson et al., Using ancient protein kinases to unravel a modern cancer drug's mechanism, *Science* **347**, 882 (2015).
- [130] S. A. Hollingsworth and R. O. Dror, Molecular dynamics simulation for all, *Neuron* **99**, 1129 (2018).
- [131] A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. W. Siderius, and D. M. Zuckerman, Best practices for quantification of uncertainty and sampling quality in molecular simulations [Article v1.0], *Living journal of computational molecular science* **1** (2018).

- [132] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, MDAnalysis: A toolkit for the analysis of molecular dynamics simulations, *Journal of Computational Chemistry* **32**, 2319 (2011).
- [133] B. J. Grant, L. Skjaerven, and X.-Q. Yao, The Bio3D packages for structural bioinformatics, *Protein Science: A Publication of the Protein Society* **30**, 20 (2021).
- [134] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, Principal component analysis for protein folding dynamics, *Journal of molecular biology* **385**, 312 (2009).
- [135] F. Sittel, A. Jain, and G. Stock, Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates, *The Journal of Chemical Physics* **141**, 07B605_1 (2014).
- [136] C. Pál, B. Papp, and M. J. Lercher, An integrated view of protein evolution, *Nature reviews genetics* **7**, 337 (2006).
- [137] R. M. Rustamov, M. Ovsjanikov, O. Azencot, M. Ben-Chen, F. Chazal, and L. Guibas, Map-based exploration of intrinsic shape differences and variability, *ACM Trans Graph* **32**, 1 (2013).
- [138] R. Huang, P. Achlioptas, L. Guibas, and M. Ovsjanikov, Limit shapes—a tool for understanding shape differences and variability in 3D model collections, *Comput Graph Forum* **38**, 187 (2019).
- [139] B. Wang, T. Sudijono, H. Kirveslahti, T. Gao, D. M. Boyer, S. Mukherjee, and L. Crawford, A statistical pipeline for identifying physical features that differentiate classes of 3D shapes, *The Annals of Applied Statistics* **15**, 638 (2021).
- [140] K. Turner, S. Mukherjee, and D. M. Boyer, Persistent homology transform for modeling shapes and surfaces, *Inf Inference* **3**, 310 (2014).
- [141] J. Curry, S. Mukherjee, and K. Turner, How many directions determine a shape and other sufficiency results for two topological transforms, *arXiv* , 1805.09782 (2019).
- [142] R. Ghrist, R. Levanger, and H. Mai, Persistent homology and Euler integral transforms, *J Appl and Comput Topology* **2**, 55 (2018).

- [143] L. Crawford, A. Monod, A. Chen, S. Mukherjee, and A. Rabadán, Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis, *J Am Stat Assoc* **115**, 1139 (2020).
- [144] Q. Jiang, S. Kurtek, and T. Needham, The weighted Euler curve transform for shape and image analysis, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.
- [145] C. Moon, Q. Li, and G. Xiao, Predicting survival outcomes using topological features of tumor pathology images, *arXiv* , 2012.12102 (2020).
- [146] B. T. Fasy, S. Micka, D. L. Millman, A. Schenfisch, and L. Williams, Challenges in reconstructing shapes from Euler characteristic curves, *arXiv* , 1811.11337 (2018).
- [147] S. Oudot and E. Solomon, Inverse problems in topological persistence, in *Topological Data Analysis*, pages 405–433, Springer, 2020.
- [148] R. M. Neal, Monte Carlo implementation of Gaussian process models for Bayesian regression and Monte Carlo implementation of Gaussian process models for Bayesian regression and classification, Technical Report 9702, Dept. of Statistics, University of Toronto, 1997.
- [149] R. M. Neal, Regression and classification using Gaussian process priors, *Bayesian Anal* **6**, 475 (1998).
- [150] C. K. I. Williams and D. Barber, Bayesian classification with Gaussian processes, *IEEE Trans Pattern Anal Mach Intell* **20**, 1342 (1998).
- [151] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, MIT Press, Cambridge, MA, 2006.
- [152] H. Nickisch and C. E. Rasmussen, Approximations for binary Gaussian process classification, *J Mach Learn Res* **9**, 2035 (2008).
- [153] B. Schölkopf, R. Herbrich, and A. J. Smola, A generalized representer theorem, in *Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pages 416–426, London, UK, UK, 2001, Springer-Verlag.

- [154] N. S. Pillai, Q. Wu, F. Liang, S. Mukherjee, and R. L. Wolpert, Characterizing the function space for Bayesian kernel models., *Journal of Machine Learning Research* **8** (2007).
- [155] Z. Zhang, G. Dai, and M. I. Jordan, Bayesian generalized kernel mixed models, *The Journal of Machine Learning Research* **12**, 111 (2011).
- [156] Y. Jiang and J. C. Reif, Modeling epistasis in genomic selection, *Genetics* **201**, 759 (2015).
- [157] L. Crawford, K. C. Wood, X. Zhou, and S. Mukherjee, Bayesian approximate kernel regression with variable selection, *J Am Stat Assoc* **113**, 1710 (2018).
- [158] L. Crawford, S. R. Flaxman, D. E. Runcie, and M. West, Variable prioritization in nonlinear black box methods: A genetic association case study, *The annals of applied statistics* **13**, 958 (2019).
- [159] A. Chaudhuri, D. Kakde, C. Sadek, L. Gonzalez, and S. Kong, The mean and median criteria for kernel bandwidth selection for support vector data description, page 842 (2017).
- [160] I. Murray, R. Prescott Adams, and D. J. MacKay, Elliptical slice sampling, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* , 541 (2010).
- [161] K. R. Singleton, L. Crawford, E. Tsui, H. E. Manchester, O. Maertens, X. Liu, M. V. Liberti, A. N. Magpusao, E. M. Stein, and J. P. Tingley, Melanoma therapeutic strategies that select against resistance by exploiting MYC-driven evolutionary convergence, *Cell Rep* **21**, 2796 (2017).
- [162] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825 (2011).
- [163] B. Xu, N. Wang, T. Chen, and M. Li, Empirical evaluation of rectified activations in convolutional network, *arXiv*, 2015.
- [164] K. Simonyan, A. Vedaldi, and A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint arXiv:1312.6034* (2013).
- [165] *SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, ACM, 2006.

- [166] E. Harder et al., OPLS3: a force field providing broad coverage of drug-like small molecules and proteins, *Journal of chemical theory and computation* **12**, 281 (2016).
- [167] P. Mark and L. Nilsson, Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K, *The Journal of Physical Chemistry A* **105**, 9954 (2001).
- [168] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, GROMACS: Fast, flexible, and free, *Journal of Computational Chemistry* **26**, 1701 (2005).
- [169] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, LINCS: A linear constraint solver for molecular simulations, *Journal of Computational Chemistry* **18**, 1463 (1997).
- [170] W. Humphrey, A. Dalke, and K. Schulten, VMD – Visual Molecular Dynamics, *Journal of Molecular Graphics* **14**, 33 (1996).
- [171] D. L. Theobald, Rapid calculation of RMSDs using a quaternion-based characteristic polynomial, *Acta Crystallographica Section A* **61**, 478 (2005).
- [172] P. Liu, D. K. Agrafiotis, and D. L. Theobald, Fast determination of the optimal rotational matrix for macromolecular superpositions, *Journal of Computational Chemistry* **31**, 1561 (2010), PMC2958452[pmcid].
- [173] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domański, D. L. Dotson, S. Buchoux, I. M. Kenney, and O. Beckstein, MDAnalysis: A python package for the rapid analysis of molecular dynamics simulations, in *Proceedings of the 15th Python in Science Conference*, pages 98–105, 2016.
- [174] M. A. Fischler and R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM* **24**, 381 (1981).
- [175] W. W. Hager, Updating the inverse of a matrix, *SIAM Review* **31**, 221 (1989).
- [176] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE transactions on information theory* **13**, 21 (1967).
- [177] T. Sellke, M. J. Bayarri, and J. O. Berger, Calibration of p-values for testing precise null hypotheses, *Am Stat* **55**, 62 (2001).

- [178] G. Gopalan and L. Bornn, FastGP: An R package for Gaussian processes, arXiv , 1507.06055 (2015).
- [179] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, UCSF chimera—a visualization system for exploratory research and analysis., *J Comput Chem* **25**, 1605 (2004).
- [180] V. Stojanoski, D.-C. Chow, L. Hu, B. Sankaran, H. F. Gilbert, B. V. Prasad, and T. Palzkill, A triple mutant in the Ω -loop of TEM-1 β -lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis, *Journal of Biological Chemistry* **290**, 10382 (2015).
- [181] A. Egorov, M. Rubtsova, V. Grigorenko, I. Uporov, and A. Veselovsky, The role of the Ω -loop in regulation of the catalytic activity of TEM-type β -lactamases, *Biomolecules* **9**, 854 (2019).
- [182] J. R. Knox, Extended-spectrum and inhibitor-resistant TEM-type beta-lactamases: mutations, specificity, and three-dimensional structure, *Antimicrobial Agents and Chemotherapy* **39**, 2593 (1995).
- [183] M. Gniadkowski, Evolution of extended-spectrum β -lactamases by mutation, *Clinical microbiology and infection* **14**, 11 (2008).
- [184] A. Brik and C.-H. Wong, HIV-1 protease: mechanism and drug discovery, *Organic & biomolecular chemistry* **1**, 5 (2003).
- [185] V. Karacostas, E. J. Wolffe, K. Nagashima, M. A. Gonda, and B. Moss, Overexpression of the HIV-1 gag-pol polyprotein results in intracellular activation of HIV-1 protease and inhibition of assembly and budding of virus-like particles, *Virology* **193**, 661 (1993).
- [186] Z. Lv, Y. Chu, and Y. Wang, HIV protease inhibitors: a review of molecular selectivity and toxicity, *HIV/AIDS (Auckland, N.Z.)* **7**, 95 (2015).
- [187] S.-Y. Rhee, J. Taylor, W. J. Fessel, D. Kaufman, W. Towner, P. Troia, P. Ruane, J. Hellinger, V. Shirvani, A. Zolopa, and R. W. Shafer, HIV-1 protease mutations and protease inhibitor cross-resistance, *Antimicrobial Agents and Chemotherapy* **54**, 4253 (2010).

- [188] O. Sheik Amamuddy, N. T. Bishop, and Ö. Tastan Bishop, Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics, *Scientific Reports* **8**, 17938 (2018), Number: 1 Publisher: Nature Publishing Group.
- [189] L. L. Palese, Conformations of the HIV-1 protease: A crystal structure data set analysis, *Biochimica Et Biophysica Acta. Proteins and Proteomics* **1865**, 1416 (2017).
- [190] V. Hornak, A. Okur, R. C. Rizzo, and C. Simmerling, HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations, *Proceedings of the National Academy of Sciences of the United States of America* **103**, 915 (2006).
- [191] F. Liu, A. Y. Kovalevsky, Y. Tie, A. K. Ghosh, R. W. Harrison, and I. T. Weber, Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir, *Journal of Molecular Biology* **381**, 102 (2008).
- [192] J. C. Adkins and D. Faulds, Amprenavir, *Drugs* **55**, 837 (1998).
- [193] K. L. Harvey, V. M. Jarocki, I. G. Charles, and S. P. Djordjevic, The diverse functional roles of elongation factor Tu (EF-Tu) in microbial pathogenesis, *Frontiers in microbiology* , 2351 (2019).
- [194] M. Warias, H. Grubmüller, and L. V. Bock, tRNA dissociation from EF-Tu after GTP hydrolysis: primary steps and antibiotic inhibition, *Biophysical journal* **118**, 151 (2020).
- [195] T. M. Schmeing, R. M. Voorhees, A. C. Kelley, Y.-G. Gao, F. V. Murphy IV, J. R. Weir, and V. Ramakrishnan, The crystal structure of the ribosome bound to EF-Tu and aminoacyl-tRNA, *Science* **326**, 688 (2009).
- [196] H. Li, X.-Q. Yao, and B. J. Grant, Comparative structural dynamic analysis of GTPases, *PLOS Computational Biology* **14**, e1006364 (2018), Publisher: Public Library of Science.
- [197] S. R. Hubbard and J. H. Till, Protein tyrosine kinase structure and function, *Annual Review of Biochemistry* **69**, 373 (2000).
- [198] E. K. Greuber, P. Smith-Pearson, J. Wang, and A. M. Pendergast, Role of ABL family kinases in cancer: from leukaemia to solid tumours, *Nature Reviews Cancer* **13**, 559 (2013).

- [199] E. P. Reddy and A. K. Aggarwal, The ins and outs of BCR-ABL inhibition, *Genes & cancer* **3**, 447 (2012).
- [200] T. Sacha, Imatinib in chronic myeloid leukemia: an overview, *Mediterranean journal of hematology and infectious diseases* **6** (2014).
- [201] D. G. Aguilera and A. M. Tsimberidou, Dasatinib in chronic myeloid leukemia: a review, *Therapeutics and Clinical Risk Management* **5**, 281 (2009).
- [202] T. Xie, T. Saleh, P. Rossi, and C. G. Kalodimos, Conformational states dynamically populated by a kinase determine its function, *Science* **370**, eabc2754 (2020).
- [203] A. Harel and D. J. Forbes, Importin beta: conducting a much larger cellular symphony, *Molecular cell* **16**, 319 (2004).
- [204] U. Zachariae and H. Grubmüller, Importin- β : structural and dynamic determinants of a molecular spring, *Structure* **16**, 906 (2008).
- [205] G. Cingolani, C. Petosa, K. Weis, and C. W. Müller, Structure of importin- β bound to the IBB domain of importin- α , *Nature* **399**, 221 (1999), Number: 6733 Publisher: Nature Publishing Group.
- [206] K. Halder, N. Dölker, Q. Van, I. Gregor, A. Dickmanns, I. Baade, R. H. Kehlenbach, R. Ficner, J. Enderlein, H. Grubmüller, and H. Neumann, MD simulations and FRET reveal an environment-sensitive conformational plasticity of importin- β , *Biophysical Journal* **109**, 277 (2015).
- [207] F. Langenfeld et al., SHREC 2020: Multi-domain protein shape retrieval challenge, *Computers & Graphics* **91**, 189 (2020).
- [208] M. Machat, F. Langenfeld, D. Craciun, L. Sirugue, T. Labib, N. Lagarde, M. Maria, and M. Montes, Comparative evaluation of shape retrieval methods on macromolecular surfaces: an application of computer vision methods in structural bioinformatics, *Bioinformatics* (2021), bt511.
- [209] L. N. Kinch, J. Pei, A. Kryshchuk, R. D. Schaeffer, and N. V. Grishin, Topology evaluation of models for difficult targets in the 14th round of the critical assessment of

- protein structure prediction (CASP14), *Proteins: Structure, Function, and Bioinformatics* **89**, 1673 (2021).
- [210] Q. Jiang, S. Kurtsek, and T. Needham, The weighted Euler curve transform for shape and image analysis, *CoRR* **abs/2004.11128** (2020).
- [211] E. Somasundaram, A. Litzler, R. Wadhwa, S. Owen, and J. Scott, Persistent homology of tumor CT scans is associated with survival in lung cancer, *Medical Physics* **48**, 7043 (2021).
- [212] O. Vipond, J. A. Bull, P. S. Macklin, U. Tillmann, C. W. Pugh, H. M. Byrne, and H. A. Harrington, Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors, *Proceedings of the National Academy of Sciences* **118** (2021).
- [213] G. Wallin, S. C. L. Kamerlin, and J. Åqvist, Energetics of activation of GTP hydrolysis on the ribosome, *Nature Communications* **4**, 1733 (2013).
- [214] D. Mondal and A. Warshel, EF-Tu and EF-G are activated by allosteric effects, *Proceedings of the National Academy of Sciences* **115**, 3386 (2018).
- [215] A. P. Kornev, N. M. Haste, S. S. Taylor, and L. F. T. Eyck, Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism, *Proc Natl Acad Sci U S A* **103**, 17783 (2006).
- [216] M. Azam, M. A. Seeliger, N. S. Gray, J. Kuriyan, and G. Q. Daley, Activation of tyrosine kinases by mutation of the gatekeeper threonine, *Nat Struct Mol Biol* **15**, 1109 (2008).
- [217] A. P. Kornev and S. S. Taylor, Defining the conserved internal architecture of a protein kinase, *Biochim Biophys Acta* **1804**, 440 (2010).
- [218] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The protein data bank, *Nucleic Acids Research* **28**, 235 (2000).
- [219] S. Belongie, Rodrigues' rotation formula, From MathWorld—A Wolfram Web Resource, created by Eric W. Weisstein.
<http://mathworld.wolfram.com/RodriguesRotationFormula.html> (1999).
- [220] H. Hasegawa and L. Holm, Advances and pitfalls of protein structural alignment, *Current opinion in structural biology* **19**, 341 (2009).

- [221] J. Ma and S. Wang, Algorithms, applications, and challenges of protein structure alignment, *Advances in Protein Chemistry and Structural Biology* **94**, 121 (2014).
- [222] J. Jumper et al., Highly accurate protein structure prediction with AlphaFold, *Nature* **596**, 583 (2021).
- [223] L. Holm and C. Sander, Protein structure comparison by alignment of distance matrices, *Journal of molecular biology* **233**, 123 (1993).
- [224] I. N. Shindyalov and P. E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path., *Protein engineering* **11**, 739 (1998).
- [225] W. R. Taylor and C. A. Orengo, Protein structure alignment, *Journal of molecular biology* **208**, 1 (1989).
- [226] J. Jung and B. Lee, Protein structure alignment using environmental profiles, *Protein Engineering* **13**, 535 (2000).
- [227] E. Krissinel and K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallographica Section D: Biological Crystallography* **60**, 2256 (2004).
- [228] Y. Ye and A. Godzik, FATCAT: a web server for flexible structure comparison and structure similarity searching, *Nucleic acids research* **32**, W582 (2004).
- [229] R. Kolodny and N. Linial, Approximate protein structural alignment in polynomial time, *Proceedings of the National Academy of Sciences* **101**, 12201 (2004).
- [230] J. Zhu and Z. Weng, FAST: a novel protein structure alignment algorithm, *PROTEINS: Structure, Function, and Bioinformatics* **58**, 618 (2005).
- [231] Y. Zhang and J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic acids research* **33**, 2302 (2005).
- [232] S. Wang and W.-M. Zheng, CLePAPS: fast pair alignment of protein structures based on conformational letters, *Journal of bioinformatics and computational biology* **6**, 347 (2008).
- [233] I. Wohlers, F. S. Domingues, and G. W. Klau, Towards optimal alignment of protein structure distance matrices, *Bioinformatics* **26**, 2273 (2010).

- [234] M. Shatsky, R. Nussinov, and H. J. Wolfson, MultiProt—a multiple protein structural alignment algorithm, in *International Workshop on Algorithms in Bioinformatics*, pages 235–250, Springer, 2002.
- [235] O. Dror, H. Benyamini, R. Nussinov, and H. Wolfson, MASS: multiple structural alignment by secondary structures, *Bioinformatics* **19**, i95 (2003).
- [236] C. Guda, S. Lu, E. D. Scheeff, P. E. Bourne, and I. N. Shindyalov, CE-MC: a multiple protein structure alignment server, *Nucleic acids research* **32**, W100 (2004).
- [237] D. Lupyan, A. Leo-Macias, and A. R. Ortiz, A new progressive-iterative algorithm for multiple structure alignment, *Bioinformatics* **21**, 3255 (2005).
- [238] Y. Ye and A. Godzik, Multiple flexible structure alignment using partial order graphs, *Bioinformatics* **21**, 2362 (2005).
- [239] A. S. Konagurthu, J. C. Whisstock, P. J. Stuckey, and A. M. Lesk, MUSTANG: a multiple structural alignment algorithm, *Proteins: Structure, Function, and Bioinformatics* **64**, 559 (2006).
- [240] M. Menke, B. Berger, and L. Cowen, Matt: local flexibility aids protein multiple structure alignment, *PLoS computational biology* **4**, e10 (2008).
- [241] W. Sheng and Z. Wei-Mou, Fast multiple alignment of protein structures using conformational letter blocks, *The Open Bioinformatics Journal* **3** (2009).
- [242] S. Wang, J. Peng, and J. Xu, Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling, *Bioinformatics* **27**, 2537 (2011).
- [243] G. Mayr, F. S. Domingues, and P. Lackner, Comparative analysis of protein structure alignments, *BMC Structural Biology* **7**, 1 (2007).
- [244] S. Wang, J. Ma, J. Peng, and J. Xu, Protein structure alignment beyond spatial proximity, *Scientific reports* **3**, 1 (2013).
- [245] M. van Kempen, S. Kim, C. Tumescheit, M. Mirdita, J. Söding, and M. Steinegger, Foldseek: fast and accurate protein structure search, *bioRxiv* (2022).

- [246] H. Cheng, B.-H. Kim, and N. V. Grishin, MALISAM: a database of structurally analogous motifs in proteins, *Nucleic acids research* **36**, D211 (2007).
- [247] H. Cheng, B.-H. Kim, and N. V. Grishin, Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets, *Journal of molecular biology* **377**, 1265 (2008).
- [248] S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein blocks, *Proceedings of the National Academy of Sciences* **89**, 10915 (1992).
- [249] S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of molecular biology* **48**, 443 (1970).
- [250] R. Guerois, J. E. Nielsen, and L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, *Journal of molecular biology* **320**, 369 (2002).
- [251] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, The FoldX web server: an online force field, *Nucleic acids research* **33**, W382 (2005).
- [252] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications, *Nature reviews Drug discovery* **3**, 935 (2004).
- [253] K. M. Ruff and R. V. Pappu, AlphaFold and implications for intrinsically disordered proteins, *Journal of Molecular Biology* **433**, 167208 (2021).
- [254] W. Shalongo, L. Dugad, and E. Stellwagen, Distribution of helicity within the model peptide acetyl(aaqaa)3amide, *Journal of the American Chemical Society* **116**, 8288 (1994).
- [255] R. Tibshirani, Regression shrinkage and selection via the lasso, *J R Stat Soc Series B Stat Methodol* **58**, 267 (1996).
- [256] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**, 55 (1970).

- [257] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J R Stat Soc Series B Stat Methodol* **67**, 301 (2005).