

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Seeing the Unseen: Incorporating Dynamics for the Understanding and Prediction of Macromolecular Properties

Permalink

<https://escholarship.org/uc/item/7kp863d4>

Author

Pecora de Barros, Emilia

Publication Date

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Seeing the Unseen:

Incorporating Dynamics for the Understanding and Prediction of Macromolecular Properties

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

in

Chemistry

by

Emília Pécora de Barros

Committee in charge:

Professor Rommie E. Amaro, Chair
Professor Susan S. Taylor, Co-Chair
Professor Elizabeth Komives
Professor J. Andrew McCammon
Professor Andrew McCulloch
Professor Francesco Paesani

2020

Copyright

Emília Pécora de Barros, 2020

All rights reserved.

The Dissertation of Emília Pécora de Barros is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2020

TABLE OF CONTENTS

Signature page	iii
Table of contents	iv
List of figures	v
List of tables	ix
Acknowledgements	x
Vita	xiii
Abstract of the dissertation.....	xiv
Chapter 1. Introduction.....	1
Chapter 2. Electrostatic Interactions as Mediators in the Allosteric Activation of Protein Kinase A RI α	38
Chapter 3. Improving the Efficiency of Ligand-Binding Protein Design with Molecular Dynamics Simulations.....	77
Chapter 4. On the Conformational Diversity within Protein Crystals.....	124
Chapter 5. Towards Mutant-Specific Therapies: Uncovering the Dynamical Landscape and Druggability of p53 DNA Binding Domain with Markov State Models	154

LIST OF FIGURES

Figure 1.1. Free energy landscape representation and associated timescales.....	4
Figure 1.2. Example of force field equation used for calculation of potential energy in MD simulations.....	7
Figure 1.3. Typical biochemical motions and associated timescales.....	9
Figure 1.4. Representation of adaptive sampling performed for MSM construction.....	10
Figure 1.5. Representation of the steps involved in MSM construction for an example model in which there is interest in characterizing the motion of a central flexible loop, highlighted.....	13
Figure 1.6. Representation of the process involved in identification of metastable states.....	16
Figure 1.7. MSM validation metrics.....	20
Figure 1.8. Diffuse scattering versus Bragg data.....	22
Figure 2.1. (a) Sequence alignment of the B/C helix of the 4 isoforms of the regulatory subunit. The positive patch in the B/C helix is colored red. (b) Representation of the regulatory subunit and B/C helix conformation at the two functional conformations of PKA.....	42
Figure 2.2. Dynamics of the wild type system. Spherical coordinate analysis of the conformational flexibility of wild type RI α	48
Figure 2.3. Histogram of CBD's centers of mass for the mutants versus wildtype	50
Figure 2.4. Free energy landscape in terms of spherical angles for the wild type and mutants.....	51
Figure 2.5. Novel conformations sampled in (a) R239A and (b) R241A simulations.....	52
Figure 2.6. B/C helical proportion analysis of wild type and mutants' systems.....	54
Figure 2.7. The network of salt bridges.....	59
Figure 2.S1. Principal components analysis of wildtype trajectory.....	63
Figure 2.S2. Mutants free energy landscape in terms of spherical angles overlaid on the wildtype sampling conformation (gray outline).....	64
Figure 2.S3. Structural assignment to the free energy landscapes of (a) R239A and (b) R241A....	65
Figure 2.S4. Per-residue analysis of helical proportion for residues located in the B/C helix.....	66

Figure 2.S5. Representation of the two residues in the B/C helix with smallest helical proportion as verified from the simulations.....	66
Figure 2.S6. Nucleotide Binding and Allosteric Activation of RI α B/C helix basic patch mutants.....	67
Figure 2.S7. Analysis of protein frustration using the protein frustratometer methodology.....	68
Figure 2.S8. R239A salt bridges' lifetime (in turquoise) for those that were altered by more than 10% with the mutation, compared to wildtype (gray).....	69
Figure 2.S9. K240A salt bridges' lifetime (in orange) for those that were altered by more than 10% with the mutation, compared to wildtype (gray).....	70
Figure 2.S10. R241A salt bridges' lifetime (in red) for those that were altered by more than 10% with the mutation, compared to wildtype (gray).....	71
Figure 2.S11. K240A salt bridges' lifetime (in blue) for those that were altered by more than 10% with the mutation, compared to wildtype (gray).....	72
Figure 2.S12. Representation of the extended electrostatic network connecting the tandem cAMP binding domains.....	73
Figure 3.1. (a) Summary of the design data set used in the simulations, constituting two protein scaffolds (β -barrels and DIG designs) designed to bind distinct ligands.....	82
Figure 3.2. Evaluation of binding determinants for DIG designs.....	90
Figure 3.3. Comparison of results from the crystal (DIG10.2 and 10.3) and modeled structure (DIG10.2a and 10.3a) simulations.....	95
Figure 3.4. β -barrel designs distribution in terms of the identified discriminative features for design screening.....	98
Figure 3.5. Unsupervised learning model for design classification.....	102
Figure 3.S1. RMSF values for the DIG designs.....	109
Figure 3.S2. POVME results for DIG designs.....	110
Figure 3.S3. Dihedral angle analysis of ligand-interacting side chains.....	111
Figure 3.S4. Protein-ligand distance for select replicates.....	111
Figure 3.S5. Convergence analysis for the DIG and β -barrel simulations.....	112
Figure 3.S6. Replicate convergence analysis for the HBI_10 and HBI_11 β -barrel simulations.....	113

Figure 3.S7. Properties of the static, Rosetta-designed protein structures used as starting conformations for the simulations.....	114
Figure 3.S8. (a) RMSF values for select non-binders and (b) tight binders of the β -barrel design dataset.....	115
Figure 3.S9. Joint unsupervised classification model for DIG and β -barrel designs.....	116
Figure 3.S10. Logistic regression feature weights from a model trained on 80% of the DIG and β -barrel design data.....	116
Figure 4.1. Representation of the reciprocal relationship between diffuse scattering and MD simulations in modeling protein motions.....	127
Figure 4.2. Representation of the staphylococcal nuclease supercell model used in this study....	128
Figure 4.3. Flexible regions MSM.....	132
Figure 4.4. Details of the different flexible regions identified in the metastable states.....	134
Figure 4.5. Chain state distributions in the supercell.....	136
Figure 4.6. Chain cross-correlations.....	138
Figure 4.7. Correlation of experimental and MSM-computed diffuse scattering.....	140
Figure 4.8. Active site MSM.....	141
Figure 4.9. Unit cell MSM.....	142
Figure 4.S1. Implied timescale plot for the flexible regions MSM.....	144
Figure 4.S2. Chapman-Komolgorov test for the flexible regions MSM.....	145
Figure 4.S3. Protein RMSF.....	146
Figure 4.S4. Correlation of flexible regions features identified by the tICA-directed procedure with tICA coordinates.....	146
Figure 4.S5. Inter-chain interactions observed for a protein chain that exhibits the extended C-terminal conformation seen in metastable state.....	147
Figure 4.S6. Validation metrics of the active site MSM.....	148
Figure 4.S7. Validation metrics of the unit cell MSM.....	149
Figure 5.1. (a) Monomeric p53 DNA-binding domain in complex with DNA (from PDB 1TSR) with important functional regions highlighted.....	162

Figure 5.2. (a) Free energy landscape of wildtype (top) and Y220C (bottom) in terms of tICA components (tICs).....	164
Figure 5.3. L1-centered MSM.....	167
Figure 5.4. Free energy landscape of wildtype and Y220C systems in terms of L6 features.....	169
Figure 5.5. L6-centered MSM.....	170
Figure 5.6. Equilibrium population and mean first passage times (MFPT) for the two major wildtype and Y220C metastable states.....	172
Figure 5.7. (a) Indication of mutant-exclusive states in the free energy landscape.....	173
Figure 5.S1. L1 model validation analysis.....	178
Figure 5.S2. L6 model validation analysis.....	179
Figure 5.S3. Alpha carbon RMSF.....	180
Figure 5.S4. tICA correlation for features incorporating functionally-important motifs in the protein (H1, H2, L2, L3, S6/7) in addition to L1 and L6.....	181
Figure 5.S5. Example of frame exhibiting most stable intra-loop hydrogen bonds, involving Ser116 in L1 and Asp228 or Thr231 in L6.....	182
Figure 5.S6. Metastable states identified via Hidden Markov models overlaid over wildtype and Y220C L6-features free energy landscape.....	183
Figure 5.S7. Representation of the Thr149-Pro222 interaction thought to stabilize the bent L6 conformation observed in the mutant-exclusive states.....	183

LIST OF TABLES

Table 2.1. Nucleotide binding and allosteric activation of RI α B/C Helix basic patch mutants and wild type.....	56
Table 2.S1. Details of the simulated systems. Arginines shown in cyan, lysines in ochre and mutated residues in orange.....	62
Table 3.1. Summary of designed protein data set ^{5,7}	84
Table 3.2. Evaluation of the supervised learning classifiers using 5-fold cross validationa.....	105
Table 3.3. Evaluation of the generality of the classifiers, with models trained exclusively on the \square -barrel designs ^a	106
Table 3.4. Evaluation of the generality of the classifiers, with models trained exclusively on the DIG designs.....	106
Table 3.S1. Evaluation of the supervised learning classifiers, using 33% or 50% of the data in the training seta.	117
Table 4.S1. Pairs selected as initial features in the tICA analysis.	150
Table 5.S1. Stepwise tICA-based selection of features for model building.	175
Table 5.S2. Pairs used for featurization of the simulations for model construction	176
Table 5.S3. Persistence of L6-S3/S4 hydrogen bonds (in % of frames in the simulation).....	177

AKNOWLEDGEMENTS

I would like to start by thanking my advisor, Rommie Amaro, who was the person responsible for showing me the wonderful world of Computational Chemistry. I was lucky enough to stumble upon it during her first-ever BioChemCoRe summer research program in the Summer of 2012, when I was doing a year of study in the United States as an undergraduate and was starting to get frustrated upon realizing that lab work wasn't for me, despite my love for Chemistry. Thank you for opening my eyes to this universe and for your constant support and belief in my abilities throughout all these years. I don't know where I would be in my scientific path now if I hadn't been lucky enough to be accepted into your program all those years ago. On the same note, a huge thank you to Márcia Fenley, who first suggested I apply to BioChemCoRe and who's been a good friend all these years.

A heartfelt thank you to the incredible mentors I've had in my scientific journey so far: first to Robert D. Malmstrom, who patiently introduced me to the concepts of Computational Chemistry as an undergraduate in BioChemCoRe all the way through my first years in graduate school, to my Master's PI, Leandro Martínez, who encouraged my scientific curiosity, to Jamie M. Schiffer, who helped me realize my independence as a researcher, and Özlem Demir, who kindly guided me through my first steps in the world of drug discovery.

Thank you to my Amaro Lab peers, who helped me with research questions and never ending troubleshooting, kept me company in lab and helped make the challenging journey through graduate school easier. Special thanks to Bryn Taylor for being a partner while we figured out Markov State Model theory, and Teri Simas and Bryan Hill, for your tireless help in sorting out office matters and computer issues.

I have been fortunate to have been able to experience computational chemistry research in a pharmaceutical industry setting during an internship at GlaxoSmithKline in Philadelphia. I would like to thank the brilliant people who shared their expertise and career advice with me during this time, particularly Ross Walker, Alan Graves, Neysa Nevins and Constantine Kreatsoulas.

Moving out of your home country to a novel place is exciting but can be terribly lonely at times. I have been lucky enough to have met incredible friends who made my years in San Diego so thoroughly enjoyable, and even more fortunate that they're too many to name here. Special thanks to Pek Jeong, Anastassia Hirlinger and Sasha Heyneman, my first close friends in San Diego who were always down for a trip, be it to explore a new restaurant or a new state, and to the wonderful friends who became my Brazilian family abroad: Leticia Nocko, Adriane Minori, Chaiane Wiggers, Danilo Gasques, Geovana Pessoa, Adrielly Razzini and Izabela Batista. You helped turn San Diego into home.

My lovely family back home was my support system all these years, cheering me on through countless Skype calls and WhatsApp messages. Thank you to my siblings, Paula Pécora de Barros and Tomás Pécora de Barros, for their good humor, advice and friendship. A heartfelt thank you to my partner of 11 years, Everton Rigotto Genari, for his unwavering support through these five years of long-distance relationship, helping me overcome the emotional challenges of this journey and exemplifying the real nature of selfless love. And my sincerest gratitude to my parents, Araújo Pécora and Celso Barros, for their boundless love and guidance all my life, for instilling in me the importance of hard work and dedication, and encouraging me to pursue my dreams. I could not have done this without you.

Chapter 2, in full, is a modified reprint of the material as it appears in “Barros, E. P., Malmstrom, R. D., Nourbakhsh, K., Del Rio, J. C., Kornev, A. P., Taylor, S. S., Amaro, R. E.,

Electrostatic interactions as mediators in the allosteric activation of protein kinase A RI α , *Biochemistry*, vol. 56, 2017”. The dissertation author was the primary investigator and primary co-author of this paper.

Chapter 3, in full, is a modified reprint of the material as it appears in “Barros, E. P., Schiffer, J.M., Vorobieva, A., Dou, J., Baker, D., Amaro, R. E., Improving the Efficiency of Ligand-Binding Protein Design with Molecular Dynamics Simulations, *Journal of Chemical Theory and Computation*, vol. 15, 2019”. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material as “Barros, E. P., Wych, D., Wall, M. E., Mobley, D., Amaro, R.E., On the conformational diversity within protein crystals”. The dissertation author is the primary investigator and author of this paper.

Chapter 5, in full, is currently being prepared for submission for publication of the material as “Barros, E. P., Demir, O., Amaro, R.E., Towards mutant-specific therapies: Uncovering the dynamical landscape and druggability of p53 DNA binding domain”. The dissertation author is the primary investigator and author of this paper.

VITA

- 2012 Visiting International Student
Barnard College of Columbia University, New York, NY
- 2014 Bachelor of Science in Chemistry
University of Campinas, Brazil
- 2015 Master of Science in Physical Chemistry
University of Campinas, Brazil
- 2019 Graduate Student Researcher
GlaxoSmithKline, Upper Providence, PA
- 2020 Doctor of Philosophy in Chemistry
University of California San Diego, San Diego, CA

ABSTRACT OF THE DISSERTATION

Seeing the Unseen:

Incorporating Dynamics for the Understanding and Prediction of Macromolecular Properties

by

Emília Pécora de Barros

Doctor of Philosophy in Chemistry

University of California San Diego, 2020

Professor Rommie E. Amaro, Chair

Professor Susan S. Taylor, Co-Chair

The passing of time is a constant. In the same way that time shapes our everyday lives, it also plays an essential role in the microscopic universe through its translation into macromolecular dynamics. Proteins, in particular, visit a variety of short-lived - or metastable - states, and the conformations and rate of transitions between these states dictate their function in the cell. However, many of these alternative states are not accessible to established biophysical techniques,

such that many puzzling results and challenges arise when trying to use these static observations to understand and predict their properties. In this dissertation, we emphasize the importance of accessing and characterizing the intrinsic dynamics of these macromolecules through the application of computational techniques in four distinct projects that allow us to observe molecular motion at the atomic level to widen our understanding of their function, and furthermore enable us to make predictions on properties that would be inaccessible from static models. In conjunction with experimental validation, we study the role of electrostatic interactions in mediating the allosteric activation of Protein Kinase A, an ubiquitous enzyme involved in key signal transduction pathways, as well as develop a methodology that incorporates dynamic descriptors to improve the efficiency of the challenging process of novel protein design. Additionally, through the application of the exciting Markov State Model methodology to access longer timescales than typically available to computational simulations, we help set the stage for the interpretation of a novice experimental technique, diffuse scattering, in terms of atomic motions in crystalline environments, and explore the conformational landscape of the essential tumor suppressor p53 and its cancer-related mutant Y220C, which leads to the identification of a novel cryptic pocket. These test cases present significant advances to our understanding of protein dynamics in fields as varied as protein design and cancer therapeutics, and taken together, stress the necessity of putting time, and dynamics, in the center stage of protein study.

Introduction

If there is one certainty in life, it is that time passes. The ever-alternating seasons, our constant aging, and even the hours that just flutter by in a day evidence how time is constantly progressing, and things are in non-stop motion. Time plays an import role in the microscopic universe as well. Proteins are the main orchestrators of life, and the effect of time, translated into their dynamics, is essential for them to carry out their important roles.

Despite an ever increasing knowledge of these protein's structures afforded from a range of experimental techniques, dynamics is essential to be incorporated in order for a mechanistic understanding of their function as well as the ability to make informed predictions of the effects that changes such as mutations have in their function or other properties. In support of this, knowledge of the specific fold adopted by a given protein does not directly imply a function¹. Dynamics need to be taken into consideration in most cases in order to fully understand processes such as protein activation, macromolecular recognition and binding, the evolution of diseases, and how to treat them. In this introductory chapter, the relationship between protein function and dynamics and the concept of free energy landscape are discussed in more detail, as well as the theory behind the computational techniques used in the dissertation to probe molecular dynamics. We finalize with a brief introduction to a novel experimental methodology, diffuse scattering, that can be useful in providing information on protein correlated motions in conjunction with computational models.

1.1 The link between protein function, dynamics, structure and sequence

Proteins are diverse, intricate and essential macromolecules present in every cell that are governed by a complex and still-elusive set of fundamental interactions². To borrow from Jane Austen, “it is a truth universally acknowledged” - at least in the scientific world - that a protein’s function is determined by its structure, which in turn, are a consequence of the specific set of interactions between its forming sequence of amino acids. In the past several decades, an increasing amount of experimental and computational data has expanded that sequence-function understanding to incorporate a fourth dimension, time, on the determination of protein function³⁻⁷.

It is now widely appreciated that function is therefore not only a consequence of protein structure, but also intimately linked to its dynamics, in the form of varied motions that span large time and spatial scales, from side chain reorientation and loop motions in the ns- μ s timescale to large domain or subunit motions in the in the ms, seconds or even longer scales. These motions include direct as well as allosteric, or “through distance”, effects⁸. Instead of a single structure at equilibrium, the current view is of proteins as flexible and mobile entities⁹, with the conformational transitions between the long-lived, or metastable states, being essential in ligand and macromolecular recognition¹⁰, catalysis¹¹⁻¹³, transcriptional activity¹⁴, signal transduction¹⁵⁻¹⁸, allostery¹⁹⁻²¹, and a wide variety of protein functions²²⁻²⁶. The dynamical states accessed by the protein are thus an intrinsic property of the macromolecule, conferring almost a “personality” to it⁴.

The sequence and structural encoding of this dynamic information²⁷ is so complex that efforts to accurately predict effects of mutations on macromolecular properties or design proteins with completely novel functions are still in their infancy. The fact that a single point mutation in

many essential enzymes can lead to life-threatening or debilitating diseases²⁸⁻³⁰ emphasizes the extent to which this sequence-function link is fine-tuned and specific. The deleterious effect to function can be achieved through a variety of ways, oftentimes interconnected, such as alterations to macromolecular stability (found for example in cancer³¹, Parkinson's³², Alzheimer's³³, muscular³⁴ and retinal³⁵ diseases), conformational dynamics (diabetes³⁶, cancer^{37,38}, leukemia³⁹ and cystic fibrosis⁴⁰), hydrogen bonding network (cataract⁴¹ and Alzheimer's disease⁴²) and functionally-important residues (cancer⁴³). The current interpretation of protein functional dynamics and the timescales involved revolves around the concept of free energy landscape, and will be discussed in the next section.

1.2 The free energy landscape and timescales of motions

The range of conformational states that a protein can access are determined by a multidimensional free energy landscape. In the same way that the time it takes for a person to get from point A to point B depends on the topology of the terrain (the presence of a mountain between them would make the journey considerably longer, for example), the energy barriers between protein states determine rates of transition and their relative populations (Figure 1.1). The landscape is an intrinsic property of the macromolecule, a consequence of the specific sets of interactions between the atoms making up its sequence and defined by local structure, but also tied to a set of conditions such as temperature and solvent, and can be significantly altered by mutations, ligand binding or interactions with partners^{19,20}. Each basin represents stable states, and the energy barrier between the states determines the rate of transition between them.

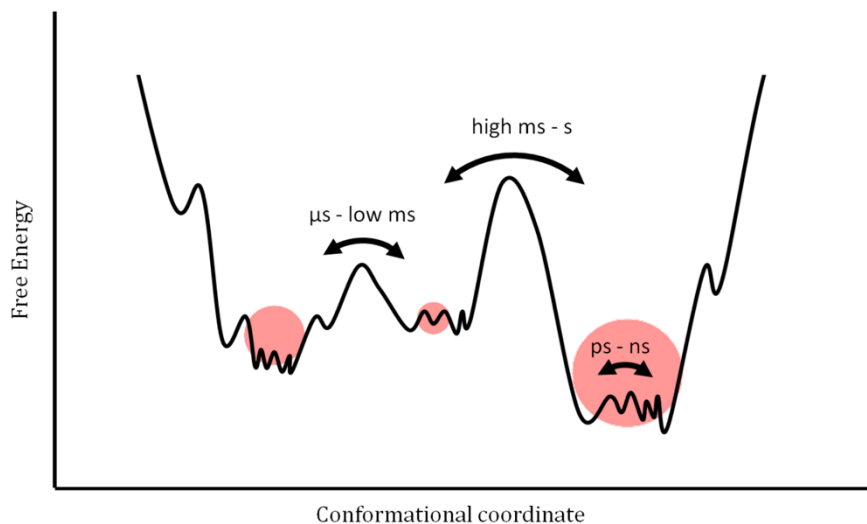


Figure 1.1. Free energy landscape representation and associated timescales. Image extracted from Göbl and Tjandra, *Entropy* (2012)⁸.

Investigation of the thermodynamics and kinetics associated with a protein's conformational landscape is at the core of biophysical techniques, since the characterization of the conformational states are what allow the understanding of protein function, mutational effects on this function, drug design efforts, and all other goals involved in biochemistry research. Each technique has their strengths and weaknesses, and incredible breakthrough can be achieved when they can be used in conjunction to solve piece by piece this complicated puzzle. Experimental techniques such as X-ray crystallography provide great understanding on the structural aspects of states that can be captured in crystals, although limited information on dynamics, while other methods such as NMR spectroscopy can inform on timescales of the transitions but with occasional loss of the atomic scale detail.

With the advancement of technology, computational methods have been added to the chemistry toolbox, providing the significant advantage of allowing the observation of all atoms in motion and thus unbeatable insights into the intrinsic dynamics of macromolecules. These methods

are not, however, without their own limitations. The remainder of this chapter will discuss in more detail two computational techniques used extensively in the projects presented in later chapters, as well as briefly introduce a relatively novel experimental technique, X-ray diffuse scattering, that is of relevance to one of the studies in this thesis and provides a novel opportunity for unifying structural and dynamic resolution for the understanding of these macromolecules' "personalities".

1.3 MD simulations to “see” protein dynamics at the atomic level

Molecular dynamics (MD) simulations are an important computational tool that allows obtaining information on the dynamical behavior of several macromolecular systems at the atomic scale⁴⁴. After over 40 years since the first protein MD simulations⁴⁵, an incredible variety of systems can today be simulated realistically, from single proteins and lipid membranes^{46,47} to large and increasingly complex multiscale assemblies⁴⁸. By providing information on dynamics at the atomic scale, MD simulations and other computational methods empower researchers to comprehend oftentimes puzzling experimental results or motivate further experimental work⁴⁹.

Classical MD simulations are based on the representation of atoms as spherical particles connected by nodes and subjected to intra- and intermolecular interactions with other atoms in the system. Through the calculation of the potential energy defined by these interactions and application of Newton's laws of motions, successive configurations of the system can be obtained. The position and velocities of the particles at each instant in time constitute a trajectory, from which the temporal evolution of properties of the system can be studied. The system dynamics are modeled by empirical force fields⁵⁰, and allow the investigation of a large range of processes, including motions that are difficult to be investigated experimentally⁵¹.

In the force fields, the atomic interactions are treated classically, based on several assumptions such as the Born-Oppenheimer approximation⁵², which states that the motion of

nuclei and electrons of an atom can be treated separately, and allow the simplification of the calculation from the quantum mechanical level to the representation of atoms as spheres with associated mass, charge and volume and connected to other atoms by springs. This reproduces in a simplified but efficient manner the interactions in a microscopic level, and allow the calculation of motions in larger systems and timescales than would be accessible with quantum mechanical calculations.

The potential energy of the system is a function of the atomic positions at each instant in time, based on intra- and interatomic interactions defined by the force field equations. Different functional forms of the force fields exist⁵³⁻⁵⁸, but they all take the general form presented in Figure 1.2⁵⁹. Intramolecular (through-bond) interactions, in the form of bond stretching, angular deformations and dihedral torsions, are represented with harmonic potentials (for bond and angle potentials) and sum of trigonometric functions to represent the periodic potential of dihedral torsions. Interactions between atoms located further away are described by the Coulombic (for electrostatic interactions) and Lennard-Jones (for van de Waals interactions) potentials⁶⁰. The equation's specific form, its constants and parameters are slightly different among the distinct force fields available, depending upon the experimental measurements and *ab initio* quantum mechanical calculations from which they are derived and validated^{61,62}.

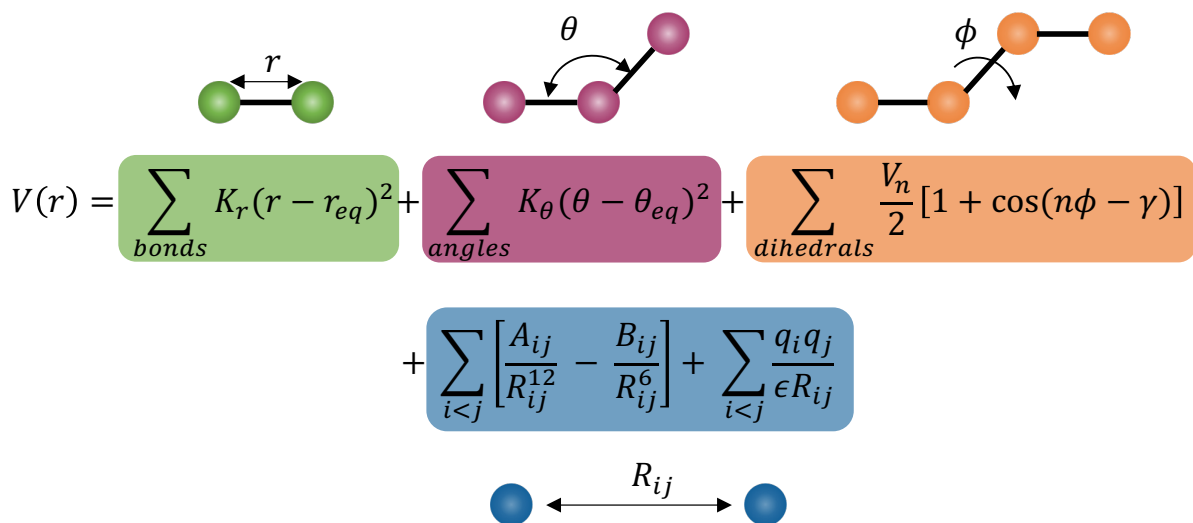


Figure 1.2. Example of force field equation used for calculation of potential energy in MD simulations. The bonded (atom, angle and dihedral) as well as non-bonded (van der Waals and electrostatic) potentials are highlighted with respective schematic representations.

An area of active research is the constant validation against experiment^{63,64} and improvement of force fields to reach ever more realistic representations of solvent and molecular interactions⁶⁵⁻⁶⁹. However, generally a more accurate force field comes at the expense of sampling time. In spite of its limitations, MD simulations have carved an important presence in biochemistry and drug discovery research, enabling for example the discovery of cryptic pockets not evident in experimental crystal structures^{70,71} which led to the development of drug leads⁷²⁻⁷⁴ and approved drugs⁷⁵, an increased understanding of aerosol chemistry⁷⁶, and many other applications.

1.4 Markov State models to extract kinetic and thermodynamic information from MD simulations

Despite incredible advances in computer software and hardware, several biological processes of relevance for proteins and macromolecular systems take place at timescales that are difficult or impossible to be accessed by classical MD simulations, given its computational cost.

A large number of methods aim to surpass this sampling limitation by promoting enhanced sampling of the system dynamics, in which the high energy barriers of the system can be more easily overcome through measures such as locally increasing the potential energy (employed in Metadynamics⁷⁷, accelerated MD⁷⁸ or Gaussian accelerated MD⁷⁹), or the temperature of the system (such as in Replica Exchange MD⁸⁰). Other computational methods employ biased or directed sampling through application of force in a determined degree of freedom, Umbrella Sampling⁸¹ being one example, or simplify the potential energy and the system representation through coarse-graining approaches⁸². These methods allow for a more thorough exploration of the conformational landscape or process under investigation, but in general have the disadvantage of losing the ability to inform on thermodynamics of the system since the potential energy is disrupted. Reweighting techniques can sometimes be used to overcome this, but in general they result in complicated data post-processing.

An alternative and emerging technique for exploration of processes that take place at long timescales are Markov State models, or MSMs (Figure 1.3), which have been used successfully to model protein folding^{83–86}, ligand binding^{87,88}, dynamics of cryptic pockets⁸⁹, and protein conformational change^{16,90,91}. In an MSM, multiple classical MD simulations are integrated in a single, statistically rigorous mathematical framework of the probability of transitions between states that can inform on motions that occur at timescales longer than the individual.

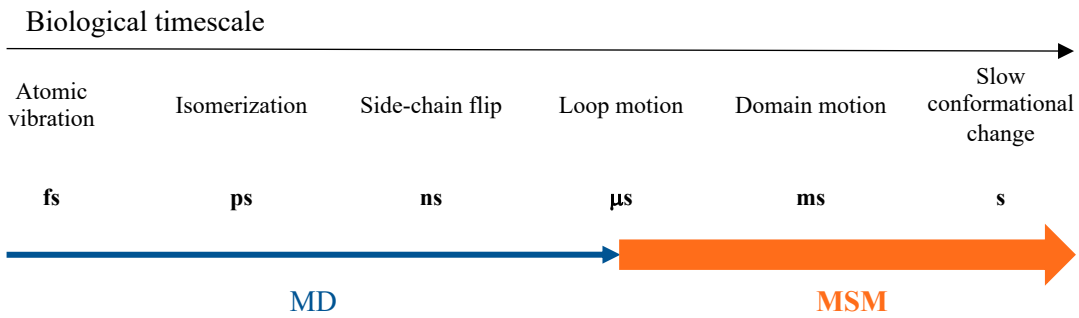


Figure 1.3. Typical biochemical motions and associated timescales. MSMs have the ability to extend the timescales that are accessible to MD simulations.

The enhanced sampling is performed by iteratively spanning multiple independent simulations from conformations located throughout the MD-accessed free energy landscape, in an effort to optimally explore different regions of the landscape⁹². In this way, the process under investigation does not have to be sampled in a single simulation (which can be hard or impossible in most cases), but can be broken down into independent trajectories that explore the local area (Figure 1.4). Because these are conventional simulations that retain the original form of the system's free energy landscape, the thermodynamics of the system are not affected, and equilibrium populations of the metastable states can be recovered. Additionally, as the method considers rate of transitions between states, kinetic information in the form of relaxation timescales or mean first passage times can also be extracted, creating the opportunity for validation and comparison with experimental observables^{93,94}.

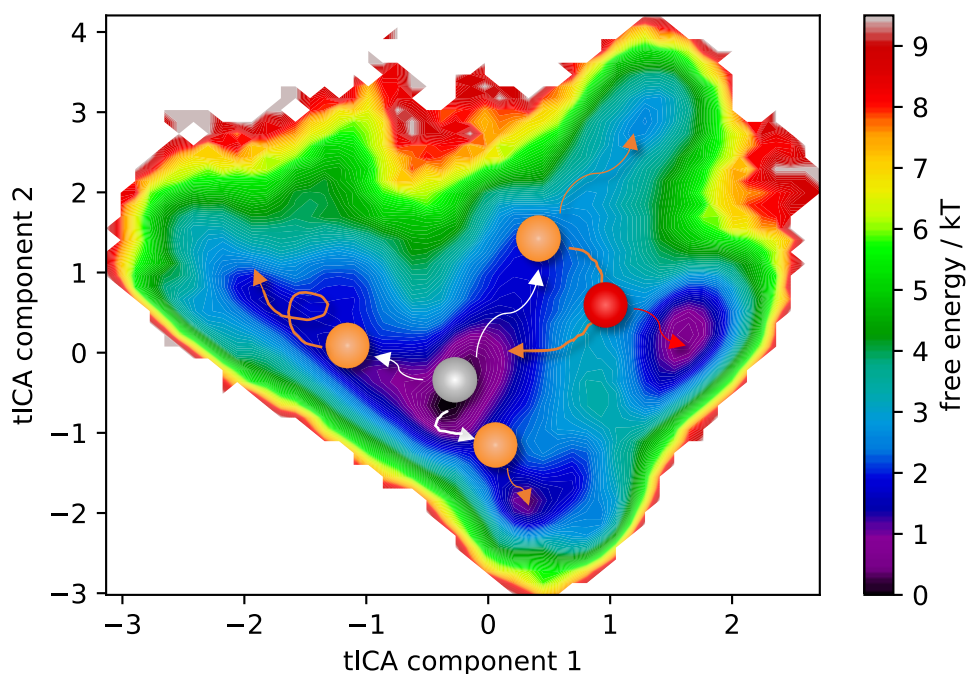


Figure 1.4. Representation of adaptive sampling performed for MSM construction. Sequential rounds of MD simulations are represented in distinct colors (white, followed by orange and then red), starting from initial configurations represented as circles. The free energy landscape is explored by performing several short simulations instead of a single very long simulation.

MSMs also present the additional advantage that, by coarse-graining the different conformations explored by the system not only structurally, but also kinetically, into metastable states, a more human-interpretable view of the dynamics under study can be obtained, a significant improvement considering the large amounts of data that is collected in long simulations. In this way, they emerge as a versatile post-processing tool for the objective extraction of meaningful information from multiple MD simulations^{95,96}.

However, despite its many advantages, construction of MSMs is not a trivial task and require a considerable degree of expertise and intuition by the modeler, not to mention significant accrued simulation time for proper statistics. Tutorials and MSM Python libraries⁹⁷⁻⁹⁹ are decreasing the barrier to access, but difficulties remain in the selection of the large number of

parameters that are taken for model construction. In the next subsections the steps involved in model construction and some of the challenges are discussed in more detail, as well as the mathematical basis behind the methodology and model validation metrics used to assure accurate modelling of the process under investigation. This summary is by no means intended to be exhaustive or mathematically heavy, for which a number of good reviews can be recommended¹⁰⁰⁻¹⁰³. Instead, I aim to provide a ‘big-picture’ overview of the methodology and the procedure involved for those not so familiar with it.

Feature selection, dimensionality reduction and space discretization

The first step in MSM construction is the decision of descriptors, or features, used to characterize the conformations the system access during the simulations. The coordinates of all atoms cannot be used for this purpose given the high dimensionality of the state space, and thus a smart decision on the best collective variables that describe the motions under investigation is essential so that the high dimensionality of the system can be accurately mapped to a low-dimensionality space. Typically used features include pairwise distances between selected residues in regions of relevance in the protein structure^{88,104}, RMSDs to reference states (such as inactive and active states⁸⁷), dihedral angles or contact maps (used commonly in protein folding studies^{86,105}) and inter-molecular distances (such as protein-ligand distances for investigation of ligand binding¹⁰⁶). Methods have been proposed to aid the selection of the features that capture the best representation of the slow transitions^{107,108}, but this is a process still greatly led by the researcher’s intuition and knowledge of the system. All subsequent steps and ultimately the MSM depend on this selection of features, as the quality of an MSM is dependent on the state

decomposition, and so it is very important to construct a state space that accurately captures the kinetics of the underlying system.¹⁰⁴

Many times, the number of features is still too large to be used as input for discretization of the conformational space. To that end, researchers typically do a subsequent step of dimensionality reduction, processing the features by Principal Components Analysis (PCA) or time-lagged Independent Component Analysis (tICA). While PCA identifies the linear combination of input parameters that best explain the variance in the data¹⁰⁹, tICA finds the degrees of freedom that maximize the covariance in the input¹⁰⁸. Ultimately, this results in the identification of tICA components (or tICs) that describe the slowest motions of the system^{104,110}. This is really useful when trying to study motions at long timescales, which is the case in most MSM applications, and the reason why tICA is becoming increasingly more popular and used even when the initial set of features is not that elevated.

Finally, once the features and possibly-transformed components have been selected, the many conformations sampled during the simulations are discretized using clustering methods, giving rise to the microstates, or clusters. With this, the different frames in the simulations can now be assigned labels according to the microstates to which they belong, and the trajectories have been converted from a series of structures over time to a series of discrete states over time. These steps are represented schematically in Figure 1.5a-c.

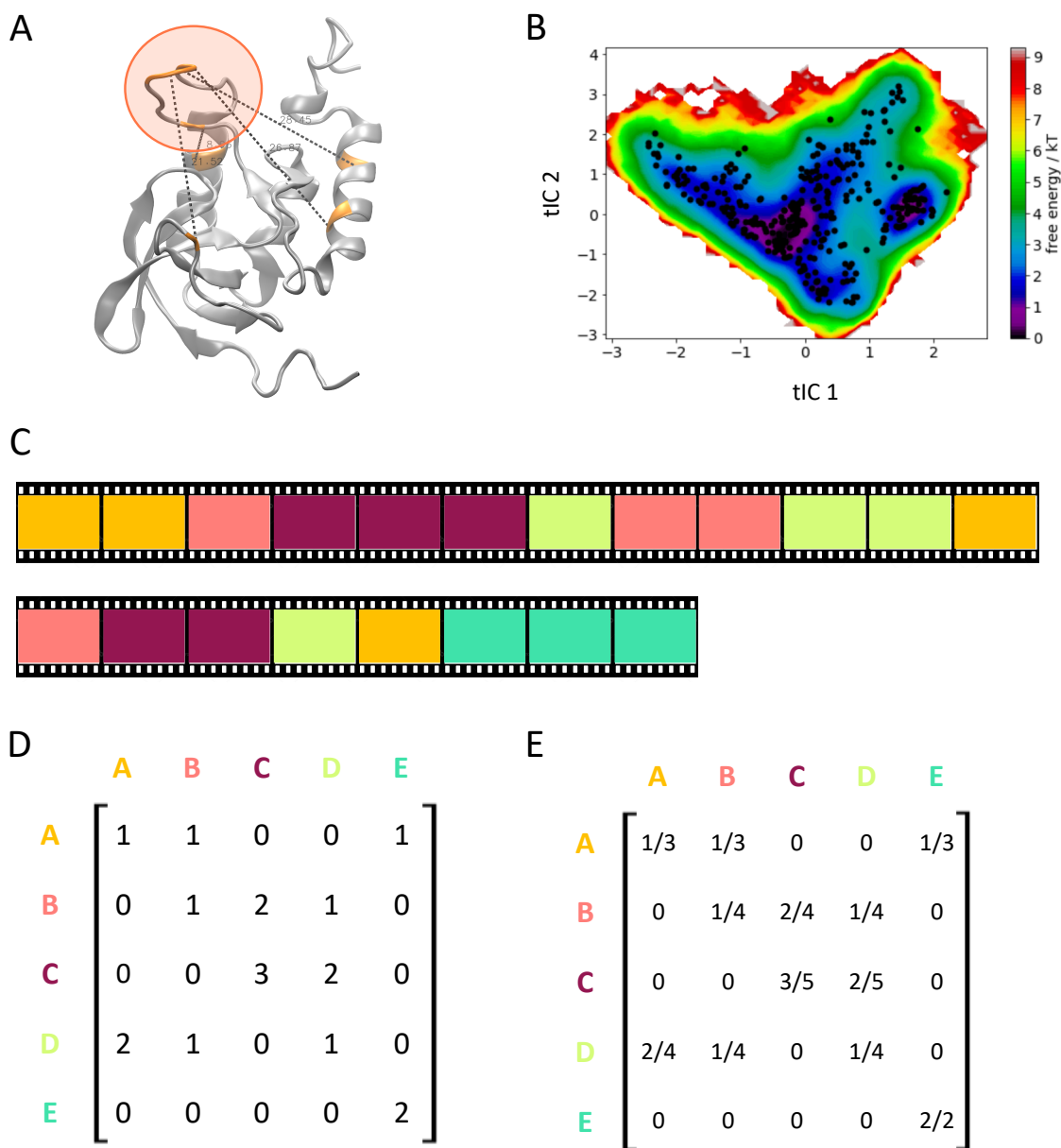


Figure 1.5. Representation of the steps involved in MSM construction for an example model in which there is interest in characterizing the motion of a central flexible loop, highlighted. (a) Features selected for model construction, which correspond to four pairwise distances involving residues in the loop and nearby secondary motifs. (b) Free energy landscape in terms of the tICA-transformed features, overlaid with the cluster centers (black circles) obtained from clustering of the trajectory data using k-means clustering. (c) Simplified representation of two trajectories in terms of the discretized cluster labels. Each frame is represented as a frame in the “movie” with colors according to the cluster to which they belong to. For simplification only 5 clusters are shown. (d) Count matrix obtained from the representative trajectories in c. Each line correspond to a origin states, and column states to the destination state in the transition. (e) Transition probability matrix calculated from the example count matrix in d.

Calculation of the transition probability matrix and implied timescales

The core of the MSM theory is the transition probability matrix. Once the simulations have been discretized into microstates, or clusters, the transition between these states are counted for each independent trajectory. To ensure that the system falls under the Markovian assumption, that is, that it has no memory and thus the probability that it transitions to a determined state is independent to which state it was in the past, and only on the current state, a lag time τ needs to be used for transition counts. The lag time basically determines the stride, or jump, between frames that are used for transition count. A large lag time is desired as it ensures that the system is memory-less, but too large a lag time results in loss of data as more and more frames are discarded for analysis. This is another critical step in model construction as different jumps result in different count matrices. Implied timescale plots are used to aid the selection of the appropriate MSM lag time, and also ensure the “Markovianity” of the system, as will be discussed below.

Transitions between microstates separated by the lag time are counted for each simulation and then summed to result in a matrix that contains total transitions between every two pair of microstates. Figure 1.5d represents the resultant count matrix for the example trajectories provided with a lag time of 1. Dividing by the total number of transitions originated from the initial macrostate, we obtain the transition probability matrix (Figure 1.5e). It is important to highlight that it is at this single stage that the data between multiple independent trajectories can be integrated, because if the Markovian assumption is satisfied, transition counts can be considered statistically independent.

The eigenvectors and eigenvalues of the discrete transition probability matrix characterize the relaxation processes involved in the system’s dynamics¹⁰¹. In this way, the complete dynamics is a composition of the dynamical processes represented by the eigenvalue and associated

eigenfunctions. The eigenvalues (λ_i) range from 0 to 1, and indicate the relaxation times of the individual transitions. The closer they are to 1, the slower is the decay of the corresponding process. The first eigenvector, with an eigenvalue of 1, thus corresponds to the microstates' equilibrium distribution. The physical timescales t_i , or the time it takes for the process to decay towards equilibrium, of all other motions can be obtained from the other eigenvalues by

$$t_i = -\frac{\tau}{\ln \lambda_i} \quad (1)$$

In practice, only a reduced number of eigenfunctions contain large eigenvalues (close to one), while the remaining represent much faster dynamics, such that we are interested in the first few m eigenvalues and eigenfunctions that represent the slowest system dynamics (Figure 1.6). Importantly, the second eigenvalue/eigenfunction pair represents the slowest motion of the system.

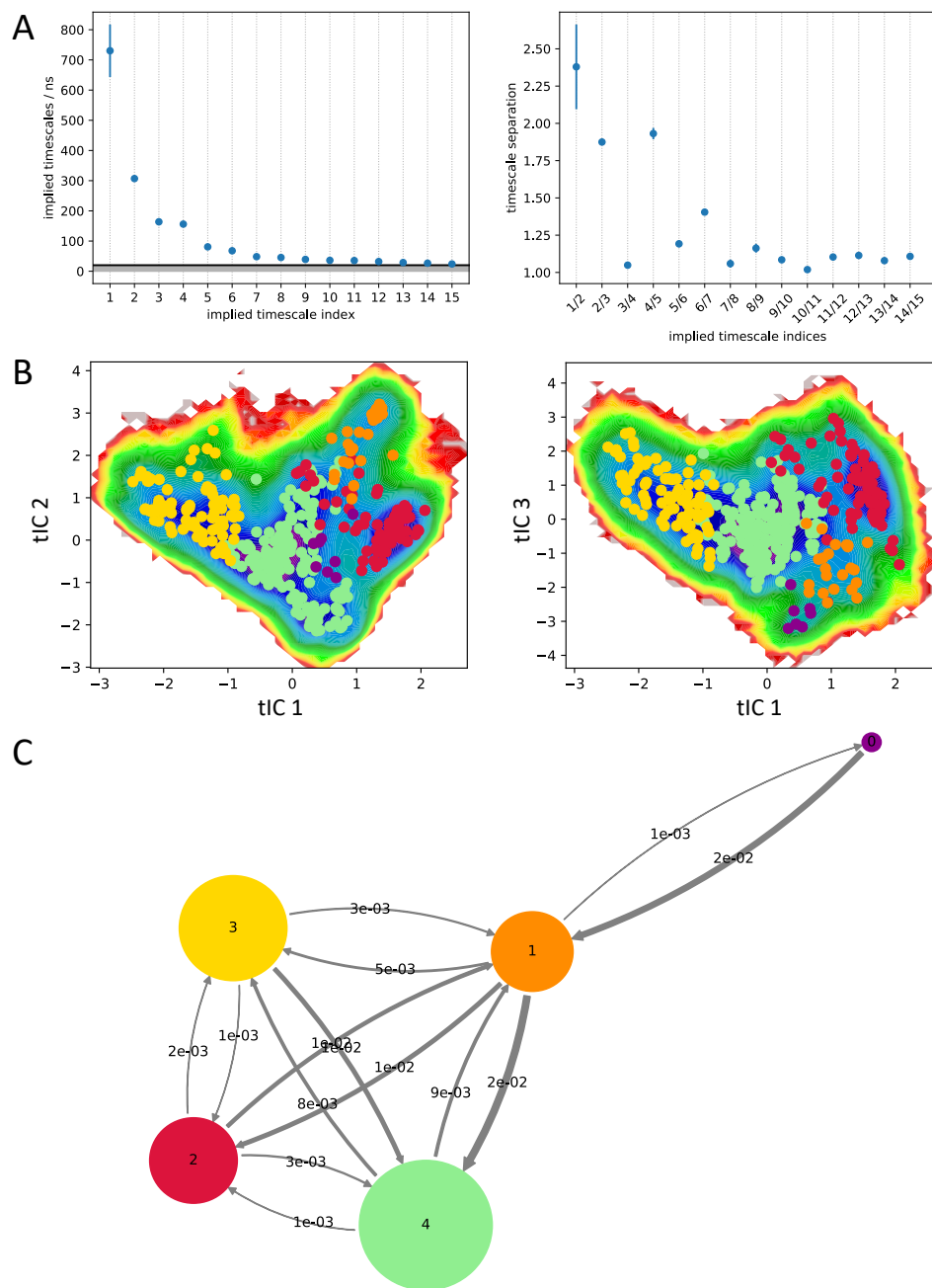


Figure 1.6. Representation of the process involved in identification of metastable states. (a) Implied timescales (left) obtained from the transition probability matrix and relative timescales between subsequent timescales (right). A gap between subsequent timescales indicates the number of metastable states in the system. In this case, a gap between the 4th and 5th timescales suggests the existence of five slow interconverting metastable states. (b) Membership of the clusters from Figure 1.5b to the metastable states. Projections of the multidimensional free energy landscape in terms of the first three components are shown. (c) Coarse-grained, Hidden Markov Model (HMM) obtained for the example system. The radius of the circles is proportional to the state's equilibrium population, and thickness of arrows to flux between states. Values above the arrows indicate number of transitions per MSM lag time.

The standard procedure for identifying the optimal choice of lag time and assuring the Markovian assumption is to monitor the relaxation timescales of the system dynamics as a function of the lag time (through the commonly named implied timescale plots). Convergence of the relaxation timescales indicates sufficiently accrued statistics on the transition regions, and the independence of the timescales on the steps between transitions, which is essentially, the loss of memory of the system.

Identification of metastable states

The number of microstates generated above is useful for the calculation of the transition probability matrix, but it is generally still too high-dimensional for appropriate interpretation of the relevant conformational changes at hand. It is thus very valuable to aggregate kinetically similar microstates into macrostates, or metastable states. However, how does one choose the number of macrostates accurately, without being biased by an often incorrect human interpretation of conformational similarities and the projection in the free energy landscape? The definition of the number of states can be aided by the implied timescales (thus, the eigenvalues) extracted from the transition probability matrix.

The eigenfunctions define the structural transitions associated with each corresponding relaxation timescale, corresponding to a transition between metastable sets¹⁰¹. The number of slow relaxations rates (or eigenvalues above some pre-defined cutoff) solved from the transition probability matrix, thus, indicates the number of metastable states accessed by the system. An appropriate number of macrostates to build can then be selected based on the location of major gaps between the implied timescales⁹⁸. The number of states will be one more than the number of timescales above the selected cutoff (Figure 1.6c). However, these gaps can often be unclear or

very small, in which case several selections of macrostates can be done and validated based on their performance on the model validation CK test (discussed below), as well as their distribution in the free energy map.

Once the number of macrostates have been defined, kinetically similar microstates can be lumped together using Perron Cluster Cluster Analysis (PCCA+)¹¹¹ or Hidden Markov Models (HMMs)^{112,113}, which identify kinetic relationships based on the eigenvalue and eigenvectors of the microstate MSM. PCCA+ assumes that there should be a separation of timescales between slow transitions across high energy barriers, and quicker transitions between microstates located within a basin are therefore coarse-grained into a single metastable state¹¹⁴. In contrast to computing memberships of microstates to metastable sets as in PCCA++, in HMMs we directly obtain a coarse transition matrix and model with fewer states.

Model validation: Implied timescale plots and CK tests

The quality of MSMs are generally analyzed using two tests, in addition to comparison of the modeled dynamics and metastability to experimental observables whenever possible. The first validation method involves the verification that the lag time is sufficiently long to allow a Markovian decomposition of the state space, using the already mentioned implied timescale plots (ITS plots). ITS plots are usually generated for a large variety of initial conformational space feature discretizations, clustering algorithms and number of clusters, since all of these parameters affect the recovered MSM. For a Markovian process, the transition matrix T needs to be independent of the lag time. In practical terms, this is performed by verifying the smoothness and leveling of the implied timescales calculated from the transition probability matrix in the ITS plots, and choosing a lag time for model construction at which constant values have been attained¹¹⁴ (Figure 1.7).

The second validation metric is the Chapman-Kolmogorov (CK) test, which verifies that the dynamical information obtained from the MSM is internally consistent with that observed in the source MD trajectories. In practice, this is done by verifying that the transition matrix estimated at a lag time τ is the same to the transition matrix estimated at a longer lag time $k\tau$:

$$T(k\tau) = T(\tau)^k \quad (2)$$

The probabilities of remaining or transitioning between states is verified at increasing timesteps, and the test is successfully passed if these fall within a $1-\sigma$ standard error of the MD data¹⁰¹.

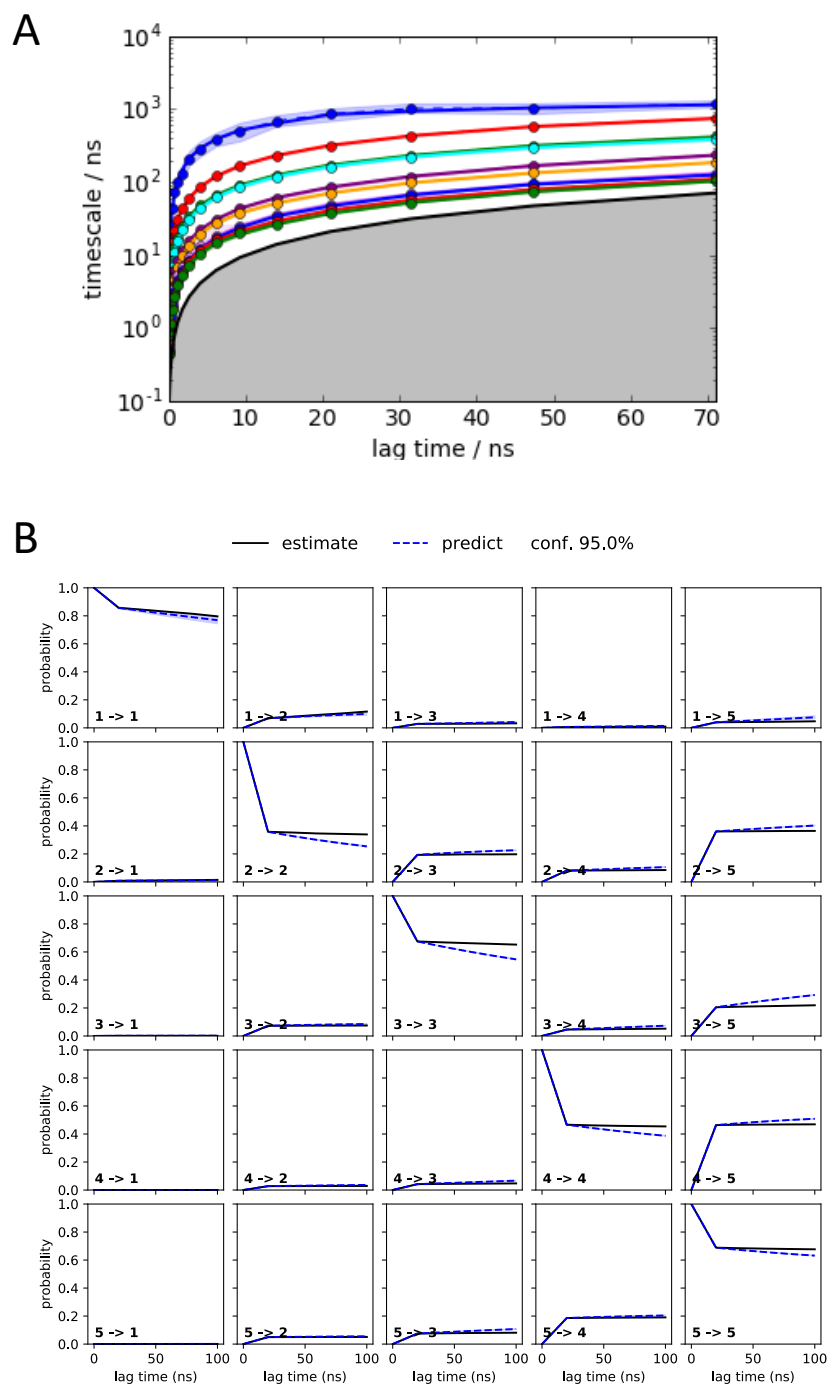


Figure 1.7. MSM validation metrics. (a) Implied timescale plot. (b) CK test. The probability values estimated from the MSM (black lines) should fall within the area of 95% confidence (in blue) extracted from the original simulations.

Extracting kinetic information

A range of kinetic information can be computed from the MSM to be compared to or help interpret experimental data. The overall relaxation timescales of the transitions can be obtained directly from the eigenvalues, as discussed above. Additionally, high-flux pathways as well as bottlenecks can be identified through transition path theory^{115,116}. Mean first passage times, the average timescale for an event (in this case a transition between metastable states) to first occur¹¹⁷, are easily computed from the coarse-grained representation of the MSM.

1.5 Diffuse scattering for studying correlated protein motions

Over the last century, the field of X-ray crystallography has had an extraordinary impact on how we understand biological function by providing perspective on protein structures and illuminating their roles in biological mechanisms¹¹⁸. The data traditionally used for these analyses are the intense Bragg peaks that carry information on the electron density and thus are used to resolve atomic structure. If crystals were perfect and static, the electron density would be identical among all the unit cells, and the Bragg peaks would be the only data obtained in such crystallographic experiments. In practice, however, the intense Bragg peaks are accompanied by cloudy and diffuse features, of much weaker intensities, which are related to imperfections in the crystal and motions that lead to divergence between unit cells^{119,120}. Although being thus a phenomenon quantified for over 100 years in such biophysical experiments, the origin of macromolecular diffuse scattering has been poorly understood, hindering its applicability to structural modeling.

In the past decades, improvements in data acquisition, leading to better signal to noise ratio in the less intense diffuse features, have renewed interest in the interpretation and establishment

of diffuse scattering as an additional biophysical experiment to aid macromolecular structure and dynamics determination¹²¹. Diffuse scattering is related to both inter- and intramolecular correlations¹²² (Figure 1.8): Single chain rotational and translational movements within the unit cell seem to be important, as well as correlated motions and fluctuations between unit cells, which can include changes in protein conformation^{123–125}. Investigation of the contributions that each of these fluctuations play in the measured diffuse intensity is an area of active research.

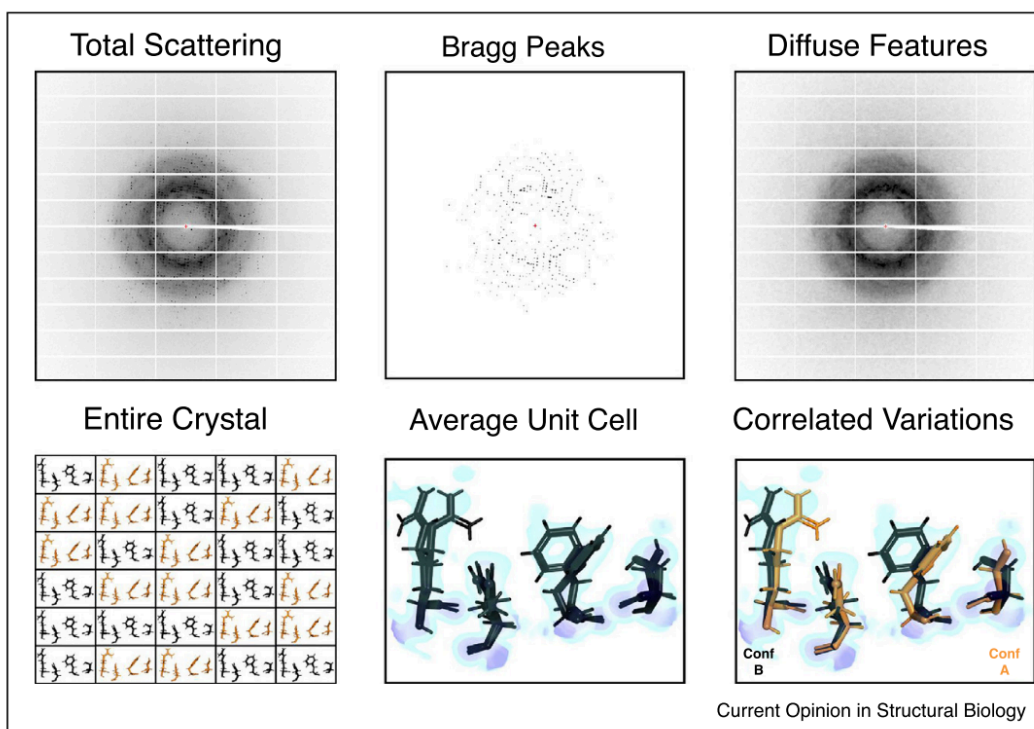


Figure 1.8. Diffuse scattering versus Bragg data. Image extracted from Wall et al, Curr. Op. Str. Biol. (2018)¹²⁰.

In order to investigate the physical origin of diffuse scattering, several models of molecular motion have been proposed and studied^{123–129}. Because of its ability to represent protein structures and intermolecular interactions in the atomic level, MD simulations have been a natural choice to

investigate these processes, although crystalline simulations are much less common than solution simulations and concerns about the accuracy of force fields are even more relevant in these compact environments^{120,122,130}. However, despite its limitations, MD simulations have successfully provided insights of protein fluctuations in crystals and recovered, at least partially, the diffuse intensity measured experimentally^{131–137}.

The diffuse intensity is computed from snapshots of the simulations using established methods as the variance of the unit cell structure factor (F_{hkl})^{137,138}:

$$I_{diffuse}(hkl) = \langle |F(hkl) - \langle F(hkl) \rangle|^2 \rangle \quad (3)$$

Or, expanding the expression:

$$I_{diffuse}(hkl) = \langle F(hkl)^2 \rangle - \langle F(hkl) \rangle^2 \quad (4)$$

The Bragg intensities are determined by the square of the mean structure factor (the second term in Eq. 4), and this equation demonstrates the distinction of the diffuse intensity from the Bragg peaks¹³³. The structure factor, an essential metric in crystallography, is determined at each Miller index hkl as the sum over the x_j , y_j and z_j positions of each atom j in the unit cell:

$$F(hkl) = \sum_{j=1}^N f_j e^{[-2\pi i(hx_j + ky_j + lz_j)]} \quad (5)$$

where f_j is the scattering factor of atom j . Diffuse scattering is emerging thus as an additional experimental technique that allows the connection to computational methods and probing of macromolecular dynamics.

1.6 Summary

In this dissertation, molecular dynamics simulations, Markov State models and other computational techniques are used to incorporate the dimension of time in the study of proteins in four distinct case studies. The dynamic information accessed by these computational tools allows not only a deeper understanding of these macromolecules' intrinsic "personalities", but also the rationalization of the connection between their sequence, dynamics and function, or prediction of key properties based on their dynamical fingerprints. Specifically, the following chapters will detail studies on:

- Understanding the effect of a positively-charged patch in the allosteric activation of the regulatory unit of protein kinase A (PKA), an ubiquitous enzyme involved in regulation of key cellular metabolism^{139,140}. Computational mutagenesis evidences the effect of the charged residues on the protein's dynamic profiles and conformational ensembles, enabling predictions of their effect on protein function which were confirmed experimentally.
- The development of a methodology for improving the efficiency of *de-novo* ligand-binding protein design for applications in biosensor and enzyme design^{141,142}. The application of machine learning approaches on the dynamical fingerprints recovered from MD simulations of the protein designs allow the prediction of the binding ability of these non-natural proteins, thus helping to increase the success rates of this challenging endeavor.

- Incorporating an ensemble-level description of the conformational flexibility accessed by proteins in a crystal using Markov State models to understand the degree of conformational diversity and long-range communication in crystalline environments. This work sets the stage for the connection between MD simulations, MSM and X-ray diffuse scattering for understanding the origins of correlated motions in protein crystals and the rationalization of experimental diffuse scattering in terms of atomic motions for better structure refinement.
- The full characterization of the conformational ensemble of the tumor suppressor p53¹⁴³⁻¹⁴⁵ and the important cancer mutant Y220C^{146,147} using MD and MSMs. The thermodynamic and kinetic information accessed by the MSMs suggests the existence of allosteric communication within the DNA binding domain that could explain the effect of the mutation on the abrogation of p53 function, as well as uncovers novel protein conformations and a cryptic pocket of relevance for cancer therapeutics and p53 reactivation efforts.

Details of the specific proteins under study will be given in the respective chapters.

1.7 References

- (1) Lee, D.; Redfern, O.; Orengo, C. Predicting Protein Function from Sequence and Structure. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 995–1005.
- (2) Huang, P.; Boyken, S. E.; Baker, D. The Coming of Age of de Novo Protein Design. *Nature* **2016**, *537*, 320–327. <https://doi.org/10.1038/nature19946>.
- (3) Changeux, J.-P.; Edelstein, S. J. Allosteric Mechanisms of Signal Transduction. *Science (80-.)*. **2005**, *308*, 1424–1428. <https://doi.org/10.1126/science.1108595>.
- (4) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450*, 964–972. <https://doi.org/10.1038/nature06522>.
- (5) Smock, R. G.; Gierasch, L. M. Sending Signals Dynamically. *Science (80-.)*. **2009**, *324*, 198–203.
- (6) Osawa, M.; Takeuchi, K.; Ueda, T.; Nishida, N.; Shimada, I. Functional Dynamics of Proteins Revealed by Solution NMR. *Curr. Opin. Struct. Biol.* **2012**, *22*, 660–669. <https://doi.org/10.1016/j.sbi.2012.08.007>.
- (7) Zhuravlev, P. I.; Papoian, G. A. *Protein Functional Landscapes, Dynamics, Allostery: A Tortuous Path towards a Universal Theoretical Framework*; Access paid by the UC San Diego Library, 2010; Vol. 43. <https://doi.org/10.1017/S0033583510000119>.
- (8) Göbl, C.; Tjandra, N. Application of Solution NMR Spectroscopy to Study Protein Dynamics. *Entropy* **2012**, *14*, 581–598. <https://doi.org/10.3390/e14030581>.
- (9) Vendruscolo, M. Determination of Conformationally Heterogeneous States of Proteins. *Curr. Opin. Struct. Biol.* **2007**, *17*, 15–20. <https://doi.org/10.1016/j.sbi.2007.01.002>.
- (10) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* **2009**, *5*, 789–796. <https://doi.org/10.1038/nchembio.232>.
- (11) Joseph, D.; A., P. G.; Karplus, M. Anatomy of a Conformational Change: Hinged “Lid” Motion of the Triosephosphate Isomerase Loop. *Science (80-.)*. **1990**, *249*, 1425–1428.
- (12) Faber, H. R.; Matthews, B. W. A Mutant T4 Lysozyme Displays Five Different Crystal Conformations. *Nature* **1990**, *348*, 263–266. <https://doi.org/10.1038/348263a0>.
- (13) Bennett, W. S.; Steitz, T. A. Glucose-Induced Conformational Change in Yeast Hexokinase. *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 4848–4852. <https://doi.org/10.1073/pnas.75.10.4848>.
- (14) Offut, T. L.; Jeong, P. U.; Demir, O.; Amaro, R. E. Dynamics and Molecular Mechanisms of P53 Transcriptional Activation. **2018**. <https://doi.org/10.1021/acs.biochem.8b01005>.
- (15) Foda, Z. H.; Shan, Y.; Kim, E. T.; Shaw, D. E.; Seeliger, M. A. A Dynamically Coupled

- Allosteric Network Underlies Binding Cooperativity in Src Kinase. *Nat. Commun.* **2015**, *5*, 5939. <https://doi.org/10.1038/ncomms6939>.
- (16) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation Pathway of Src Kinase Reveals Intermediate States as Targets for Drug Design. *Nat. Commun.* **2014**, *5*, 3397. <https://doi.org/10.1038/ncomms4397>.
- (17) Kim, C.; Cheng, C. Y.; Saldanha, S. A.; Taylor, S. S. PKA-I Holoenzyme Structure Reveals a Mechanism for CAMP-Dependent Activation. *Cell* **2007**, *130*, 1032–1043.
- (18) Boras, B. W.; Kornev, A.; Taylor, S. S.; McCulloch, A. D. Using Markov State Models to Develop a Mechanistic Understanding of Protein Kinase A Regulatory Subunit RI α Activation in Response to CAMP Binding. *J. Biol. Chem.* **2014**, *289* (43), 30040–30051. <https://doi.org/10.1074/jbc.M114.568907>.
- (19) Popovych, N.; Sun, S.; Ebright, R. H.; Kalodimos, C. G. Dynamically Driven Protein Allostery. *Nat. Struc. Mol. Biol.* **2006**, *13*, 831–838. <https://doi.org/10.1038/nsmb1132>.
- (20) Manley, G.; Loria, J. P. NMR Insights into Protein Allostery. *Arch. Biochem. Biophys.* **2012**, *519*, 223–231. <https://doi.org/10.1016/j.abb.2011.10.023>.
- (21) Tsai, C.-J.; Nussinov, R. A Unified View of “How Allostery Works.” *PLoS Comput. Biol.* **2014**, *10*, e1003394. <https://doi.org/10.1371/journal.pcbi.1003394>.
- (22) Min, W.; Luo, G.; Cherayil, B. J.; Kou, S. C.; Xie, X. S. Observation of a Power-Law Memory Kernel for Fluctuations within a Single Protein Molecule. *Phys. Rev. Lett.* **2005**, *94*, 198302. <https://doi.org/10.1103/PhysRevLett.94.198302>.
- (23) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. Intrinsic Dynamics of an Enzyme Underlies Catalysis. *Nature* **2005**, *438*, 117–121. <https://doi.org/10.1038/nature04105>.
- (24) Neubauer, H.; Gaiko, N.; Berger, S.; Schaffer, J.; Eggeling, C.; Tuma, J.; Verdier, L.; Seidel, C. A. M.; Griesinger, C.; Volkmer, A. Orientational and Dynamical Heterogeneity of Rhodamine 6G Terminally Attached to a DNA Helix Revealed by NMR and Single-Molecule Fluorescence Spectroscopy. *J. Am. Chem. Soc.* **2007**, *129*, 12746–12755. <https://doi.org/10.1021/ja0722574>.
- (25) Gansen, A.; Valeri, A.; Hauger, F.; Felekyan, S.; Kalinin, S.; Tóth, K.; Langowski, J.; Seidel, C. A. M. Nucleosome Disassembly Intermediates Characterized by Single-Molecule FRET. *Proc. Natl. Acad. Sci.* **2009**, *106*, 15308–15313. <https://doi.org/10.1073/pnas.0903005106>.
- (26) Gebhardt, J. C. M.; Bornschlöggl, T.; Rief, M. Full Distance-Resolved Folding Energy Landscape of One Single Protein Molecule. *Proc. Natl. Acad. Sci.* **2010**, *107*, 2013–2018. <https://doi.org/10.1073/pnas.0909854107>.
- (27) Bahar, I.; Lezon, T. R.; Yang, L.-W.; Eyal, E. Global Dynamics of Proteins: Bridging

between Structure and Function. *Annu. Rev. Biophys.* **2010**, *39*, 23–42.

- (28) Stefl, S.; Nishi, H.; Petukh, M.; Panchenko, A. R.; Alexov, E. Molecular Mechanisms of Disease-Causing Missense Mutations. *J. Mol. Biol.* **2013**, *425*, 3919–3936. <https://doi.org/10.1016/j.jmb.2013.07.014>.
- (29) Linglart, A.; Menguy, C.; Couvineau, A.; Auzan, C.; Gunes, Y.; Cancel, M.; Motte, E.; Pinto, G.; Chanson, P.; Bougnères, P.; et al. Recurrent PRKAR1A Mutation in Acrodysostosis with Hormone Resistance. *N. Engl. J. Med.* **2011**, *364*, 2218–2226.
- (30) Salpea, P.; Stratakis, C. A. Carney Complex and McCune Albright Syndrome: An Overview of Clinical Manifestations and Human Molecular Genetics. *Mol. Cell. Endocrinol.* **2014**, *386*, 85–91.
- (31) Stehr, H.; Jang, S.-H. J.; Duarte, J. M.; Wierling, C.; Lehrach, H.; Lappe, M.; Lange, B. M. H. The Structural Impact of Cancer-Associated Missense Mutations in Oncogenes and Tumor Suppressors. *Mol. Cancer* **2011**, *10* (54). <https://doi.org/10.1186/1476-4598-10-54>.
- (32) Morais, V. A.; Verstreken, P.; Roethig, A.; Smet, J.; Snellinx, A.; Vanbrabant, M.; Haddad, D.; Frezza, C.; Mandemakers, W.; Vogt-Weisenhorn, D.; et al. Parkinson's Disease Mutations in PINK1 Result in Decreased Complex I Activity and Deficient Synaptic Function. *EMBO Mol. Med.* **2009**, *1*, 99–111. <https://doi.org/10.1002/emmm.200900006>.
- (33) Grant, M. A.; Lazo, N. D.; Lomakin, A.; Condrón, M. M.; Arai, H.; Yamin, G.; Rigby, A. C.; Teplow, D. B. Familial Alzheimer's Disease Mutations Alter the Stability of the Amyloid β -Protein Monomer Folding Nucleus. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 16522–16527. <https://doi.org/10.1073/pnas.0705197104>.
- (34) Zwerger, M.; Jaalouk, D. E.; Lombardi, M. L.; Isermann, P.; Mauermann, M.; Dialynas, G.; Herrmann, H.; Wallrath, L. L.; Lammerding, J. Myopathic Lamin Mutations Impair Nuclear Stability in Cells and Tissue and Disrupt Nucleo-Cytoskeletal Coupling. *Hum. Mol. Genet.* **2013**, *22*, 2335–2349. <https://doi.org/10.1093/hmg/ddt079>.
- (35) Rakoczy, E. P.; Kiel, C.; McKeone, R.; Stricher, F.; Serrano, L. Analysis of Disease-Linked Rhodopsin Mutations Based on Structure, Function, and Protein Stability Calculations. *J. Mol. Biol.* **2011**, *405*, 584–606. <https://doi.org/10.1016/j.jmb.2010.11.003>.
- (36) George Priya Doss, C.; Rajith, B. A New Insight into Structural and Functional Impact of Single-Nucleotide Polymorphisms in PTEN Gene. *Cell Biochem. Biophys.* **2013**, *66*, 249–263. <https://doi.org/10.1007/s12013-012-9472-9>.
- (37) Kumar, A.; Rajendran, V.; Sethumadhavan, R.; Purohit, R. Evidence of Colorectal Cancer-Associated Mutation in MCAK: A Computational Report. *Cell Biochem. Biophys.* **2013**, *67*, 837–851. <https://doi.org/10.1007/s12013-013-9572-1>.
- (38) Uversky, V. N.; Oldfield, C. J.; Midic, U.; Xie, H.; Xue, B.; Vucetic, S.; Iakoucheva, L. M.; Obradovic, Z.; Keith, A. K. Unfoldomics of Human Diseases: Linking Protein Intrinsic Disorder with Diseases. *BMC Genomics* **2009**, *10*, S7. <https://doi.org/10.1186/1471-2164->

10-S1-S7.

- (39) Rehman, A. U.; Rafiq, H.; Rahman, M. U.; Li, J.; Liu, H.; Luo, S.; Arshad, T.; Wadood, A.; Chen, H. F. Gain-of-Function SHP2 E76Q Mutant Rescuing Autoinhibition Mechanism Associated with Juvenile Myelomonocytic Leukemia. *J. Chem. Inf. Model.* **2019**, *59*, 3229–3239. <https://doi.org/10.1021/acs.jcim.9b00353>.
- (40) Kass, I.; Knaupp, A. S.; Bottomley, S. P.; Buckle, A. M. Conformational Properties of the Disease-Causing Z Variant of A1-Antitrypsin Revealed by Theory and Experiment. *Biophys. J.* **2012**, *102*, 2856–2865. <https://doi.org/10.1016/j.bpj.2012.05.023>.
- (41) Wang, S.; Zhao, W. J.; Liu, H.; Gong, H.; Yan, Y. Bin. Increasing BB1-Crystallin Sensitivity to Proteolysis Caused by the Congenital Cataract-Microcornea Syndrome Mutation S129R. *Biochim. Biophys. Acta* **2013**, *1832*, 302–311. <https://doi.org/10.1016/j.bbadis.2012.11.005>.
- (42) Attanasio, F.; Convertino, M.; Magno, A.; Caflisch, A.; Corazza, A.; Haridas, H.; Esposito, G.; Cataldo, S.; Pignataro, B.; Milardi, D.; et al. Carnosine Inhibits A β 42 Aggregation by Perturbing the H-Bond Network in and around the Central Hydrophobic Cluster. *ChemBioChem* **2013**, *14*, 583–592. <https://doi.org/10.1002/cbic.201200704>.
- (43) Eldar, A.; Rozenberg, H.; Diskin-posner, Y.; Rohs, R.; Shakked, Z. Structural Studies of P53 Inactivation by DNA-Contact Mutations and Its Rescue by Suppressor Mutations via Alternative Protein – DNA Interactions. *Nucleic Acids Res.* **2013**, *41*, 8748–8759. <https://doi.org/10.1093/nar/gkt630>.
- (44) Karplus, M.; Kuriyan, J. Molecular Dynamics and Protein Function. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6679–6685. <https://doi.org/10.1073/pnas.0408930102>.
- (45) Mccammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585–590.
- (46) Schiffer, J. M.; Luo, M.; Dommer, A. C.; Thoron, G.; Pendergraft, M.; Santander, M. V.; Lucero, D.; Barros, E. P.; Prather, K. A.; Grassian, V. H.; et al. Impacts of Lipase Enzyme on the Surface Properties of Marine Aerosols. *J. Phys. Chem. Lett.* **2018**, *9*, 3839–3849. <https://doi.org/10.1021/acs.jpcclett.8b01363>.
- (47) Luo, M.; Dommer, A. C.; Schiffer, J. M.; Rez, D. J.; Mitchell, A. R.; Amaro, R. E.; Grassian, V. H. Surfactant Charge Modulates Structure and Stability of Lipase-Embedded Monolayers at Marine-Relevant Aerosol Surfaces. *Langmuir* **2019**, *35*, 9050–9060. <https://doi.org/10.1021/acs.langmuir.9b00689>.
- (48) Durrant, J. D.; Kochanek, S. E.; Casalino, L.; Jeong, P. U.; Dommer, A. C.; Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Cent. Sci.* **2020**, *6*, 189–196. <https://doi.org/10.1021/acscentsci.9b01071>.
- (49) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143. <https://doi.org/10.1016/j.neuron.2018.08.011>.

- (50) Leach, A. R. *Molecular Modelling - Principles and Applications*, Second edi.; Peason Education: Harlow, England, 2001.
- (51) Karplus, M.; Mccammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9* (9), 646–652.
- (52) Martínez, L.; Borin, I. A.; Skaf, M. S. Fundamentos de Simulação Por Dinâmica Molecular. In *Métodos de química teórica e modelagem molecular*; Morgon, N. H., Coutinho, K., Eds.; Editora Livraria da Física: São Paulo, 2007.
- (53) Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (54) Mackerell, A.D.; Bashford, D.; Dunbrack, R.L.; Evanseck, J.D.; Fiel, M.J.; Fischer, S.; Gao, J.; Guo, H.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.T.K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.T.; Prodhom, B.; Teiher, W.E.; Roux, B.; Sch, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J Phys Chem B* **1998**, *102*, 3586–3616. <https://doi.org/10.1021/jp973084f>.
- (55) Jorgesen, W. L.; Mazwell, D. S.; Tirado-River, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Propoerties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (56) Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
- (57) Hermans, J.; Berendsen, H. J. C.; Van Gunsteren, W. F.; Postina, J. P. M. A Consistent Empirical Potential for Water-Protein Interactions. *Biopolymers* **1984**, *23*, 1513–1518.
- (58) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (2), 198–210. <https://doi.org/10.1002/wcms.1121>.
- (59) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9*, 71. <https://doi.org/10.1080/17460441.2018.1403419>.
- (60) Lennard-Jones, J. E. On the Determination of Molecular Fields - II. From the Equation of State of a Gas. *Proc. R. Soc. Lond. A* **1924**, *106*, 463–477.
- (61) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Adv. Prot. Chem.* **2003**, *66*, 27–85. [https://doi.org/10.1016/S0065-3233\(03\)66002-X](https://doi.org/10.1016/S0065-3233(03)66002-X).
- (62) Martín-García, F.; Papaleo, E.; Gomez-Puertas, P.; Boomsma, W.; Lindorff-Larsen, K. Comparing Molecular Dynamics Force Fields in the Essential Subspace. *PLoS One* **2015**, *10*, e0121114. <https://doi.org/10.1371/journal.pone.0121114>.

- (63) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* **2012**, *7*, e32131. <https://doi.org/10.1371/journal.pone.0032131>.
- (64) Vitalini, F.; Mey, A. S. J. S.; Noé, F.; Keller, B. G. Dynamic Properties of Force Fields. *J. Chem. Phys.* **2015**, *142*, 084101. <https://doi.org/10.1063/1.4909549>.
- (65) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034–2045. <https://doi.org/10.1021/ct700127w>.
- (66) Baker, C. M. Polarizable Force Fields for Molecular Dynamics Simulations of Biomolecules. *WIREs Comput. Mol. Sci.* **2015**, *5*, 241–254. <https://doi.org/10.1002/wcms.1215>.
- (67) Wang, L. P.; Chen, J.; Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *J. Chem. Theory Comput.* **2013**, *9*, 452–460. <https://doi.org/10.1021/ct300826t>.
- (68) Babin, V.; Leforestier, C.; Paesani, F. Development of a “First Principles” Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient. *J. Chem. Soc., Faraday Trans.* **2013**, *9* (12), 5395–5403. <https://doi.org/10.1021/ct400863t>.
- (69) Cisneros, G. A.; Wikfeldt, K. T.; Ojamäe, L.; Lu, J.; Xu, Y.; Torabifard, H.; Bartók, A. P.; Csányi, G.; Molinero, V.; Paesani, F. Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions. *Chem. Rev.* **2016**, *116*, 7501–7528. <https://doi.org/10.1021/acs.chemrev.5b00644>.
- (70) Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.* **2004**, *47*, 1879–1881. <https://doi.org/10.1021/jm0341913>.
- (71) Wassman, C. D.; Baronio, R.; Demir, Ö.; Wallentine, B. D.; Chen, C.-K.; Hall, L. V; Salehi, F.; Lin, D.; Chung, B. P.; Hatfield, G. W.; et al. Computational Identification of a Transiently Open L1/S3 Pocket for Reactivation of Mutant P53. *Nat. Commun.* **2013**, *4*, 1407. <https://doi.org/10.1038/ncomms2361>.
- (72) Amaro, R. E.; Schnaufer, A.; Interthal, H.; Hold, W.; Stuart, K. D.; McCammon, J. A. Discovery of Drug-like Inhibitors of an Essential RNA-Editing Ligase in *Trypanosoma Brucei*. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17278–17283. <https://doi.org/10.1073/pnas.0805820105>.
- (73) Wang, Y.; Hess, T. N.; Jones, V.; Zhou, J. Z.; McNeil, M. R.; Andrew McCammon, J. Novel Inhibitors of Mycobacterium Tuberculosis DTDP-6-Deoxy-l-Lyx-4- Hexulose Reductase (RmlD) Identified by Virtual Screening. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 7064–7067. <https://doi.org/10.1016/j.bmcl.2011.09.094>.

- (74) Grant, B. J.; Lukman, S.; Hocker, H. J.; Sayyah, J.; Brown, J. H.; Mccammon, J. A.; Gorfe, A. A. Novel Allosteric Sites on Ras for Lead Generation. **2011**, *6* (10), 1–10. <https://doi.org/10.1371/journal.pone.0025711>.
- (75) Savarino, A. A Historical Sketch of the Discovery and Development of HIV-1 Integrase Inhibitors. *Expert Opin. Investig. Drugs* **2006**, *15*, 1507–1522. <https://doi.org/10.1517/13543784.15.12.1507>.
- (76) Schiffer, J. M.; Mael, L. E.; Prather, K. A.; Amaro, R. E.; Grassian, V. H. Sea Spray Aerosol: Where Marine Biology Meets Atmospheric Chemistry. *ACS Cent. Sci.* **2018**, *4* (12), 1617–1623. <https://doi.org/10.1021/acscentsci.8b00674>.
- (77) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.
- (78) Pierce, L. C. T.; Salomon-Ferrer, R.; Augusto F. De Oliveira, C.; McCammon, J. A.; Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 2997–3002. <https://doi.org/10.1021/ct300284c>.
- (79) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.* **2015**, *11*, 3584–3595. <https://doi.org/10.1021/acs.jctc.5b00436>.
- (80) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151. [https://doi.org/10.1016/S0009-2614\(99\)01123-9](https://doi.org/10.1016/S0009-2614(99)01123-9).
- (81) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (82) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824. <https://doi.org/10.1021/jp071097f>.
- (83) Beauchamp, K. A.; McGibbon, R.; Lin, Y. S.; Pande, V. S. Simple Few-State Models Reveal Hidden Complexity in Protein Folding. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17807–17813. <https://doi.org/10.1073/pnas.1201810109>.
- (84) De Sancho, D.; Mittal, J.; Best, R. B. Folding Kinetics and Unfolded State Dynamics of the GB1 Hairpin from Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 1743–1753. <https://doi.org/10.1021/ct301033r>.
- (85) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-Level Description of Ubiquitin Folding. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5915–5920. <https://doi.org/10.1073/pnas.1218321110>.
- (86) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov State

- Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413–18419. <https://doi.org/10.1021/ja207470h>.
- (87) Malmstrom, R. D.; Kornev, A. P.; Taylor, S. S.; Amaro, R. E. Allostery through the Computational Microscope: CAMP Activation of a Canonical Signalling Domain. *Nat. Commun.* **2015**, *6* (May), 7588. <https://doi.org/10.1038/ncomms8588>.
- (88) Taylor, B. C.; Lee, C. T.; Amaro, R. E. Structural Basis for Ligand Modulation of the CCR2 Conformational Landscape. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (17), 8131–8136. <https://doi.org/10.1073/pnas.1814131116>.
- (89) Bowman, G. R.; Geissler, P. L. Equilibrium Fluctuations of a Single Folded Protein Reveal a Multitude of Potential Cryptic Allosteric Sites. **2012**, *109* (29), 11681–11686. <https://doi.org/10.1073/pnas.1209309109>.
- (90) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-Based Simulations on Google Exacycle Reveal Ligand Modulation of GPCR Activation Pathways. *Nat. Chem.* **2014**, *6*, 15–21. <https://doi.org/10.1038/nchem.1821>.
- (91) Zimmerman, M. I.; Hart, K. M.; Sibbald, C. A.; Frederick, T. E.; Jimah, J. R.; Knoverek, C. R.; Tolia, N. H.; Bowman, G. R. Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS Cent. Sci.* **2017**, *3*, 1311–1321. <https://doi.org/10.1021/acscentsci.7b00465>.
- (92) Zimmerman, M. I.; Bowman, G. R. FAST Conformational Searches by Balancing Exploration/ Exploitation Trade-Offs. *J. Chem. Theory Comput.* **2015**, *11*, 5747–5757. <https://doi.org/10.1021/acs.jctc.5b00737>.
- (93) Prinz, J. H.; Keller, B.; Noé, F. Probing Molecular Kinetics with Markov Models: Metastable States, Transition Pathways and Spectroscopic Observables. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912–16927. <https://doi.org/10.1039/c1cp21258c>.
- (94) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48* (2), 414–422. <https://doi.org/10.1021/ar5002999>.
- (95) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126*, 155101. <https://doi.org/10.1063/1.2714538>.
- (96) Noé, F. Probability Distributions of Molecular Observables Computed from Markov Models. *J. Chem. Phys.* **2008**, *128*, 244103. <https://doi.org/10.1063/1.2916718>.
- (97) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noe, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**,

- 11, 5525–5542. <https://doi.org/10.1021/acs.jctc.5b00743>.
- (98) Bowman, G. R.; Huang, X.; Pande, V. S. Using Generalized Ensemble Simulations and Markov State Models to Identify Conformational States. *Methods* **2009**, *49* (2), 197–201. <https://doi.org/10.1016/j.ymeth.2009.04.013>.
- (99) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419. <https://doi.org/10.1021/ct200463m>.
- (100) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99–105. <https://doi.org/10.1016/j.ymeth.2010.06.002>.
- (101) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105. <https://doi.org/10.1063/1.3565032>.
- (102) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396. <https://doi.org/10.1021/jacs.7b12191>.
- (103) Bowman, G. R.; Pande, V. S. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*.
- (104) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9* (4), 2000–2009. <https://doi.org/10.1021/ct300878a>.
- (105) Chong, S. H.; Ham, S. Examining a Thermodynamic Order Parameter of Protein Folding. *Sci. Rep.* **2018**, *8*, 7148. <https://doi.org/10.1038/s41598-018-25406-8>.
- (106) Bernetti, M.; Masetti, M.; Recanatini, M.; Amaro, R. E.; Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach to Characterize Drug Binding Processes. *J. Chem. Theory Comput.* **2019**, *15*, 5689–5702. <https://doi.org/10.1021/acs.jctc.9b00450>.
- (107) Wu, H.; Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *J. Nonlinear Sci.* **2020**, *30*, 23–66. <https://doi.org/10.1007/s00332-019-09567-y>.
- (108) Hyvärinen, A.; Oja, E. Independent Component Analysis: Algorithms and Applications. *Neural Networks* **2000**, *13*, 411–430. [https://doi.org/10.1016/S0893-6080\(00\)00026-5](https://doi.org/10.1016/S0893-6080(00)00026-5).
- (109) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412–425. <https://doi.org/10.1002/prot.340170408>.
- (110) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102. <https://doi.org/10.1063/1.4811489>.

- (111) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179. <https://doi.org/10.1007/s11634-013-0134-6>.
- (112) Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77*, 257–286. <https://doi.org/10.1109/5.18626>.
- (113) Noé, F.; Wu, H.; Prinz, J. H.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139*, 184114. <https://doi.org/10.1063/1.4828816>.
- (114) Malmstrom, R. D.; Lee, C. T.; Wart, A. T. Van; Amaro, R. E. Application of Molecular-Dynamics Based Markov State Models to Functional Proteins. **2014**.
- (115) E, W.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503–523. <https://doi.org/10.1007/s10955-005-9003-9>.
- (116) Metzner, P.; Schutte, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7*, 1192–1219.
- (117) Polizzi, N. F.; Therien, M. J.; Beratan, D. N. Mean First-Passage Times in Biology. *Isr. J. Chem.* **2016**, *56*, 816–824. <https://doi.org/10.1002/ijch.201600040>.
- (118) Sumner, T. Dazzling History. *Science (80-.)*. **2014**, *343*, 1092–1093. <https://doi.org/10.1126/science.343.6175.1092>.
- (119) Meisburger, S. P.; Thomas, W. C.; Watkins, M. B.; Ando, N. X-Ray Scattering Studies of Protein Structural Dynamics. *Chem. Rev.* **2017**, *117*, 7615–7672. <https://doi.org/10.1021/acs.chemrev.6b00790>.
- (120) Wall, M. E.; Wolff, A. M.; Fraser, J. S. Bringing Diffuse X-Ray Scattering into Focus. *Curr. Opin. Struct. Biol.* **2018**, *50*, 109–116. <https://doi.org/10.1016/j.sbi.2018.01.009>.
- (121) Wall, M. E.; Adams, P. D.; Fraser, J. S.; Sauter, N. K. Diffuse X-Ray Scattering to Model Protein Motions. *Structure* **2014**, *22*, 182–184.
- (122) Meisburger, S. P.; Case, D. A.; Ando, N. Diffuse X-Ray Scattering from Correlated Motions in a Protein Crystal. *bioRxiv* **2019**. <https://doi.org/10.1101/805424>.
- (123) Riccardi, D.; Cui, Q.; Phillips Jr., G. N. Evaluating Elastic Network Models of Crystalline Biological Molecules with Temperature Factors, Correlated Motions, and Diffuse X-Ray Scattering. *Biophys. J.* **2010**, *99* (8), 2616–2625. <https://doi.org/10.1016/j.bpj.2010.08.013>.
- (124) Polikanov, Y. S.; Moore, P. B. Acoustic Vibrations Contribute to the Diffuse Scatter Produced by Ribosome Crystals. *Acta Crystallogr., Sect. D Biol. Crystallogr* **2015**, *71*, 2021–2031. <https://doi.org/10.1107/S1399004715013838>.
- (125) Van Benschoten, A. H.; Liu, L.; Gonzalez, A.; Brewster, A. S.; Sauter, N. K.; Fraser, J. S.;

- Wall, M. E. Measuring and Modeling Diffuse Scattering in Protein X-Ray Crystallography. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (15), 4069–4074. <https://doi.org/10.1073/pnas.1524048113>.
- (126) Cleary, M. Molecular Dynamics Studied by Analysis of the X-Ray Diffuse Scattering from Lysozyme Crystals. *Doucet, J. Benoit, J. P.* **1987**, *325*, 643–646. <https://doi.org/10.1017/CBO9781107415324.004>.
- (127) Clarage, J. B.; Clarage, M. S.; Phillips, W. C.; Sweet, R. M.; Caspar, D. L. D. Correlations of Atomic Movements in Lysozyme Crystals. *Proteins* **1992**, *12*, 145–157. <https://doi.org/10.1002/prot.340120208>.
- (128) Moore, P. B. On the Relationship between Diffraction Patterns and Motions in Macromolecular Crystals. *Structure* **2009**, *17*, 1307–1315. <https://doi.org/10.1016/j.str.2009.08.015>.
- (129) Peck, A.; Poitevin, F.; Lane, T. J. Intermolecular Correlations Are Necessary to Explain Diffuse Scattering from Protein Crystals. *IUCrJ* **2018**, *5*, 211–222. <https://doi.org/10.1107/S2052252518001124>.
- (130) Meinhold, L.; Smith, J. C. Fluctuations and Correlations in Crystalline Protein Dynamics: A Simulation Analysis of Staphylococcal Nuclease. *Biophys. J.* **2005**, *88* (4), 2554–2563. <https://doi.org/10.1529/biophysj.104.056101>.
- (131) Héry, S.; Genest, D.; Smith, J. C. X-Ray Diffuse Scattering and Rigid-Body Motion in Crystalline Lysozyme Probed by Molecular Dynamics Simulation. *J. Mol. Biol.* **1998**, *279*, 303–319. <https://doi.org/10.1006/jmbi.1998.1754>.
- (132) Meinhold, L.; Smith, J. C. Protein Dynamics from X-Ray Crystallography: Anisotropic, Global Motion in Diffuse Scattering Patterns. *Proteins* **2007**, *66*, 941–953. <https://doi.org/10.1002/prot>.
- (133) Wall, M. E.; Van Benschoten, A. H.; Sauter, N. K.; Adams, P. D.; Fraser, J. S.; Terwilliger, T. C. Conformational Dynamics of a Crystalline Protein from Microsecond-Scale Molecular Dynamics Simulations and Diffuse X-Ray Scattering. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (50), 17887–17892. <https://doi.org/10.1073/pnas.1416744111>.
- (134) Chan, E. J. On the Use of Molecular Dynamics Simulation to Calculate X-Ray Thermal Diffuse Scattering from Molecular Crystals. *J. Appl. Cryst.* **2015**, *48*, 1420–1428.
- (135) Wall, M. E. Internal Protein Motions in Molecular-Dynamics Simulations of Bragg and Diffuse X-Ray Scattering. *IUCrJ* **2018**, 172–181.
- (136) Wych, D. C.; Fraser, J. S.; Mobley, D. L.; Wall, M. E. Liquid-like and Rigid-Body Motions in Molecular-Dynamics Simulations of a Crystalline Protein. *Struct. Dyn* **2019**, *6*, 064704. <https://doi.org/10.1063/1.5132692>.
- (137) Cerutti, D. S.; Case, D. A. Molecular Dynamics Simulations of Macromolecular Crystals.

Wires Comput. Mol. Sci. **2019**, *9*, e1402. <https://doi.org/10.1002/wcms.1402>.

- (138) Guinier, A. *X-Ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies*; W. H. Freeman and Company: San Francisco, 19963.
- (139) Sjoberg, T. J.; Kornev, A. P.; Taylor, S. S. Dissecting the CAMP-Inducible Allosteric Switch in Protein Kinase A RIalpha. *Protein Sci.* **2010**, *19*, 1213–1221. <https://doi.org/10.1002/pro.400>.
- (140) Das, R., Esposito, V., Abu-Abed, M., Anand, G. S., Taylor, S. S., Melacini, G. CAMP Activation of PKA Defines an Ancient Signaling Mechanism. *PNAS* **2007**, *104* (1), 93–98. <https://doi.org/10.1073/pnas.0609033103>.
- (141) Bick, M. J.; Greisen, P. J.; Morey, K. J.; Antunes, M. S.; La, D.; Sankaran, B.; Reymond, L.; Johnsson, K.; Medford, J. I.; Baker, D. Computational Design of Environmental Sensors for the Potent Opioid Fentanyl. *Elife* **2017**, *6*, e28909.
- (142) Feng, J.; Jester, B. W.; Tinberg, C. E.; Mandell, D. J.; Antunes, M. S.; Chari, R.; Morey, K. J.; Rios, X.; Medford, J. I.; Church, G. M.; et al. A General Strategy to Construct Small Molecule Biosensors in Eukaryotes. *Elife* **2015**, *4* (2015), e10606. <https://doi.org/10.7554/eLife.10606>.
- (143) Vogelstein, B.; Lane, D.; Levine, A. J. Surfing the P53 Network. *Nature* **2000**, *408*, 307–310.
- (144) Vousden, K. H.; Lu, X. Live or Let Die: The Cell's Response to P53. *Nature* **2002**, *2*, 594–604. <https://doi.org/10.1038/nrc864>.
- (145) Biegging, K. T.; Attardi, L. D. Deconstructing P53 Transcriptional Networks in Tumor Suppression. *Trends Cell Biol.* **2012**, *22* (2), 97–106. <https://doi.org/10.1016/j.tcb.2011.10.006>.
- (146) Basse, N.; Kaar, J. L.; Settanni, G.; Joerger, A. C.; Rutherford, T. J.; Fersht, A. R. Toward the Rational Design of P53-Stabilizing Drugs: Probing the Surface of the Oncogenic Y220C Mutant. *Chem. & Biol.* **2010**, *17*, 46–56. <https://doi.org/10.1016/j.chembiol.2009.12.011>.
- (147) Freed-Pastor, W. A.; Prives, C. Mutant P53 : One Name , Many Proteins. *Genes Dev.* **2012**, *26*, 1268–1286. <https://doi.org/10.1101/gad.190678.112.1268>.

Electrostatic interactions as mediators in the allosteric activation of PKA RI α

Emília P. Barros^{‡,1}, Robert D. Malmstrom^{‡,1,2}, Kimya Nourbakhsh¹, Jason C. Del Rio³, Alexandr P. Kornev³, Susan S. Taylor^{1,3}, Rommie E. Amaro^{*1,2}.

[‡] These authors contributed equally to this work.

Author Affiliations:

1 – Department of Chemistry and Biochemistry; University of California, San Diego

2 – National Biomedical Computation Resource; University of California, San Diego

3 – Department of Pharmacology; University of California, San Diego

2.1 Abstract

Close-range electrostatic interactions that form salt bridges are key components of protein stability. Here we investigate the role of these charged interactions in modulating the allosteric activation of Protein Kinase A (PKA) via computational and experimental mutational studies of a conserved basic patch located in the regulatory subunit's B/C helix. Molecular dynamics simulations evidenced the presence of an extended network of fluctuating salt bridges spanning the helix and connecting the two cAMP binding domains in its extremities. Distinct changes in the flexibility and conformational free energy landscape induced by the separate mutations of Arg239 and Arg241 suggested alteration of cAMP-induced allosteric activation and were verified through *in vitro* fluorescent polarization assays. These observations suggest a mechanical aspect to the allosteric transition of PKA, with Arg239 and Arg241 acting in competition to promote the transition between the two protein functional states. The simulations also provide a molecular explanation to the essential role of Arg241 in allowing cooperative activation, by evidencing the existence of a stable interdomain salt bridge with Asp267. Our integrated approach points to the role of salt bridges not only in protein stability but also in promoting conformational transition and function.

2.2 Introduction

Protein structure is essential for protein function and is a result of interactions between neighboring as well as spatially distant residues relative to their primary sequence. Among the wide range of possible intra- and intermolecular interactions, salt bridges are defined by non-covalent charged interactions between acidic and basic residues¹ and are critical to the folding, stability and function of most proteins²⁻⁵. Salt bridges are also key interactions in areas such as

drug design and protein engineering⁶⁻⁸. In addition to playing important roles in structural stability, salt bridges can also mediate protein conformational change, allostery, and dynamics^{9,10}.

The majority of the studies of salt bridges have involved static representations of biomolecules using structures resolved by X-ray crystallography. Molecular dynamics (MD) simulations, however, provide a means of studying these proteins at a dynamic and molecular level. As a model system for the investigation of the role of salt bridges on protein structure and function using all-atom MD simulations, we turned to the flexible type I α regulatory subunit (RI α) of cAMP-dependent protein kinase A (PKA). The use of MD simulations coupled with experiments has recently successfully allowed the identification of allosteric networks on other protein kinases¹¹⁻¹³.

PKA is an ubiquitous protein kinase that is important in many key cellular signaling pathways¹⁴. In its basal state, PKA exists in an inactive holoenzyme conformation containing a regulatory (R) subunit dimer and two catalytic (C) subunits and is activated by the second-messenger cyclic adenosine monophosphate (cAMP)¹⁵. The R subunits bind to cAMP cooperatively and allosterically activate the holoenzyme by unleashing the catalytically active C subunits¹⁶. All four R subunit isoforms (RI α and β , RII α and β) share a conserved domain architecture, containing two tandem cyclic-nucleotide binding domains (CBD-A and CBD-B) connected via a long helical segment, known as the B/C helix¹⁷. The B/C helix incorporates portions of both CBDs, including the α B and α C helices from CBD-A and the α N-helix from the N3A-motif of CBD-B¹⁸. The binding of cAMP and release of C are associated with a dramatic conformational change in R, with residues in the CBDs moving up to 30 Å from their position in the holoenzyme conformation¹⁴. The two conformations are termed the “H-conformation” and “B-conformation”, with the former being the holoenzyme and the latter bound to cAMP.

The B/C helix plays a critical role in the regulation of PKA activity and is proposed to be essential for allosteric signal transduction^{19,20}. Several major structural changes occur in this helix upon protein activation. In the four R subunits isoforms, a conserved patch of four positively charged residues is located in the center of this important structural motif (Figure 2.1)^{17,21–23}, and mutational studies of Arg241 showed it to be important for cooperative activation²⁴. While differences in some PKA mutant's dynamics and function have been attributed to disruptions in individual salt bridges within R^{24,25}, no systematic study has investigated the role of salt bridge networks in the function and stability of the RI α subunit. We therefore sought to evaluate the role of the positively charged basic patch within the B/C helix in the allosteric activation of PKA, using alanine mutagenesis to analyze the dynamic formation (or disruption) of salt bridge networks through MD simulations and validated by experimental *in vitro* activation assays.

A

RI α	2 2 6	—	RDSYRRILMGSTL	RKRK	MYEEFLSK	—	2 5 0
RI β	2 2 6	—	RDSYRRILMGSTL	RKRK	MYEEFLSK	—	2 5 0
RII α	2 3 0	—	RVTFRRIIVKNN	KRKR	MFESFIES	—	2 5 4
RII β	2 4 7	—	RVTFRRIIVKNN	KRKR	MYESFIES	—	2 7 1

B

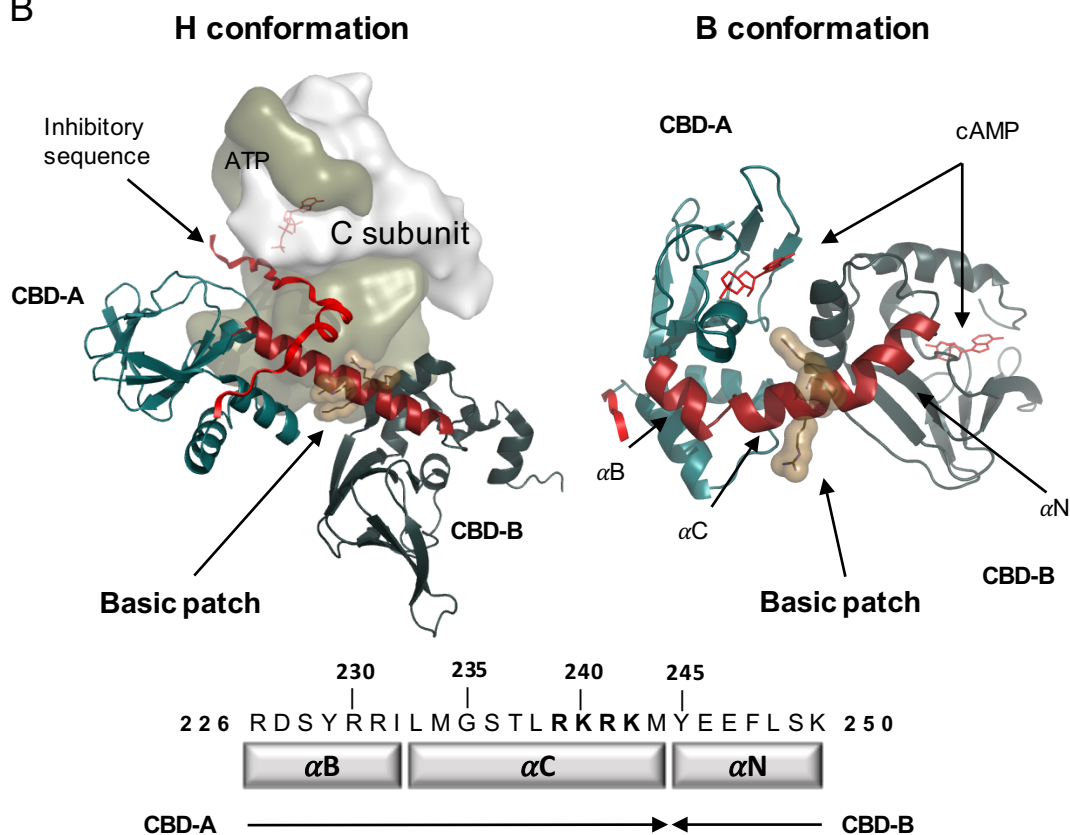


Figure 2.1. (a) Sequence alignment of the B/C helix of the 4 isoforms of the regulatory subunit. The positive patch in the B/C helix is colored red. (b) Representation of the regulatory subunit and B/C helix conformation at the two functional conformations of PKA. The side chains of the basic patch residues are colored ochre, the B/C helix is highlighted in red, the N lobe of the C subunit in white and the C lobe in tan.

With this aim, we have performed molecular dynamics simulations of wild type and four mutants of RI α (R239A, K240A, R241A and K242A) in the absence of the catalytic subunit and cAMP. The apo state was simulated to provide a molecular level view of the dynamics of RI α in

between its two structurally characterized states. We discuss the general features of the systems observed in the trajectories, including the flexibility of the B/C helix, and the differences in R conformational ensembles upon introduction of the mutations. In conjunction with *in vitro* fluorescence polarization assays, our results provide insights into the intrinsic flexibility of R and indicate that the basic patch in the B/C helix is important for stabilization of the H-conformation and in governing conformational dynamics and allosteric regulation. Our analysis also suggests the existence of an extended electrostatic network connecting the two cAMP binding domains, with the Arg241-Asp267 and Arg239-Glu143 salt bridges in particular playing key roles in the activation and stabilization of PKA. Finally, this work constitutes another example of the role of close-range electrostatic interactions in the stabilization and function of macromolecules.

2.3 Materials and Methods

System set up

The heavy atom coordinates for the five systems (wild type and mutants R239A, K240A, R241A and K242A) of PKA RI α were obtained from the crystallographic structure of the holoenzyme (PDB code 2QCS¹⁴), ranging from residue 113 to 379, which omits the flexible dimerization/docking domain (residues 11-61), the inhibitory site (residues 94-98) and linker regions¹⁵. Residues were protonated at pH 7.0 using Maestro-integrated PROPKA and mutations made using Schrödinger's Maestro (version 10.4, Schrödinger, LLC, New York, NY). The proteins were solvated in water boxes with counterions and 150 mM NaCl to simulate physiological conditions. The Amber14SB²⁶ force field was used for the protein and NaCl with TIP3P waters²⁷.

Molecular dynamics simulations

Simulations were performed using GPU accelerated Amber 14²⁶. The system was minimized in four stages: proton only, solvent, solvent and side chains, and the full system totaling 11,000 cycles using a combination of steepest decent and conjugate gradient methods. Equilibration involved a heating step to 100 K at constant volume for 50 ps followed by heating to 310 K at constant pressure, 1 bar, for 200 ps. The system was further equilibrated at 310 K and 1 bar NPT 750 ps. Molecular dynamics simulations were run as an ensemble with periodic boundary conditions at 1bar and 310 K. We used a non-bonded short range interactions cutoff of 10 Å, and the long-range electrostatic interactions were approximated by particle mesh Ewald²⁸. The simulations used a 2 fs time step with the SHAKE algorithm to constrain hydrogen atoms. Each protein system was simulated in 5 parallel runs of 1,000 ns each, each run being assigned new starting velocities, resulting in 5 μs of total sampling time for each system (Supplementary Table 2.S1).

Trajectory analysis

Trajectories were visualized using VMD²⁹. Structures obtained in the trajectory were aligned to the β-barrel of CBD-A (residues 152-225) and frames sampled for analyses every 100 ps. Secondary structure assignment and pairwise distance calculation for the salt bridge analysis were performed using functions from MDTraj³⁰, and all other analysis involved in-house programs. For the secondary structure analysis, bootstrapping sampling was done in order to obtain estimates of the variances of alpha helical proportions among the systems. A total of 20,000 bootstrapping samplings were performed, in each of which 1,000 frames of the simulation were randomly selected. The secondary structure averages of each bootstrapping sampling were then

used to create the histograms. The salt bridge distance cut-off was taken as nitrogen-oxygen distance of 4 Å.

Data Sharing

All MD input files, MD trajectories, and ipython notebooks used for the analyses presented in the paper are available for download at <http://doi.org/10.6075/J07D2S2X>

Purification of Regulatory Subunits and Generation of Mutants

The basic patch mutants (R239A, K240A, R241A, K242A) were generated using QuickChange site-directed mutagenesis. Wild type and mutant RI α proteins were purified as previously described^{16,31}. Proteins were expressed in *Eschericia coli* BL21 (DE3) from Novagen for 20-24 h at 15° C in TB medium. In brief, ammonium sulfate precipitation of the soluble lysate supernatants was batch bound overnight to a cAMP resin to purify via affinity chromatography. Proteins were eluted from the resin, using 40 mM cGMP, and applied to a Superdex 200 gel filtration column for final purification in gel filtration buffer [50 mM MES (pH 5.8), 200 mM NaCl, 2 mM EGTA, 2 mM EDTA, and 5 mM DTT].

Fluorescent Polarization Allosteric Activation Assay

Allosteric activation of Type I α PKA holoenzyme basic patch mutants was evaluated using a fluorescence polarization assay as previously described³¹. PKA holoenzymes were formed *in vitro* using a 1.2:1 (RI α :C) molar ratio in FP assay buffer [50 mM MOPS (pH 7.0), 35 mM NaCl, 10 mM MgCl₂, 0.005% (v/v) TritonX-100, 1 mM ATP, and 1 mM DTT]. The PKA inhibitory peptide conjugated to 5/6-carboxyfluoroscein (5/6-FAM-IP20) fluorophore was added to wells

containing the PKA holoenzyme at a final concentration of 2 nM. The C subunit concentration was kept constant at 12 nM and titrated with various concentrations of cAMP (0 nM to 8000 nM) to induce dissociation and allow binding of 5/6-FAM-IP20 to the C subunit. After 30 minutes of incubation to reach equilibrium, fluorescent polarization (excitation at 485 ± 20 nm, emission at 535 ± 25 nm) was measured using a GENios Pro microplate reader (TECAN) in black, flat bottom 96-well low-binding polypropylene assay plates (Greiner). Three independent experiments were performed, with each measurement being the mean of triplicate samples \pm the standard deviation. Graphs were generated and analyzed in Graphpad Prism 7.0a (La Jolla, CA), using a sigmoidal dose response curve of variable slope.

8-[Fluo]-cAMP Fluorescent Polarization Binding Assay

The ability of the R subunit basic patch mutants to bind cAMP in the absence of the C subunit was tested using the fluorescent cAMP analogue, 8-[fluo]-cAMP (Biolog), used at a final concentration of 10 nM. Wells were titrated with a range of R subunit concentrations (0 nM to 125 nM), which were diluted with FP assay buffer. Polarization readings were measured and analyzed as described above.

2.4 Results

Wild type conformational dynamics

The wild type and mutant systems were simulated in 5 parallel runs of 1 μ s each. Despite the time scales being shorter than those of the majority of biologically relevant processes, such as protein folding and domain mobility, we nonetheless observed great flexibility of the R subunit. To quantify this flexibility, we aligned all of the frames to the relatively rigid β -barrel of CBD-A

in the H conformation and measured the displacement of the center of mass of CBD-B relative to the principal moments of inertia of CBD-A's β -barrel using spherical coordinates. In this new reference system, the distance between the centers of mass of CBD-A and CBD-B is given by d , and the displacement in the x - y plane and relative to the z axis is given by the angles ϕ and θ , respectively (Figure 2.2a).

The histogram of distances between the CBD's centers of mass for the wild type system provides one measure of the flexibility of the system (Figure 2.2b). Most of the structures sampled adopted distances similar to that of the crystallographic H conformation, which was also the initial position of our simulations. Interestingly, we found a great proportion of configurations that extended beyond the known B and H conformations' CBDs distances. Larger distances were observed, as well as a small fraction of conformations with the centers of mass of the two lobes even closer together than in the globular, collapsed B conformation.

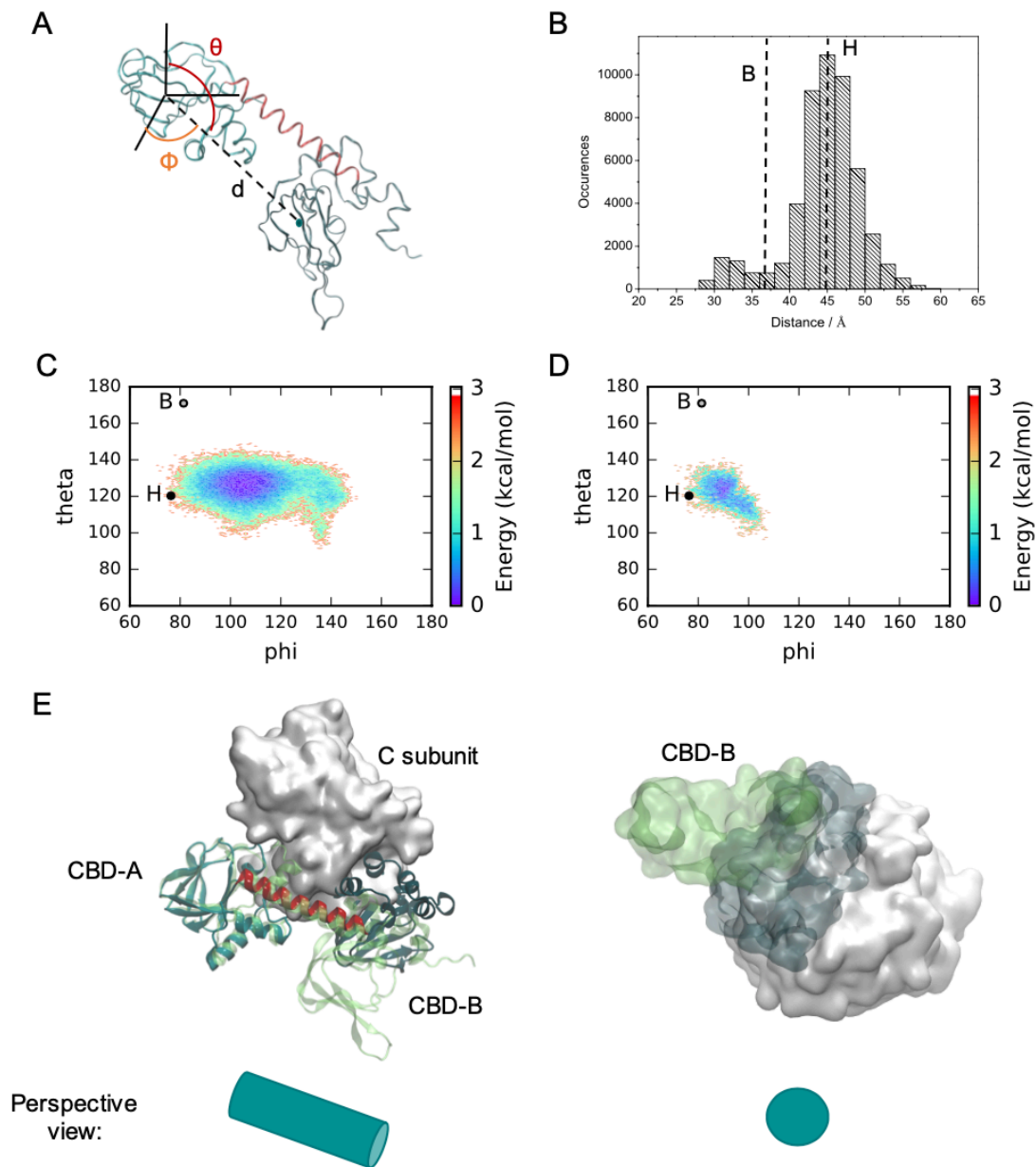


Figure 2.2. Dynamics of the wild type system. Spherical coordinate analysis of the conformational flexibility of wild type RI α . (a) Representation of the spherical coordinates of the center of mass of CBD-B β barrel relative to the principal moments of inertia of the H conformation CBD-A β barrel. (b) Histogram of the center of mass distances. (c) Free energy landscape in terms of spherical angles ϕ and θ for the complete set of sampled conformations. (d) Free energy landscape of the structures showing no overlap with the coordinates of the C subunit in the holoenzyme crystallographic structure. The spherical coordinates corresponding to the crystallographic structures are also shown. (e) Two views of the most probable conformation in the wildtype ensemble (CBD-A colored cyan, B/C helix red, and CBD-B dark green) compared to crystallographic H conformation (regulatory subunit in green, catalytic subunit in white). Structures were aligned in their CBD-A β barrel.

The spherical angles describe the orientation of the B domain and enrich the three-dimensional quantification of the RI α conformational ensemble. We performed Principal Component Analysis (PCA) on the coordinates sampled and found that the first two principal components are very similar to the chosen spherical angles (Supplementary Figure 2.S1). We thus extended our analysis using these spherical parameters because they provide a more intuitive representation of the ensemble, with the metrics directly indicating the relative position of CBD-A and CBD-B contrary to the more abstract representation given by principal components. With the most probable state being taken as a reference, the free energy landscape of the wild type ensemble in terms of these two angles is shown in Figure 2.2c, as well as the values corresponding to the H and B crystallographic conformations. On the timescale of the simulations, the transition between the H and B crystallographic conformations was not sampled, indicating that longer sampling would be required to observe the conformational change, as expected. However, this analysis shows that the most probable wildtype apo R conformation, located at the well in Figure 2.2c, does not correspond to the crystallographic H conformation. It differs from the H conformation mainly by a rotation of the B/C helix, which results in a roughly 45°-rotated B domain relative to its position in the X-ray crystal structure (Figure 2.2e). The rotation of CBD-B in the apo conformation suggests an overlap with the C subunit position (Figure 2.2d); thus, we would not expect the CBD-B to adopt this rotational position in the presence of the C subunit.

Mutant conformational ensembles

The spherical coordinate analysis was extended to the simulations of the four alanine mutations of the B/C helix basic patch. R239A, K240A and K242A displayed behavior similar to that of the wildtype in terms of the distance between the CBD's centers of mass, with the exception

of a slight increase in the number of observed conformations that have more proximal lobes, or smaller values of d (Figure 2.3). R241A, however, showed an opposite trend, with no sampling of these more collapsed conformations and instead a significant number of structures in which the centers of mass were further apart than in H.

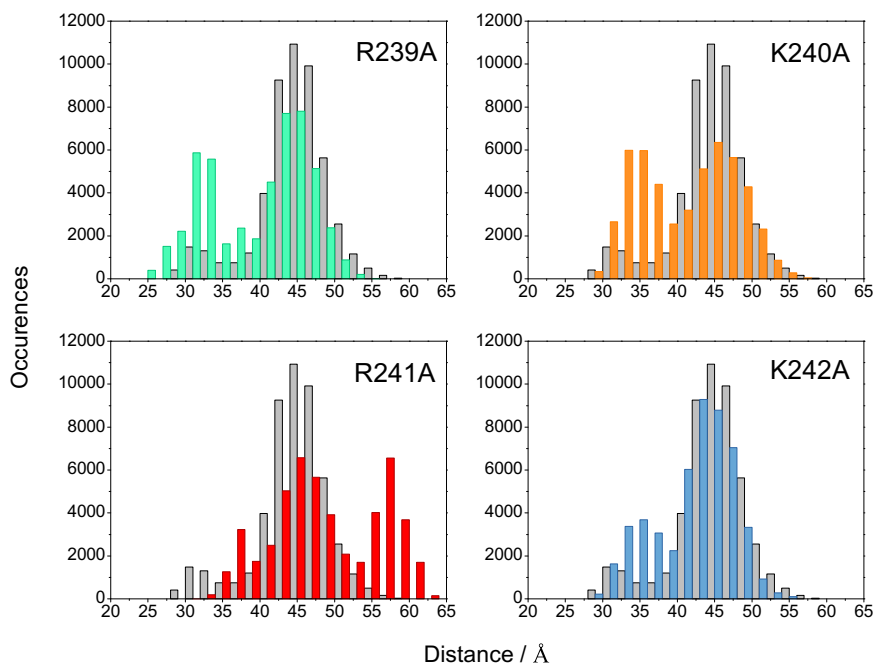


Figure 2.3. Histogram of CBD's centers of mass for the mutants (colors) versus wildtype (gray).

The free energy landscape in terms of the spherical angles for the mutants is shown in Figure 2.4. Comparison with that of wild type (Figure 2.2c) shows that the mutations of K240 and K242 did not significantly affect the sampled conformational ensemble of R, while the removal of the basic residues R239 and R241 resulted in a markedly different distribution of structures and the sampling of novel conformations (See Supplementary Figure 2.S2). These conformations displayed dramatic bends and deformations of the B/C helix (Figure 2.5 and Supplementary Figure 2.S3), and were more regularly sampled than in the wild type, K240A and K242A simulations.

These observations and the distribution of CBD distances suggest that, for R239A and R241A (especially the latter), there is uncoupling between the two domains upon the removal of the basic residue, resulting in much more freedom in conformational exploration. Nonetheless, all of these mutants still sample for a significant part of the simulation conformations similar to the wildtype's most probable conformation, as evidenced by the presence of the wells in the same values of ϕ and θ .

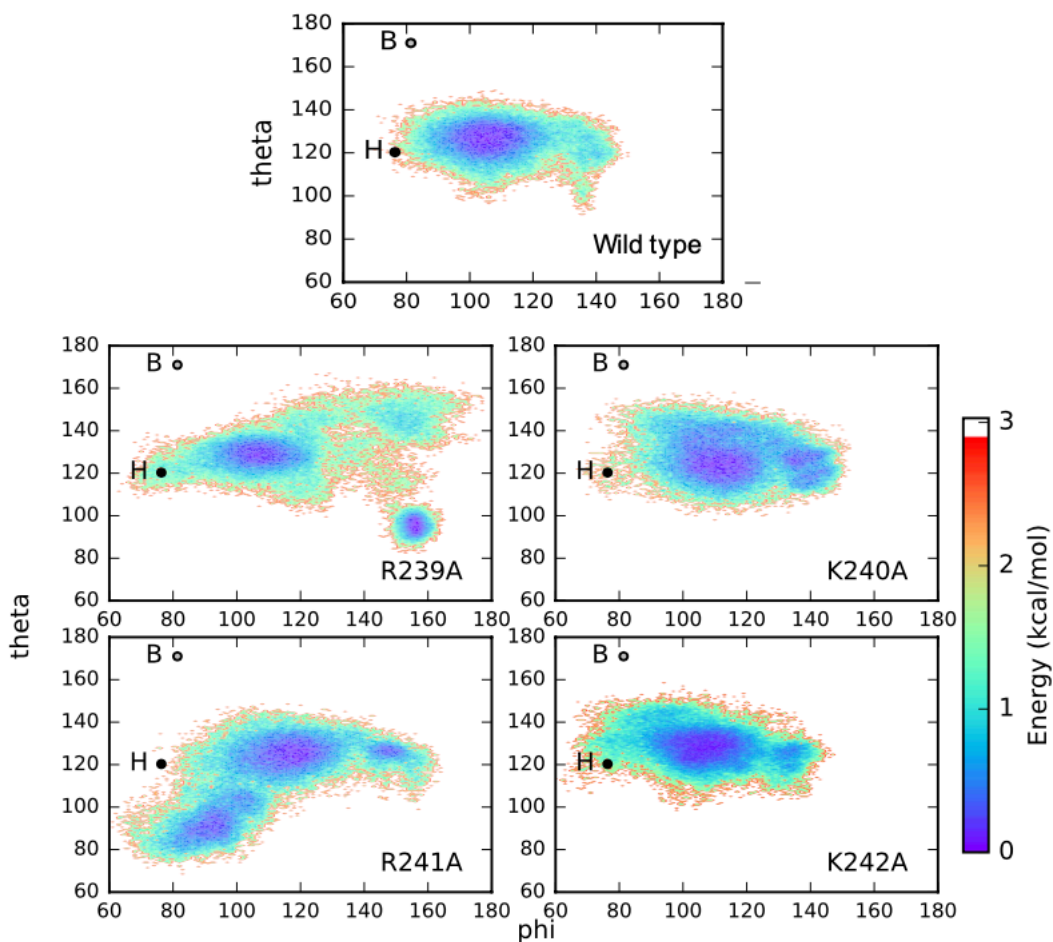


Figure 2.4. Free energy landscape in terms of spherical angles for the wild type and mutants. The coordinates for the crystallographic structures are also shown.

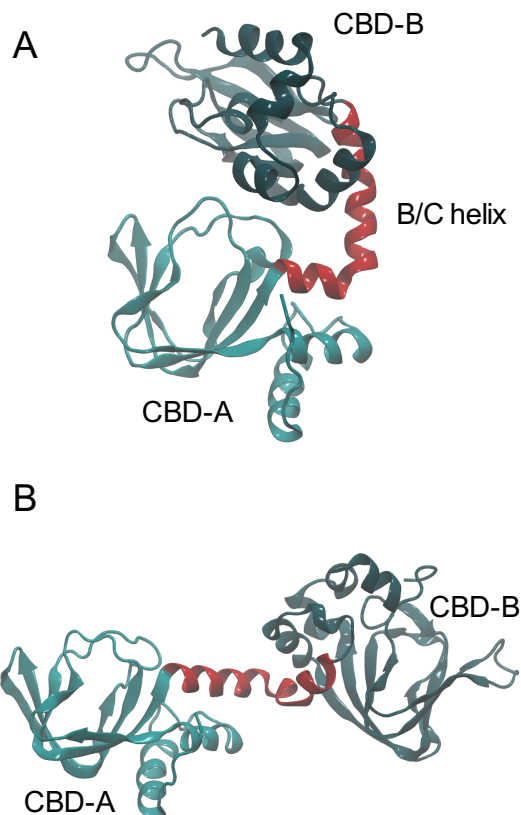


Figure 2.5. Novel conformations sampled in (a) R239A and (b) R241A simulations. CBD-A is colored cyan, CBD-B dark green and B/C helix highlighted red.

Flexibility of the B/C helix

Visual inspection suggested a high degree of flexibility in the B/C helix for all systems (Figure 2.5). Analysis of the spherical coordinates provided an indirect indication of this flexibility, because the displacement of the CBD-B relative to CBD-A's principal moments of inertia was mainly caused by movements in the B/C helix. To directly quantify the plasticity of the helix and identify regions with greater propensity for deformation, we calculated the fraction of residues in the B/C helix (residues 226-250) that displayed an α -helical secondary structure at each analyzed frame (Figure 2.6a). The wildtype, K240A and K242A showed similar helix

proportions, in agreement with the similarity in their conformational ensembles. The B/C helix in R239A and R241A was less well formed, having a lower helix proportion, as would be expected from the greater sampling of bent structures (Figure 2.5).

We further refined this analysis by calculating, for each residue in the helix, the fraction of frames in which it possesses an α -helical structure (Figure 2.6b). Most of the residues had very high helical proportion, with the notable exception of the C-terminal residues Ser249 and Lys250 (not shown in Figure 2.6b, with helical proportions ranging from 12.0 to 22.2 % and 3.8 to 11.0 %, respectively, see Supplementary Figure 2.S4). These residues, in all of the systems, were assigned in the majority of the frames as hydrogen-bonded turns³². In addition, R239A and R241A showed smaller helical proportions for other residues in the helix. Residues 226-236, located in the N-terminal part of the helix (in the CBD-A), were less ordered in R239A, with Leu233 being the most flexible of these. Oppositely, for R241A, it was the C-terminal part of the helix, comprising residues 233-248 in CBD-A and CBD-B, which displayed most pronounced flexibility. In this case, Leu238 had the greatest diversion from the wildtype and other mutants' helical proportion. Both Leu233 and Leu238 are found in the interface with the C subunit in the holoenzyme structure (Supplementary Figure 2.S5). Mutational studies of Leu233 have suggested that this residue is important for allowing the formation of the holoenzyme, with L233A displaying a 3-fold decrease in cAMP activation constant and a 3-fold increase in R-C dissociation constant¹⁹.

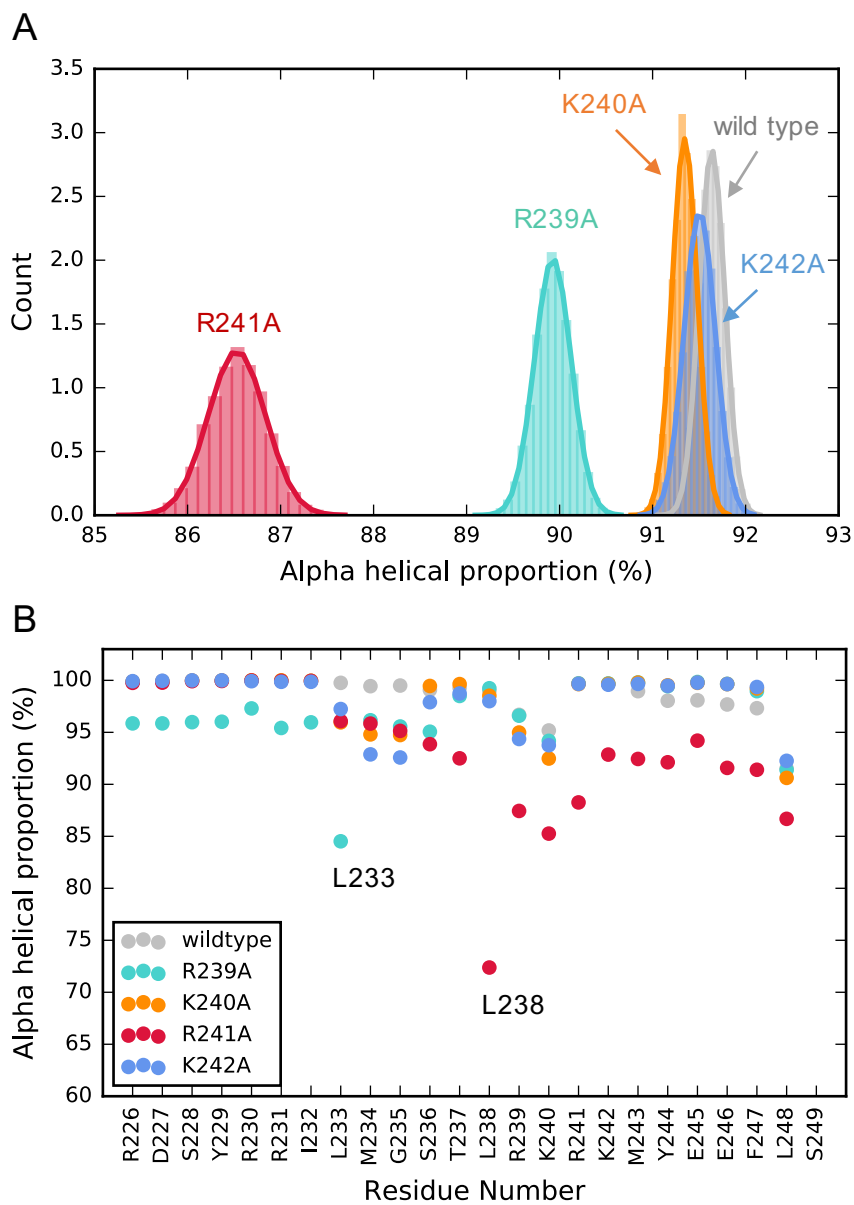


Figure 2.6. B/C helical proportion analysis of wild type and mutants' systems. (a) Full helix analysis and (b) per residue analysis for residues located in the B/C helix.

Allosteric Activation and cAMP Binding of B/C Helix Mutants

MD simulations suggested that mutations of the arginine residues of the basic patch induce perturbations in the dynamics and conformational ensemble of the R subunit. Given the dependence of enzyme function on protein structure and the overall shape of the free energy

landscape, the observations described above implicated altered function of R239A and R241A compared to wildtype RI α .

To validate our MD simulations of the basic patch mutants, we used two separate *in vitro* fluorescence polarization assays to address binding of cAMP to R subunits and allosteric activation of holoenzyme complexes (Supplementary Figure 2.S6). We initially speculated that because the mutations do not directly interact with cAMP in either the H or B conformations, cAMP binding would not be significantly impacted. As expected, using a fluorescent cAMP analogue, 8-[fluo]-cAMP, we found only minor differences in K_d values for R239A and K242A compared to wildtype RI α , with no change in cooperativity for any mutant (Table 2.1 and Supplementary Figure 2.S6a). We assessed the allosteric activation of PKA using an assay to measure dissociation of the C subunit from mutant holoenzyme complexes in response to increasing concentrations of cAMP, by measuring polarization of the fluorescent 5/6-FAM-IP20 peptide which binds to free C subunit in solution (Table 2.1 and Supplementary Figure 2.S6b). As anticipated, we found that R241A was >20-fold less sensitive to cAMP-stimulated activation and less cooperative than the wildtype, but R239A was slightly more sensitive to cAMP and exhibited greater cooperativity. Furthermore, K240A, but not K242A, showed a modest decrease in sensitivity to cAMP with no change in cooperativity.

Table 2.1. Nucleotide binding and allosteric activation of RI α B/C Helix basic patch mutants and wildtype.

System	cAMP binding		Activation of PKA	
	Kd (nM)	Hill Coefficient	EC50 (nM)	Hill Coefficient
Wildtype	7.30 \pm 0.11	1.65 \pm 0.04	23.36 \pm 0.66	2.11 \pm 0.11
R239A	6.72 \pm 0.11	1.71 \pm 0.04	17.64 \pm 0.67	2.54 \pm 0.22
K240A	7.32 \pm 0.12	1.83 \pm 0.05	29.98 \pm 1.23	2.15 \pm 0.18
R241A	7.29 \pm 0.10	1.80 \pm 0.04	543.07 \pm 27.40	1.44 \pm 0.09
K242A	7.71 \pm 0.11	1.80 \pm 0.04	24.93 \pm 0.85	1.86 \pm 0.10

2.5 Discussion

The use of all-atom MD simulations evidenced the pronounced flexibility of the apo regulatory subunit of PKA. Free wildtype PKA adopts a variety of conformations, with the extended, crystallographic H conformation only rarely sampled (Figure 2.2). Instead, a structure with a slight torsion on the B/C helix constitutes the most probable conformation, which disrupts the C subunit binding interface (Figure 2.2e). This finding suggests that the binding to C and inactivation of the enzyme require straining of the B/C helix and the R subunit undergoing the transition into a mechanically “frustrated” state. The protein frustratometer Web server (frustratometer.tk)³³ was used to verify this hypothesis, and the contact interactions in the B/C helix are predicted to be frustrated compared to the energetics of other residues in the same location (mutational frustration, Supplementary Figure 2.S7a) or the same interactions in other configurations (configurational frustration, Supplementary Figure 2.S7b)³⁴. This structural

frustration of R in the holoenzyme provides a molecular explanation for the quick activation of PKA upon cAMP binding and its ability to act as a dynamic allosteric switch. Despite favorable interactions in the R-C interface, the strain in the B/C helix favors the release of C and may be one of the factors that results in the shallow free energy landscape observed using long-timescale MD simulations and Markov State Models²⁰.

Mutation of Arg239 or Arg241 to alanine greatly perturbs the conformational ensemble of the regulatory subunit. R241A, in particular, differs from the other systems in that the CBD's dynamics seem to be decoupled, resulting in conformations in which they are separated by large distances. Similarly, the greatest variations in helix flexibility, as measured by the helix proportion, were seen in the R239A and R241A simulations. Our analysis further allowed the identification of the areas in the B/C helix that have a stronger propensity to be deformed and found that Leu233 in the case of R239A and Leu238 in R241A, both located at the interface with the catalytic subunit, are the most affected residues.

To relate the observed mutationally driven perturbation of the ensembles of R239A and R241A to the role of the basic residues and their involvement in electrostatic interactions, we calculated the total survival time of all of the salt bridges established within $RI\alpha$ from the simulations. This metric corresponds to the total fraction of frames in the simulation in which each salt bridge is formed. A qualitative representation of the network of salt bridges in the wildtype is given in Figure 2.7a. The great majority of the salt bridges are intra-domain, formed exclusively within CBD-A and within CBD-B. Interestingly, the only stable inter-domain salt bridge, formed for approximately 80% of the time in the wildtype simulation, is between Arg241 and Asp267. The same calculation was performed for the mutants, and the total survival time of salt bridges

that showed a change of more than 10% compared to wildtype are shown in Supplementary Figures 2.S8-11.

The survival times of the salt bridges in the mutant systems indicates that the disruption of a very reduced number of salt bridges by a single mutation affects a variety of others, emphasizing the fact that there is communication between the charged residues throughout RI α and that they are involved in an extended network. Moreover, the mutations in Arg239 and Arg241 involved deletion of stable, long-lived salt bridges, while Lys240A and Lys242A removed only transient, short-lived interactions. Using the identified salt bridges in the simulations, an electrostatic network can be established from the cAMP binding site in domain A, extending through the B/C helix and reaching the CBD-B (Figure 2.7b).

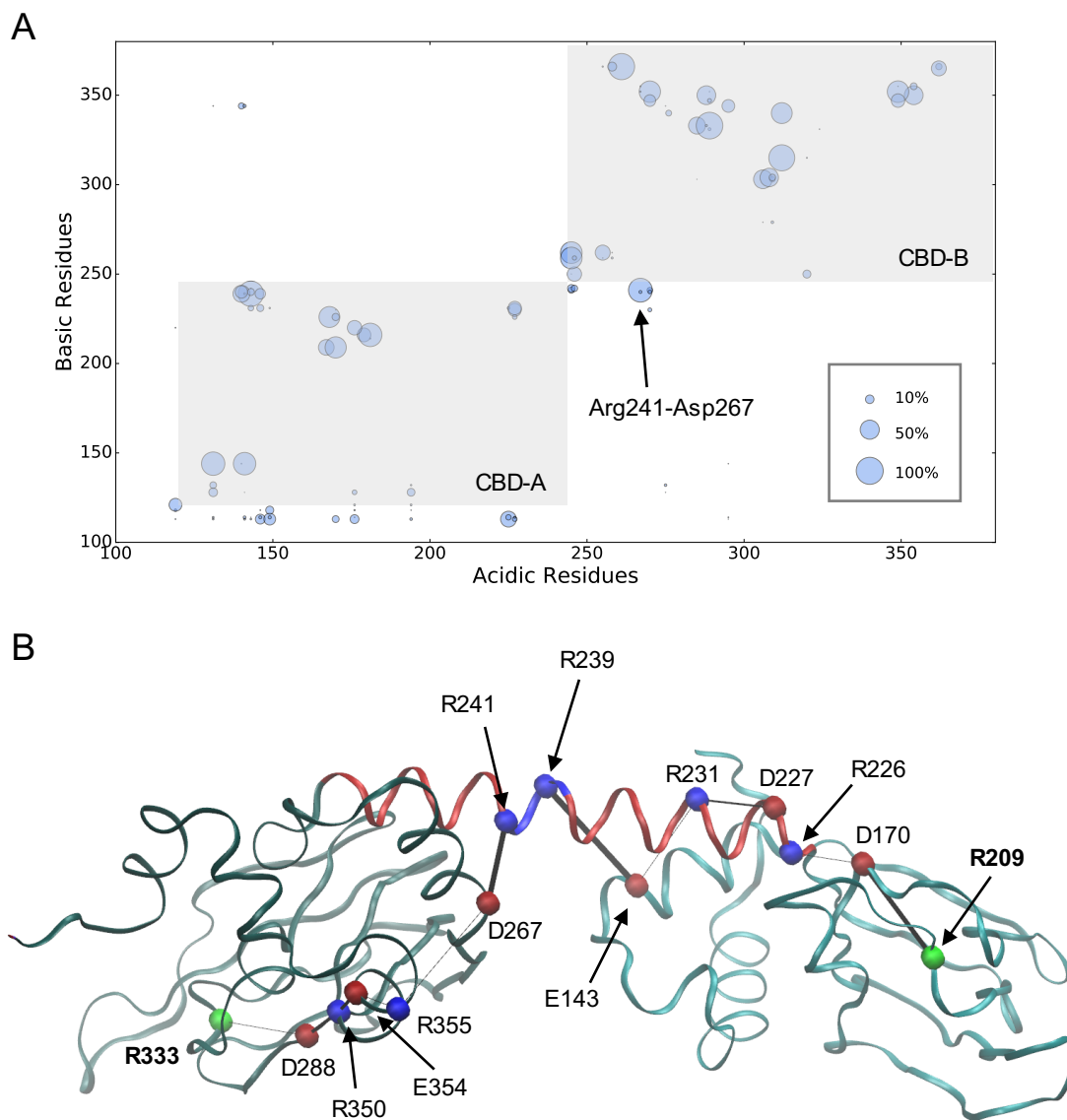


Figure 2.7. The network of salt bridges. (a) survival times of salt bridges formed in wildtype simulations. The size of the circles is proportional to the total survival time of the salt bridge. (b) Scheme of an electrostatic network connecting the two cAMP-binding domains. Basic residues are colored blue and acidic residues red, with the exception of residues in the cAMP binding sites, Arg209 and Arg233, which are colored green. The thickness of the black lines represents the lifetime of the salt bridge as measured in the wild type simulation.

Arg241-Asp267 functions as the main inter-domain salt bridge, allowing communication between the two binding sites. Crystal structures of the cAMP-bound R subunit and previous

structural models have suggested that PKA activation involved an interaction between Glu200 and Arg241, with Glu200 interacting with the 2'-hydroxyl group in cAMP²⁴. Our simulations of wild type and the mutants, however, did not sample the B conformation, with Arg241 and Glu200 not coming into close contact with each other. Our analysis therefore shows that there is no such salt bridge formed between these residues, indicating instead that the decoupling of the two domains happens due to the breakage of the Arg241-Asp267 salt bridge. The loss of allosteric activation seen experimentally is therefore a result of the removal of the intra-domain interaction, which breaks the electrostatic communication between the CBDs and disrupts the propagation of the allosteric signal.

The fluorescent polarization assays, on the other hand, indicate that Arg239 is involved in interactions of a different nature, contributing to the stability of the regulatory subunit in the extended, H-like conformation. The increased level of R-C dissociation may be caused by the destabilization of the binding interface, particularly Leu233, resulting in an effect similar to the L233A mutation¹⁹. In this way, the salt bridges involving Arg239 may function as “anchors” to keep the helix extended and allow the formation of the binding interface.

The coupling of computational and experimental analysis suggests that Arg239 and Arg241 play competing roles and that their modulation is an important factor for the regulation of PKA. We propose, in this way, that the extended salt bridge network is a key component of the allosteric mechanism and that there is a mechanical aspect to the conformational change caused by activation, with the salt bridges exerting a torque on the flexible B/C helix. More specifically, Arg239 and Arg241 seem to have essential roles in the stabilization of the H conformation and in the allosteric transduction upon activation, respectively.

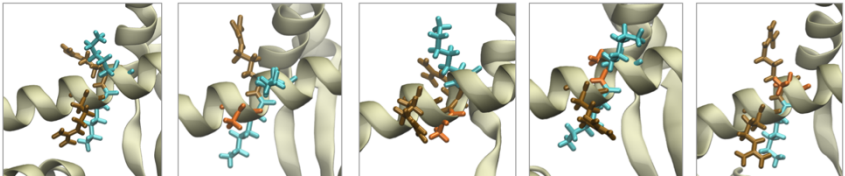
In conclusion, the investigation of the dynamics of ion pair interactions in PKA using MD simulations and experimental assays allowed the identification of several complex salt bridges and how they modulate the dynamics and function of PKA. Because there is compensation between the electrostatic interactions and variations in the pairs throughout the simulations due to side chain flexibility and greater-scale multi-domain motion, the use of computer simulations to investigate these interactions can greatly enrich the structural or ensemble-averaged observations achieved with other methods.

2.6 Acknowledgments

Chapter 2, in full, is a modified reprint of the material as it appears in “Barros, E. P., Malmstrom, R. D., Nourbakhsh, K., Del Rio, J. C., Kornev, A. P., Taylor, S. S., Amaro, R. E., Electrostatic interactions as mediators in the allosteric activation of protein kinase A RI α , *Biochemistry*, vol. 56, 2017. The dissertation author was the primary investigator and primary co-author of this paper.

2.7 Supporting Information

Table 2.S1. Details of the simulated systems. Arginines shown in cyan, lysines in ochre and mutated residues in orange



System	wildtype	R239A	K240A	R241A	K242A
# of atoms	82,561	82,569	82,568	82,581	82,571
# of independent runs	5	5	5	5	5
Simulation length (μ s)	1	1	1	1	1
Total sim. time (μ s)	5	5	5	5	5

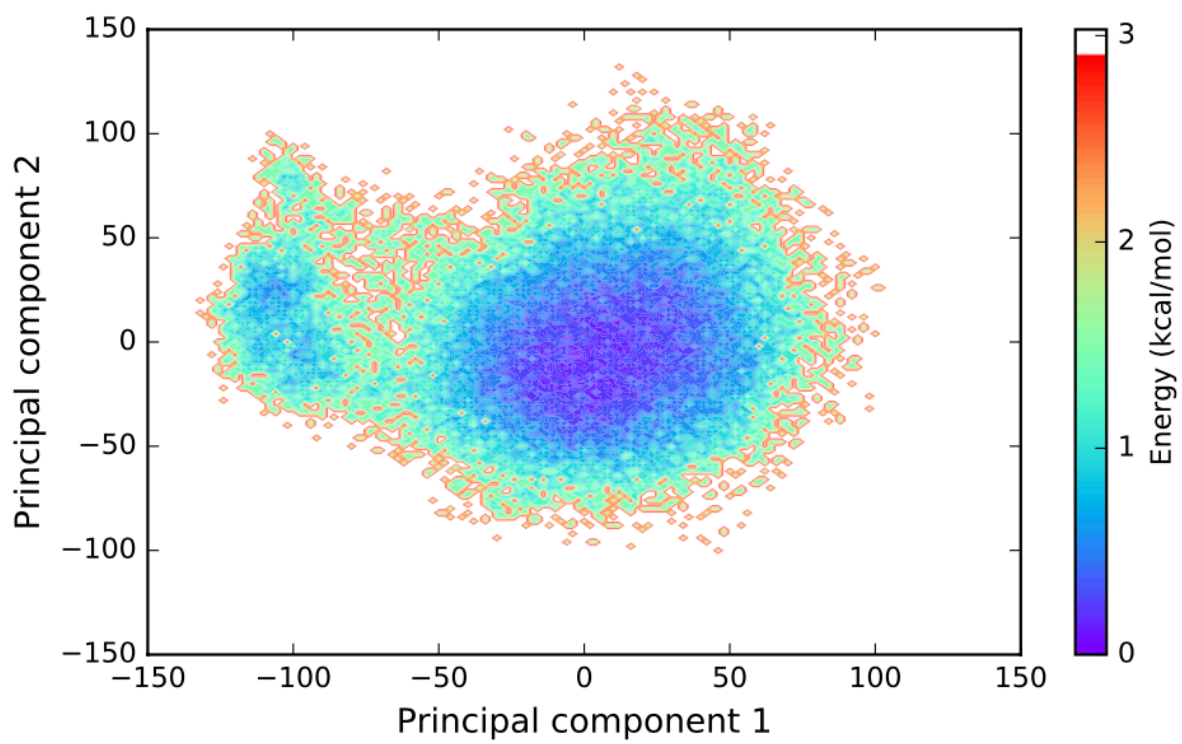


Figure 2.S1. Principal components analysis of wildtype trajectory

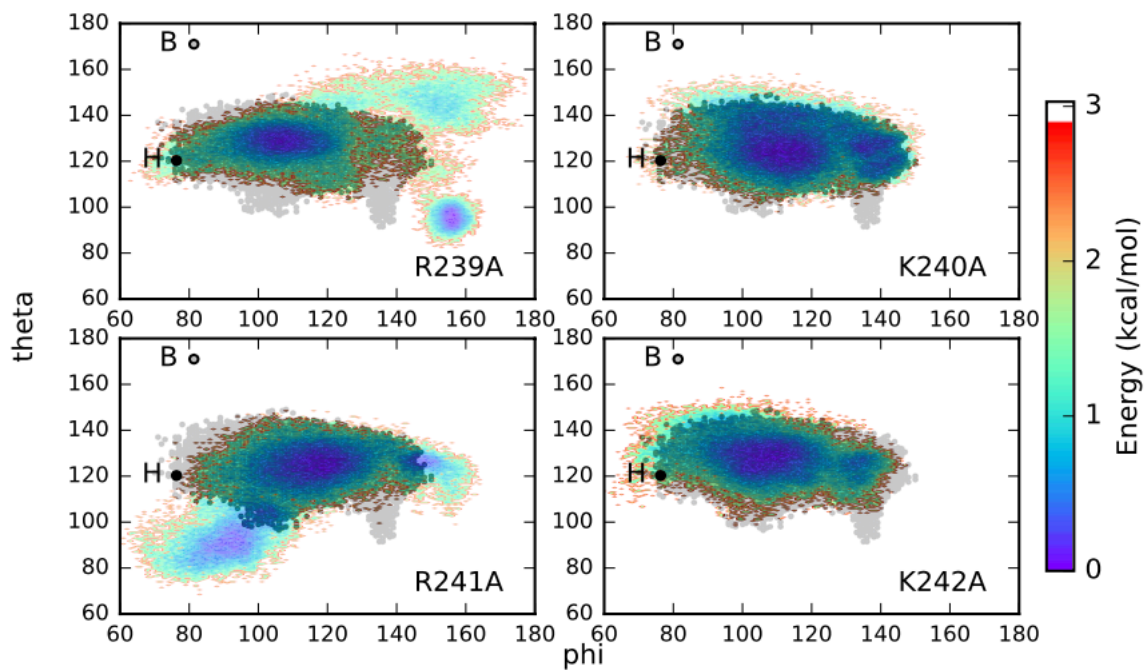


Figure 2.S2. Mutants free energy landscape in terms of spherical angles overlaid on the wildtype sampling conformation (gray outline).

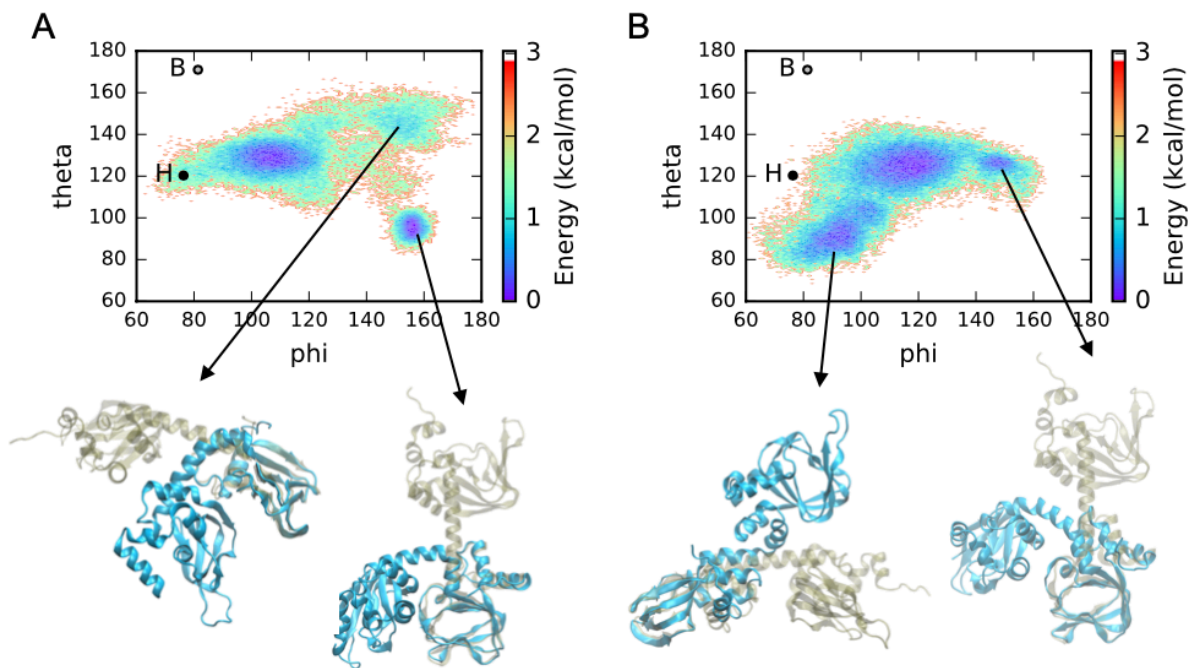


Figure 2.S3. Structural assignment to the free energy landscapes of (a) R239A and (b) R241A.

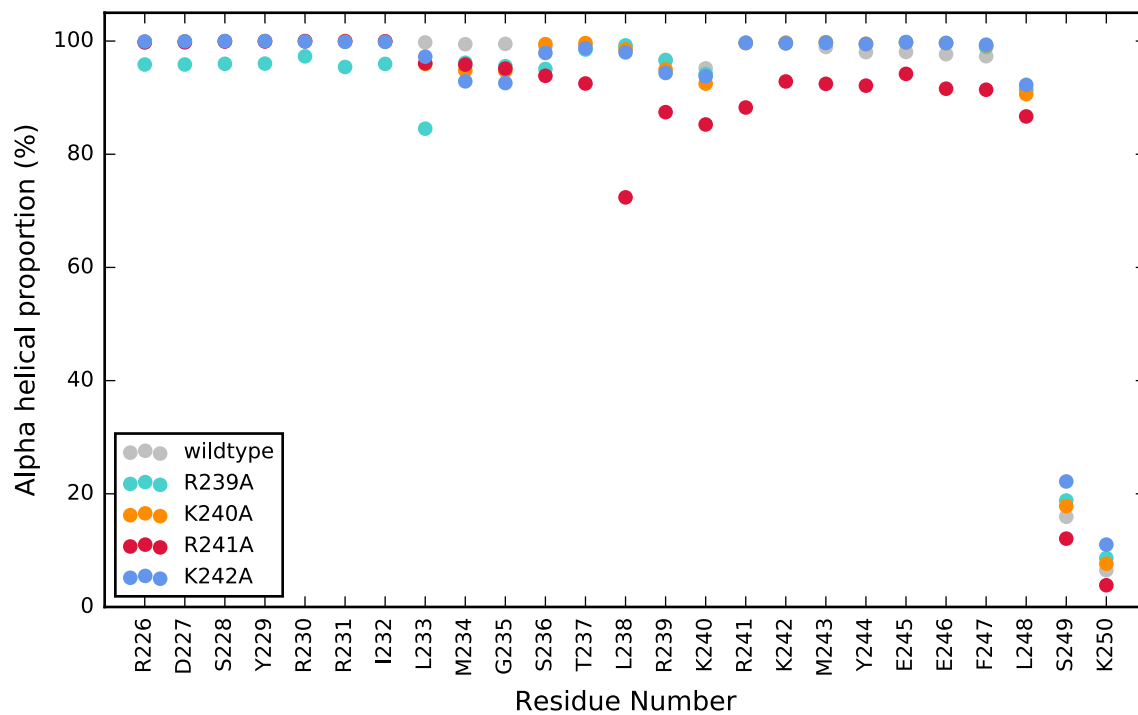


Figure 2.S4. Per-residue analysis of helical proportion for residues located in the B/C helix.

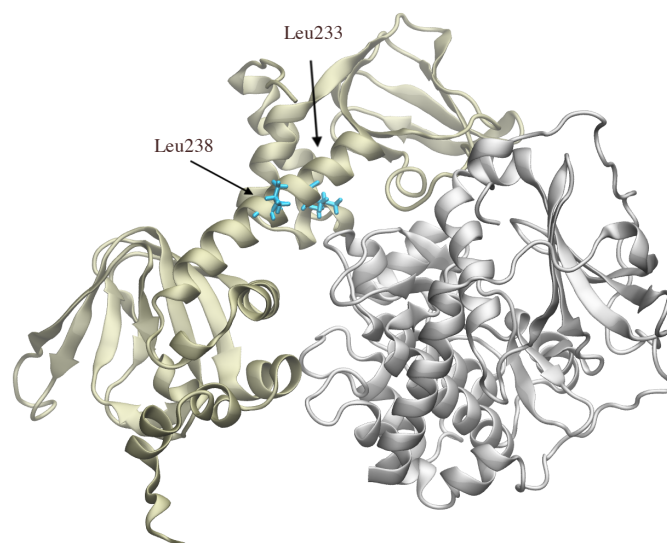


Figure 2.S5. Representation of the two residues in the B/C helix with smallest helical proportion as verified from the simulations (Leu233 in R239A and Leu238 in R241A simulations) overlaid on the holoenzyme crystal structure.

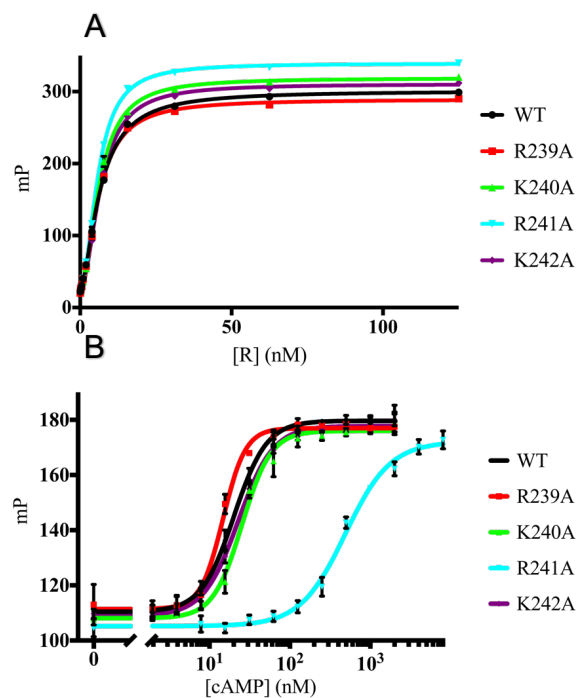


Figure 2.S6. Nucleotide Binding and Allosteric Activation of RI α B/C helix basic patch mutants. (a) To assess cyclic nucleotide binding to RI α , the fluorescent cAMP analogue, 8-[fluor]-cAMP, was used to measure fluorescent polarization to R subunits titrated at various concentrations (0 nM – 125 nM). (b) To evaluate allosteric activation of RI α mutant holoenzymes, polarization of the fluorescent PKA inhibitory peptide, 5/6-FAM-IP20, by binding to dissociated C subunit was assessed in response to titrating concentrations of cAMP (0 nM – 8000 nM). (a-b) The corresponding results summaries are shown in the tables right of the respective graphs, and represent the weighted mean of three independent experiments, each containing 3-4 sample replicates. Graphs were generated and analyzed using Graphpad Prism 6.

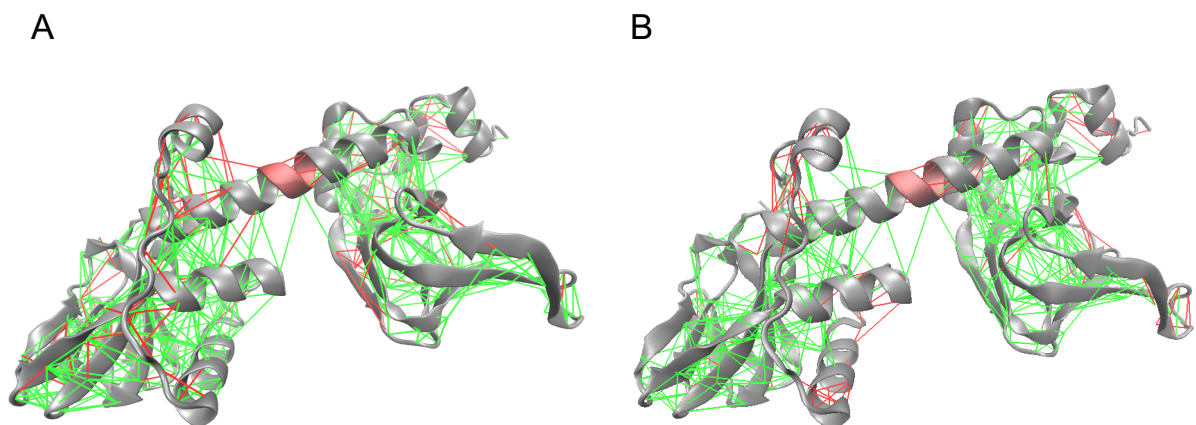


Figure 2.S7. Analysis of protein frustration using the protein frustratometer methodology. (a) mutational frustration and (b) configurational frustration. Highly frustrated contacts are represented as red lines, neutral contacts as grey line and minimally frustrated contacts as green lines. The backbone of the residues in the basic patch of the B/C helix are highlited in red.

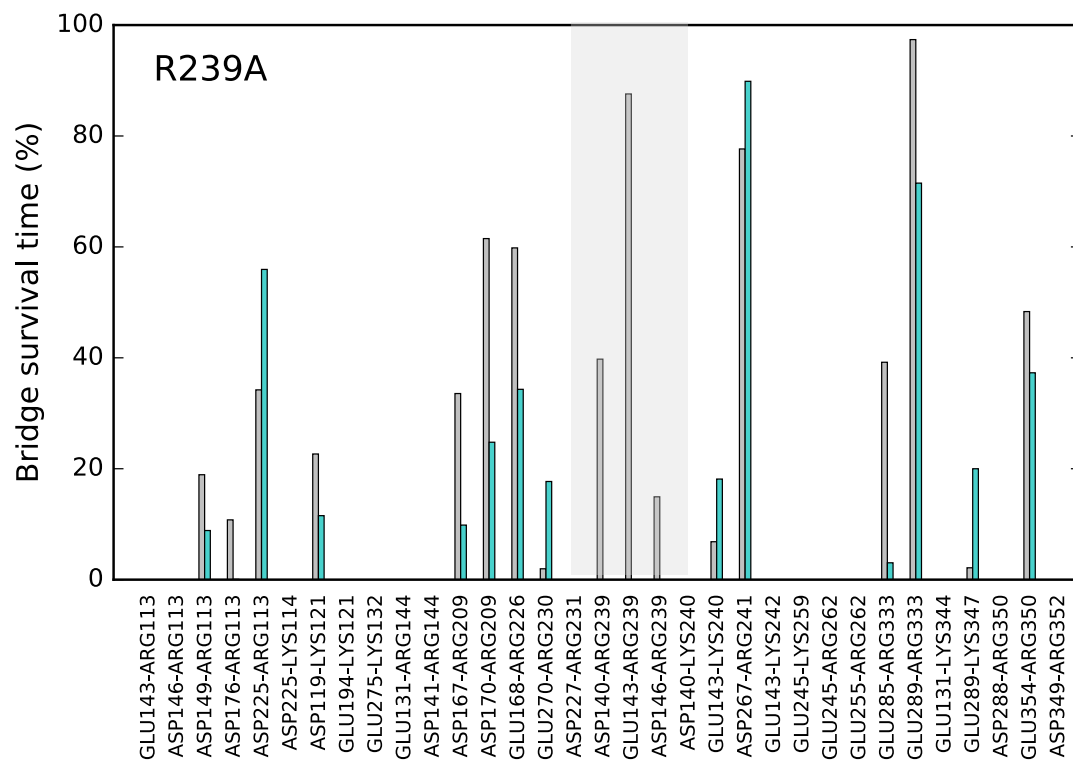


Figure 2.S8. R239A salt bridges' lifetime (in turquoise) for those that were altered by more than 10% with the mutation, compared to wildtype (gray). The abolished salt bridges are highlighted.

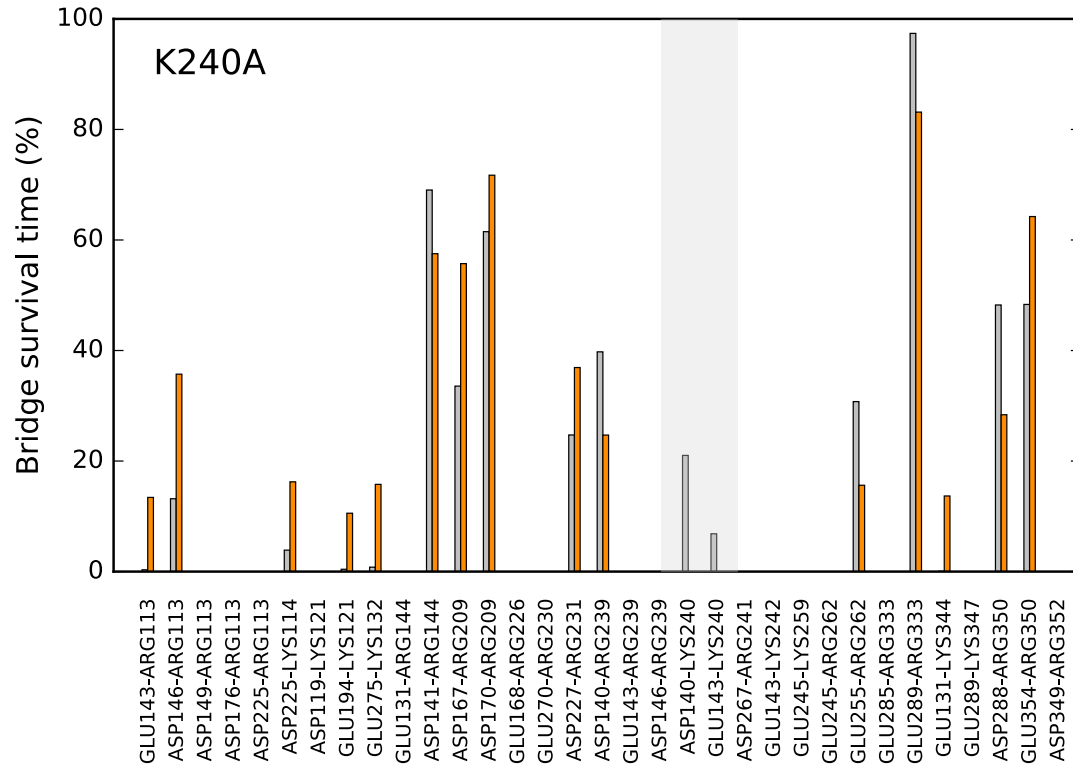


Figure 2.S9. K240A salt bridges' lifetime (in orange) for those that were altered by more than 10% with the mutation, compared to wildtype (gray). The abolished salt bridges are highlighted.

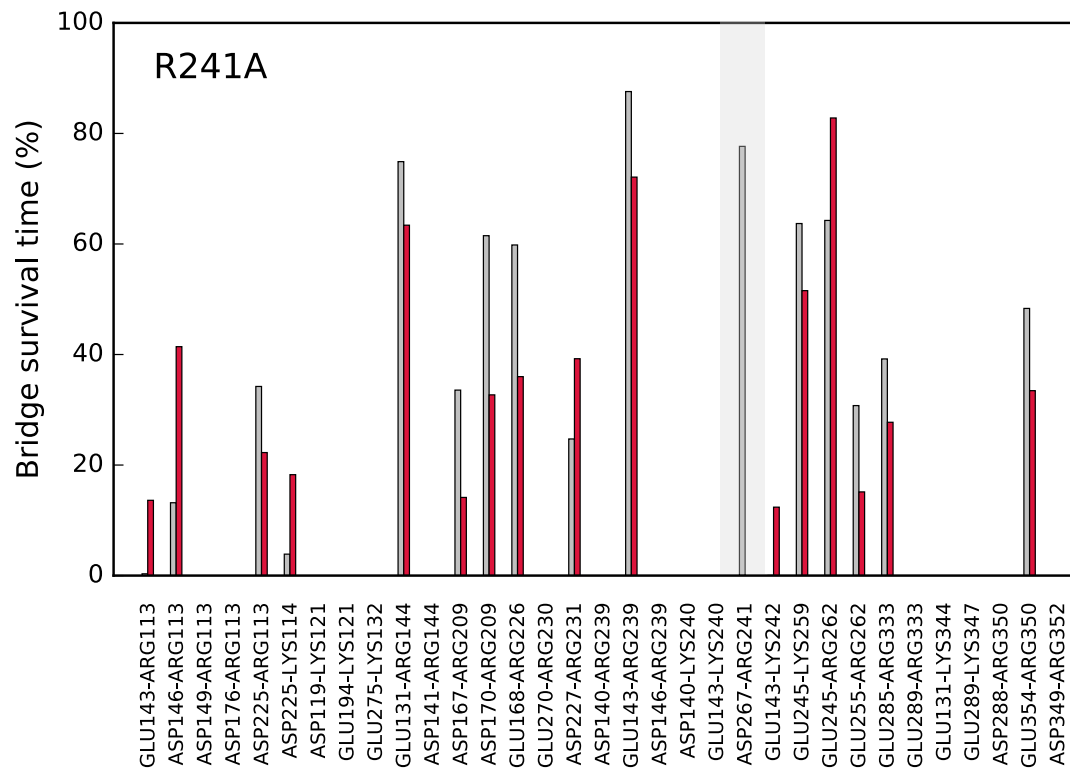


Figure 2.S10. R241A salt bridges' lifetime (in red) for those that were altered by more than 10% with the mutation, compared to wildtype (gray). The abolished salt bridge is highlighted.

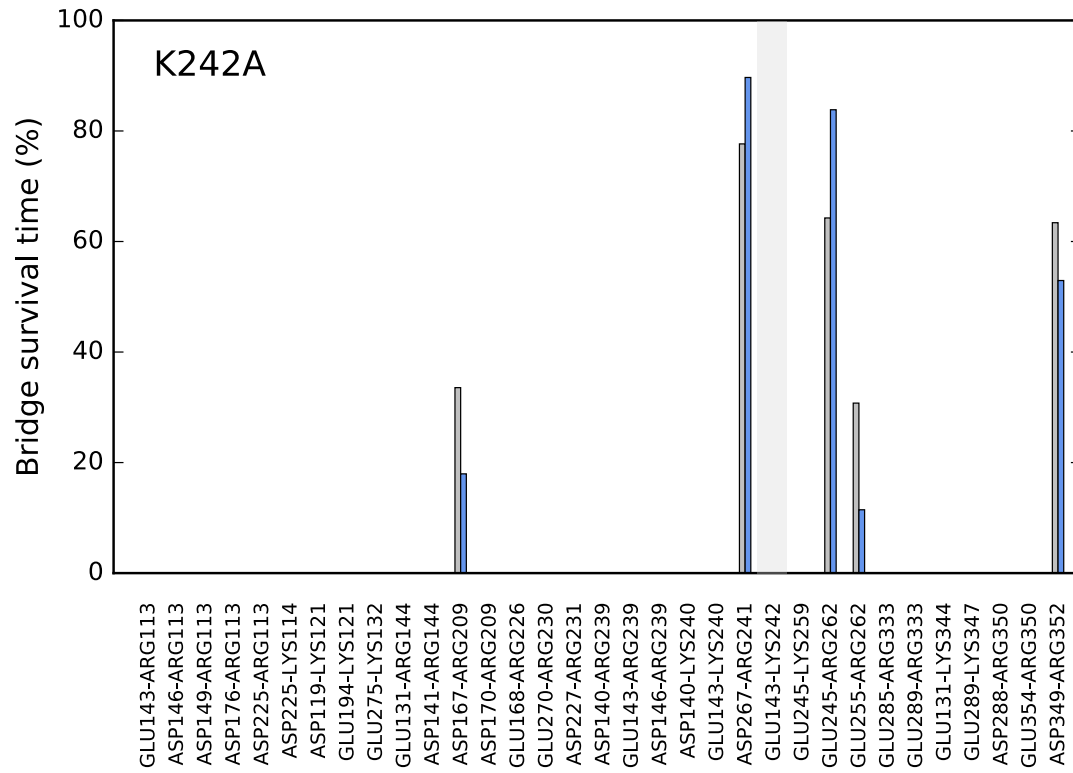


Figure 2.S11. K240A salt bridges' lifetime (in blue) for those that were altered by more than 10% with the mutation, compared to wildtype (gray). The abolished salt bridge is highlighted.

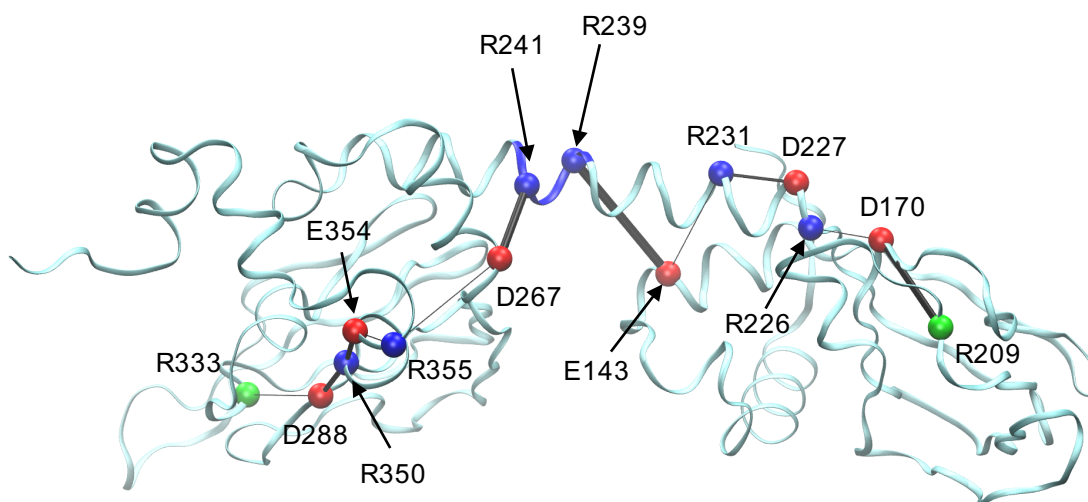


Figure 2.S12. Representation of the extended electrostatic network connecting the tandem cAMP binding domains. Basic residues are shown in blue and acidic residues, in red, with the exception of residues in the cAMP binding sites, Arg209 and Arg233, which are shown in green. The thickness of the black lines represent the lifetime of the salt bridge as measured in the wildtype simulation.

2.8 References

- (1) Kumar, S.; Nussinov, R. Close-Range Electrostatic Interactions in Proteins. *Chembiochem* **2002**, *3*, 604–617.
- (2) Tissot, A. C.; Vuilleumier, S.; Fersht, A. R. Importance of Two Buried Salt Bridges in the Stability and Folding Pathway of Barnase. *Biochemistry* **1996**, *35*, 6786–6794.
- (3) Elcock, A. H.; McCammon, J. A. Electrostatic Contributions to the Stability of Halophilic Proteins. *J. Molec. Biol.* **1998**, *280* (4), 731–748.
- (4) Jonsdottir, L. B.; Ellertsson, B. O.; Invernizzi, G.; Magnusdottir, M.; Thorbjarnardottir, S. H.; Papaleo, E.; Krisjansson, M. M. The Role of Salt Bridges on the Temperature Adaptation of Aqualysin I, a Thermostable Subtilisin-like Proteinase. *Biochim. Biophys. Acta* **2014**, *1884*, 2174–2181.
- (5) Zima, V.; Witschas, K.; Hynkova, A.; Zimová, L.; Barvík, I.; Vlachova, V. Structural Modeling and Patch-Clamp Analysis of Pain-Related Mutation TRPA1-N855S Reveal

- Inter-Subunit Salt Bridges Stabilizing the Channel Open State. *Neuropharmacology* **2015**, *93*, 294–307. <https://doi.org/10.1016/j.neuropharm.2015.02.018>.
- (6) Cui, G.; Freeman, C. S.; Knotts, T.; Prince, C. Z.; Kuang, C.; McCarty, N. A. Two Salt Bridges Differentially Contribute to the Maintenance of Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Channel Function. *J. Biol. Chem.* **2013**, *288* (28), 20758–20767. <https://doi.org/10.1074/jbc.M113.476226>.
 - (7) Bairagya, H. R.; Mukhopadhyay, B. P.; Bera, A. K. Role of Salt Bridge Dynamics in Inter Domain Recognition of Human IMPDH Isoforms: An Insight to Inhibitor Topology for Isoform-II. *J. Biomol. Struct. Dyn.* **2011**, *29* (3), 441–462. <https://doi.org/10.1080/07391102.2011.10507397>.
 - (8) Makhatadze, G. I.; Loladze, V. V.; Ermolenko, D. N.; Chen, X.; Thomas, S. T. Contribution of Surface Salt Bridges to Protein Stability: Guidelines for Protein Engineering. *J. Mol. Biol.* **2003**, *327* (5), 1135–1148. [https://doi.org/10.1016/S0022-2836\(03\)00233-X](https://doi.org/10.1016/S0022-2836(03)00233-X).
 - (9) Gur, M.; Madura, J. D.; Bahar, I. Global Transitions of Proteins Explored by a Multiscale Hybrid Methodology: Application to Adenylate Kinase. *Biophys. J.* **2013**, *105* (7), 1643–1652. <https://doi.org/10.1016/j.bpj.2013.07.058>.
 - (10) Zhang, L.; Buck, M. Molecular Simulations of a Dynamic Protein Complex: Role of Salt-Bridges and Polar Interactions in Configurational Transitions. *Biophys. J.* **2013**, *105* (10), 2412–2417. <https://doi.org/10.1016/j.bpj.2013.09.052>.
 - (11) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation Pathway of Src Kinase Reveals Intermediate States as Targets for Drug Design. *Nat. Commun.* **2014**, *5*, 3397. <https://doi.org/10.1038/ncomms4397>.
 - (12) Foda, Z. H.; Shan, Y.; Kim, E. T.; Shaw, D. E.; Seeliger, M. A. A Dynamically Coupled Allosteric Network Underlies Binding Cooperativity in Src Kinase. *Nat. Commun.* **2015**, *5*, 5939. <https://doi.org/10.1038/ncomms6939>.
 - (13) Lu, S.; Deng, R.; Jiang, H.; Song, H.; Li, S.; Shen, Q.; Huang, W.; Nussinov, R.; Yu, J.; Zhang, J. The Mechanism of ATP-Dependent Allosteric Protection of Akt Kinase Phosphorylation. *Structure* **2015**, *23*, 1725–1734. <https://doi.org/10.1016/j.str.2015.06.027>.
 - (14) Kim, C.; Cheng, C. Y.; Saldanha, S. A.; Taylor, S. S. PKA-I Holoenzyme Structure Reveals a Mechanism for CAMP-Dependent Activation. *Cell* **2007**, *130*, 1032–1043.
 - (15) Boettcher, A. J.; Wu, J.; Kim, C.; Yang, J.; Bruystens, J.; Cheung, N.; Pennypacker, J. K.; Blumenthal, D. a; Kornev, A. P.; Taylor, S. S. Realizing the Allosteric Potential of the Tetrameric Protein Kinase A R1 α Holoenzyme. *Structure* **2011**, *19* (2), 265–276. <https://doi.org/10.1016/j.str.2010.12.005>.
 - (16) Su, Y., Dostmann, R. G., Herberg, F. W., Durick, K., Xuong, N-h., Eyck, L. Ten, Taylor, S. S., Varughese, K. I. Regulatory Subunit of Protein Kinase A: Structure of Deletion Mutant with CAMP Binding Domains. *Science (80-.)*. **1995**, *269*, 807–813.

- (17) Bruystens, J. G. H., Wu, J., Fortezzo, A., Kornev, A. P., Blumenthal, D. K., Taylor, S. S. PKA RI α Homodimer Structure Reveals an Intermolecular Interface with Implications for Cooperative CAMP Binding and Carney Complex Disease. *Structure* **2014**, *22* (1), 59–69. <https://doi.org/10.1016/j.str.2013.10.012>.
- (18) Kornev, A. P.; Taylor, S. S.; Eyck, L. F. T. A Generalized Allosteric Mechanism for Cis-Regulated Cyclic Nucleotide Binding Domains. *PLoS* **2008**, *4* (4), 1–9. <https://doi.org/10.1371/journal.pcbi.1000056>.
- (19) Sjoberg, T. J.; Kornev, A. P.; Taylor, S. S. Dissecting the CAMP-Inducible Allosteric Switch in Protein Kinase A RI α . *Protein Sci.* **2010**, *19* (6), 1213–1221. <https://doi.org/10.1002/pro.400>.
- (20) Malmstrom, R. D.; Kornev, A. P.; Taylor, S. S.; Amaro, R. E. Allosterism through the Computational Microscope: CAMP Activation of a Canonical Signalling Domain. *Nat. Commun.* **2015**, *6* (May), 7588. <https://doi.org/10.1038/ncomms8588>.
- (21) Wu, J.; Brown, S. H. J.; von Daake, S.; Taylor, S. S. PKA Type II α Holoenzyme Reveals a Combinatorial Strategy for Isoform Diversity. *Science (80-.)*. **2007**, *318*, 274–279.
- (22) Zhang, P., Smith-Nguyen, E. V., Keshwani, M. M., Deal, M. S., Kornev, A. P., Taylor, S. S. Structure and Allosterism of the PKA RI β Tetrameric Holoenzyme. *Science (80-.)*. **2012**, *335* (6069), 712–716. <https://doi.org/10.1126/science.1213979>.
- (23) Ilouz, R., Bubis, J., Wu, J., Yim, Y. Y., Deal, M. S., Kornev, A. P. Ma, Y., Blumenthal, D. K., Taylor, S. S. Localization and Quaternary Structure of the PKA RI β Holoenzyme. *PNAS* **2012**, *109* (31), 12443–12448. <https://doi.org/10.1073/pnas.1209538109>.
- (24) Symcox, M. M.; Cauthron, R. D.; Ogreid, D.; Steinberg, R. A. Arg-242 Is Necessary for Allosteric Coupling of Cyclic AMP-Binding Sites A and B of RI Subunit of Cyclic AMP-Dependent Protein Kinase. *J. Biol. Chem.* **1994**, *269* (37), 23025–23031.
- (25) Gibson, R. M.; Ji-Buechler, Y.; Taylor, S. S. Interaction of the Regulatory and Catalytic Subunits of CAMP-Dependent Protein Kinase. **1997**, *272* (26), 16343–16350.
- (26) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; T.E. Cheatham, I.; Darden, T. A.; Duke, R. E.; Gohlke, H.; et al. AMBER 14. University of California, San Francisco 2014.
- (27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys* **1983**, *79*, 926–932.
- (28) Darden, T. A.; York, D.; Pedersen, L. Particle-Mesh Ewald: An N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (29) Humphrey, W.; Dalke, A.; Schulten, K. VMD-Visual Molecular Dynamics. *J. Molec. Graph.* **1996**, *14*, 33–38.

- (30) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532.
- (31) Wu, J.; Jones, J. M.; Nguyen-Huu, X.; Ten Eyck, L. F.; Taylor, S. S. Crystal Structures of RIalpha Subunit of Cyclic Adenosine 5'-Monophosphate (cAMP)-Dependent Protein Kinase Complexed with (Rp)-Adenosine 3',5'-Cyclic Monophosphothioate and (Sp)-Adenosine 3',5'-Cyclic Monophosphothioate, the Phosphothioate Analogues of cAMP. *Biochemistry* **2004**, *43* (21), 6620–6629. <https://doi.org/10.1021/bi0302503>.
- (32) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.
- (33) Parra, R. G.; Schafer, N. P.; Radusky, L. G.; Tsai, M.-Y.; Guzovsky, A. B.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer 2: A Tool to Localize Energetic Frustration in Protein Molecules, Now with Electrostatics. *Nucleic Acids Res.* **2016**, *44*, W356–W360. <https://doi.org/10.1093/nar/gkw304>.
- (34) Jenik, M.; Parra, R. G.; Radusky, L. G.; Turjanski, A.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer: A Tool to Localize Energetic Frustration in Protein Molecules. *Nucleic Acids Res.* **2012**, *40*, W348–W351. <https://doi.org/10.1093/nar/gks447>.

Improving the efficiency of ligand-binding protein design with molecular dynamics simulations

Emilia P. Barros¹, Jamie M. Schiffer², Anastassia Vorobieva^{3,4}, Jiayi Dou^{4,5}, David Baker^{3,4},
Rommie E. Amaro^{1,6}

Author affiliations:

1 – Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA

2 – Janssen Pharmaceuticals, Inc, San Diego, CA, USA

3 - Department of Biochemistry, University of Washington, Seattle, WA, USA

4 – Institute for Protein Design, University of Washington, Seattle, WA, USA

5 – Current address: Department of Bioengineering, Stanford University, Stanford, CA, USA

6 - National Biomedical Computation Resource, University of California, San Diego, La Jolla, CA, USA

3.1 Abstract

Custom-designed ligand-binding proteins represent a promising class of macromolecules with exciting applications toward the design of new enzymes or the engineering of antibodies and small-molecule recruited proteins for therapeutic interventions. However, several challenges remain in designing a protein sequence such that the binding site organization results in high affinity interaction with a bound ligand. Here, we study the dynamics of explicitly-solvated designed proteins through all-atom molecular dynamics (MD) simulations to gain insight into the causes that lead to the low affinity or instability of most of these designs, despite the prediction of their success by the computational design methodology. Simulations ranging from 500 to 1000 ns per replicate were conducted on 37 designed protein variants encompassing two distinct folds and a range of ligand affinities, resulting in more than 180 μ s of combined sampling. The simulations provide retrospective insights into the properties affecting ligand affinity that can prove useful in guiding further steps of design optimization. Features indicate that entropic components are particularly important for affinity, which are not easily incorporated in the empirical models often used in design protocols. Additionally, we demonstrate that the application of machine learning approaches built upon the output from the simulations can help discriminate between successful and failed binders, such that MD could act as a screening step in protein design, resulting in a more efficient process.

3.2 Introduction

Protein design is a young and ambitious field that aims to expand beyond naturally-occurring proteins to explore the massive protein sequence and fold spaces in the search for novel and customized structures^{1,2}. Successes in the design of novel folds^{3,4}, ligand-binding proteins⁵⁻⁷,

enzymes^{8,9}, antibodies¹⁰⁻¹² and self-assembling supra-molecular structures¹³⁻¹⁶ underscore this field's progress and growing potential. However, despite an increasing number of achievements, the protein design process remains very challenging and time consuming, with usually low success rates in initial design rounds^{11,17}.

Molecular recognition and protein-ligand binding are universally important processes that are however not yet fully understood or emulated. The development of novel molecules to treat diseases rely on the understanding of these interactions, and the improvement of protein-ligand affinity is far from being a negligible task¹⁸. In this context, the design of ligand-binding proteins offers the opportunity to better investigate the fundamentals affecting high affinity binding and selectivity^{1,5}, as well as laying out the foundations for custom design of *de novo* enzymes¹⁹, biosensors^{20,21} and antibody engineering^{22,23}. Designing ligand-binding proteins poses the extra challenge that protein scaffolds not only need to be structurally stable and fold in the intended conformation, but also include residues lining up the binding cavity that result in high-affinity interactions with the ligand. Thus, the functionalization of the binding site, generally with polar residues for the establishment of hydrogen bonds with the ligand, has to be balanced with the hydrophobicity of the protein core to maintain an energetically favorable folded state⁷ and the desolvation cost of the polar cavity upon ligand binding²⁴.

The general ligand-binding design protocol involves initial sampling of disembodied amino acids to create a binding site with specific protein-ligand interactions. The binding site is then positioned in a protein scaffold, and surrounding residues are further optimized to generate the desired interactions or to buttress the interactions in the secondary shell⁵. While tight ligand binders have been successfully generated by computational design, in a recent study 17 pre-selected designs of the nuclear transport factor 2 (NTF2) scaffold had to be expressed and tested

to yield two successful micromolar binders⁵, while the pool of tested candidates for the more hydrophobic fentanyl ligand involved 62 candidates, among which only three were successful first generation μM to nM binders²¹. More recently, the first completely *de-novo* designed β -barrel binding proteins required the generation of thousands of designs and experimental characterization of 56 high-scoring sequences to yield two successful binders in the first round of design generation⁷. This constitutes an expensive and lengthy process, as the computational design generation needs to be followed by expression of the highest-ranking candidates and experimental characterization, which includes assays to test proper protein folding and stability (such as circular dichroism and yeast-surface display), and ligand binding (e.g. fluorescence activation or polarization and isothermal titration calorimetry). Several challenges, including the evaluation of desolvation energies and sampling of alternate backbone conformations²⁵, affect the design accuracy, and thus the majority of proteins in the initial rounds of computational design end up failing the experimental validation. The most common sources of failure are due to improper protein folding (leading to aggregation and insolubility in many cases), or the absence of high-affinity interactions with the ligand. The few promising candidates from the first round of design can then be optimized by techniques such as site-saturation mutagenesis to yield tighter binding proteins, further lengthening the design process.

Protein function is directly determined by the macromolecule's structure and dynamics, and in this sense molecular dynamics (MD) simulations are uniquely poised to assist in the protein design process as the simulations can inform on the designs' dynamics beyond the static models generated by the empirical design protocols^{26,27}. MD simulations have been successfully applied at several different steps in the protein design methodology, such as in the refinement of predicted protein structures²⁸⁻³⁰ or in the identification of flexible regions in designed proteins. The

enumeration of mutations for the rigidification of these flexible sites³¹ led to verified increase in thermal stability experimentally³²⁻³⁴. Additionally, investigation of designed enzymes using short simulations provided indication of key disrupted catalytic interactions in unsuccessful designs, and was proposed as a screening method for enzyme design³⁵.

Here, we investigate the dynamics of 37 ligand-binding designs from two different scaffolds and designed to bind to distinct ligands using MD simulations (Figure 3.1a). We survey the dynamical properties that reveal rational explanations for the unexpected failure of some of these designs and explore the applicability of the simulations as a screening tool for early identification of the most promising designs prior to experimental validation (Figure 3.1b). Differences in protein structural flexibility, ligand dynamics, pocket pre-organization, and water dynamics inside the cavity provide evidence of the predictive power of MD simulations when used concomitantly with the protein design process. We find that the application of machine learning approaches to the descriptors generated from simulations of the 37 designs in a retrospective analysis allows for accurate classification of the models and identification of the tight binder designs. This reinforces the potential of using MD simulations in the protein design pipeline to achieve higher efficiency and success rates in the design of novel functional proteins.

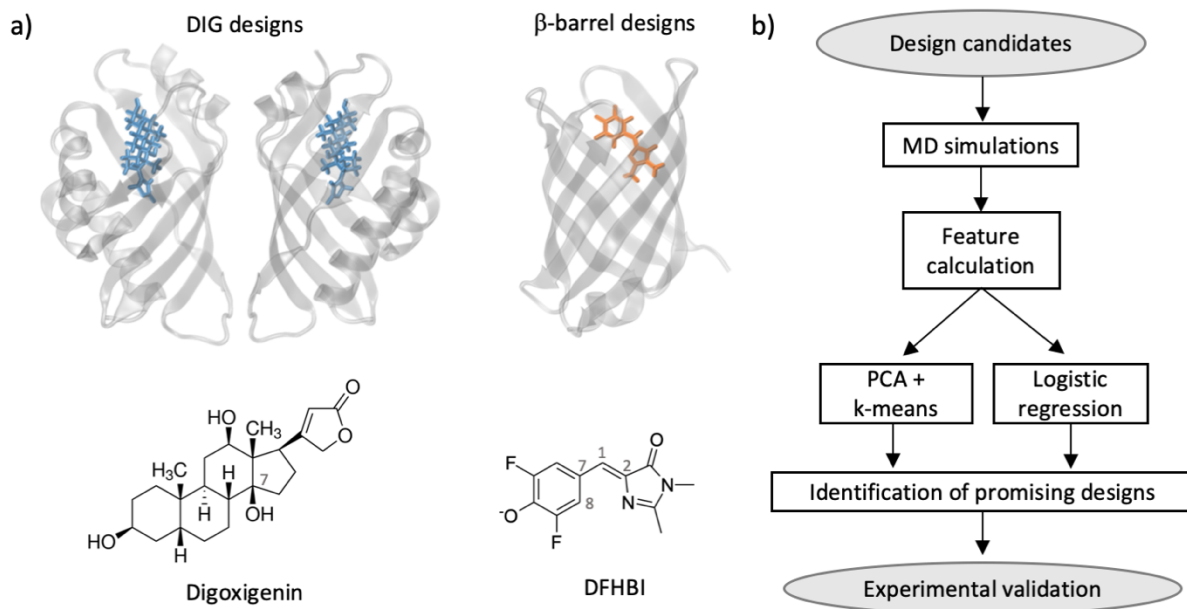


Figure 3.1. (a) Summary of the design data set used in the simulations, constituting two protein scaffolds (β -barrels and DIG designs) designed to bind distinct ligands. Representative designs are shown on the top with ligands highlighted, and the structure of the ligands are shown on the bottom panels. DFHBI stands for fluorogenic 3,5-difluoro-4-hydroxybenzylidene imidazolinone. Key atoms mentioned later in the text are indicated by their respective numbering in the ligand molecule. (b) Schematics of the prosed use of MD as a screening tool in the protein design process.

3.3 Methods

System selection and preparation

The starting structures for the simulations consisted of Rosetta-modeled ligand-binding proteins published previously^{5,7}. Thorough descriptions of the design methodologies and experimental characterization assays employed, including ligand affinity and selectivity measurements, can be found in references 5 and 7. Four designs of the digoxigenin-binders based on Nuclear Transport Factor 2 (NTF2) folds were selected (DIG10.2, DIG10.3, DIG12 and DIG16, here referred to as DIG designs)⁵, as well as 33 designs of the fluorogenic 3,5-difluoro-4-hydroxybenzylidene imidazolinone (DFHBI) binders based on a *de novo* β -barrel scaffold (Figure 3.1a)⁷. This set included examples of tight binders as well as unsuccessful designs, thus classified

due to failures to fold properly or to bind to the ligand with high affinity (Table 3.1). Besides the modeled structures, we also performed simulations starting from the X-ray crystal structures of DIG10.2 (PDB code 4J8T, chains A and B) and DIG10.3 (PDB code 4J9A, chains C and F) to investigate possible errors introduced by the use of design models instead of experimentally-validated structures. Missing terminal residues in the crystal structures of DIG10.2 and DIG10.3 were modeled with Schrödinger's Maestro (version 10.4, Schrödinger, LLC, New York, NY) based on the known protein sequence, and missing side chains were added with Schrödinger's Prime^{36,37}.

The digoxigenin and DFHBI ligands were parametrized using Antechamber and the generalized Amber force field (GAFF)^{38,39}, with geometry optimization performed with Gaussian 09⁴⁰. For DFHBI, the torsional parameters of the C2-C1-C7-C8 dihedral were increased to model the molecule's expected planarity due to its aromaticity in the bound fluorophore state. All starting structures were processed with Maestro-integrated PROPKA to assign protonation states at pH 7. Dowser⁴¹ was used to hydrate the protein cavity following removal of the ligand's coordinates for apo simulations. The proteins were solvated in water boxes with a buffer distance of 13 Å (for the β -barrels) or 15 Å (DIG designs) to the box edge with counter ions for charge neutrality, and 150 mM NaCl to simulate the experimental ionic concentration. The Amber14SB force field^{42,43} was used for the protein and NaCl, with TIP3P for the water molecules⁴⁴. As a note, we re-checked the protein protonation in the 33 β -barrel simulations after 500 ns of sampling and observed that about 30% of the histidine residues were assigned a different protonation state due to structural rearrangements, evidencing the limitations of conventional MD when it comes to fixed protonation states⁴⁵.

Table 3.1. Summary of designed protein data set^{5,7}

Design name	K _D	Classification*	Number of apo and holo replicates	Replicate simulation length (ns)
DIG10.2	8.9 nM	Binder	5	1,000
DIG10.2a	8.9 nM	Binder	5	500
DIG10.3	541 pM	Binder	5	1,000
DIG10.3a	541 pM	Binder	5	500
DIG12	-	Non-binder	5	1,000
DIG16	-	Non-binder	5	1,000
HBI_06	-	Non-binder	3	500
HBI_10	-	Non-binder	5	1,000
HBI_11	12.8 μM	Binder	5	1,000
HBI_15	-	Non-binder	3	500
HBI_19	-	Non-binder	3	500
HBI_21	-	Non-binder	3	500
HBI_22	-	Non-binder	3	500
HBI_24	-	Non-binder	3	500
HBI_26	-	Non-binder	5	1,000
HBI_27	-	Non-binder	3	500
HBI_32	49.8 μM	Binder	5	1,000
HBI_33	-	Non-binder	3	500
HBI_34	-	Non-binder	3	500
HBI_36	-	Non-binder	3	500
HBI_38	-	Unstable	3	500
HBI_41	-	Unstable	3	500
HBI_42	-	Non-binder	3	500
HBI_48	-	Non-binder	3	500
HBI_49	-	Non-binder	5	1,000
HBI_50	-	Non-binder	3	500
HBI_52	-	Non-binder	3	500
HBI_54	-	Non-binder	3	500
HBI_55	-	Non-binder	3	500
HBI_56	-	Non-binder	3	500
b11_loop	~0.5 μM**	Binder	3	500
b11L5F.1	~0.5 μM**	Binder	3	500
b11L5F_nC1	~0.5 μM**	Binder	3	500
b11L5F_nC2	~0.5 μM**	Binder	3	500
b11L5F_nC3	~0.5 μM**	Binder	3	500
b11L5F_nC4	~0.5 μM**	Binder	3	500
b11L5F.2	~0.5 μM**	Binder	3	500
mFAP0	~0.5 μM**	Binder	3	500
mFAP1	0.56 μM	Binder	3	500

* Unsuccessful designs are subdivided into two categories: “Unstable” for designs that showed improper folding or aggregation and “Non-binder” for folded designs that did not show ligand affinity within the sensitivity of the binding assays^{5,7}.

** estimated K_D values based on rough titration data from Dou, J. *et al*⁷

Molecular dynamics simulation protocol

All systems were simulated in their apo and ligand-bound (holo) states. We used the MD Kepler Workflow developed by our lab to enable automated MD minimization and equilibration for such a large number of systems⁴⁶. Minimization consisted of five stages: hydrogen only, solvent, solvent and ligand, side chains, and the full system resulting in 13,000 cycles using a combination of steepest decent and conjugate gradient methods. Since the majority of the starting structures were not experimentally-resolved conformations, we performed a long equilibration protocol and verified RMSD evolution to ensure system relaxation and convergence. Equilibration involved an initial heating to 100 K at constant volume for 50 ps followed by heating to 298 K at constant pressure, 1 bar, for 200 ps. The systems were further equilibrated at 298 K and 1 bar for 2.25 ns.

Molecular dynamics simulations were run using GPU accelerated Amber14^{42,47} as an NPT ensemble with periodic boundary conditions at 1 bar and 298 K to simulate experimental conditions. We used a non-bonded short-range interaction cutoff of 10 Å, and the long-range electrostatic interactions were approximated by particle mesh Ewald⁴⁸. The simulations used a 2 fs time step with the SHAKE algorithm to constrain hydrogen atoms. The initial design data set was simulated for five replicas of 1,000 ns each in the apo and holo states, while additional validation systems were simulated in three 500 ns replicas of each state (Table 3.1), resulting in a total sampling of 184 μs. MD input files are available for download at https://github.com/emiliapb/Design_screening.

Analysis methods

Trajectory files were visualized in VMD⁴⁹ and analysis were conducted using Jupyter notebooks⁵⁰ and in-house scripts, using a variety of MD analysis functions from MDTraj⁵¹,

CPPTRAJ⁵², PyEMMA⁵³ and MDAnalysis^{54,55}. The Jupyter notebooks are available for download on GitHub (https://github.com/emiliapb/Design_screening). Specifics of the different analysis methods conducted are discussed below.

Protein structural flexibility

Root mean square fluctuation (RMSF) of C α carbons was calculated using CPPTRAJ, following structural alignment of the protein to backbone atoms. To obtain a single value informative of structural flexibility for each design, we used PyEMMA's regular space clustering of the C α coordinates with RMSD metric and a cutoff of 1.5 Å to obtain the number of clusters (NOC) sampled. Simulations were analyzed every 100 frames.

To inform on the solvent accessibility of hydrophobic residues, we calculated solvent accessible surface area (SASA) of Ala, Ile, Leu, Phe, Val, Pro, Gly, Met and Trp residues for every frame in the simulations, using MDTraj's SASA function.

Ligand dynamics

We investigated ligand displacement in the holo simulations through the calculation of ligand root mean square deviation (RMSD) from the starting conformation in the designed models. Each trajectory was aligned to the protein coordinates of the respective starting structure, and RMSD values calculated using CPPTRAJ.

Pocket organization

Cavity volume was investigated using POVME (Pocket Volume Measurer), version 2.0⁵⁶. Volume calculations were performed for every 100 frames of the aligned trajectories. Inclusion

spheres were defined to encompass the binding site, and seed spheres were selected to include the minimal definition of the pocket, which were placed roughly at the center of the ligand position in the binding pocket. To allow for comparison across designs, the same POVME spheres were used for each scaffold and POVME's convex hull option was turned off.

The side chain chi1 dihedral angles of residues designed to interact with the ligand were investigated using MDTraj. For protein-ligand hydrogen bond analysis, MDTraj was used to calculate the distance between the hydrogen and acceptor atoms, and the angle formed between donor, H and acceptor atoms. H-bonds that fell within the definition of strong and moderately-strong bonds were counted (Strong = XH --- Y bond length of 1.2-1.5 Å and X-H---Y angle of 170-180°, moderate = bond length between 1.5-2.2 Å and angle 130-170°)^{35,57}.

Water analysis

To investigate the presence of water molecules inside the protein cavity, we counted the number of water molecules within a sphere delimiting the binding site using MDAnalysis. The delimiting region was selected as a sphere of radius 8 Å from the coordinates of the C7 atom in the digoxigenin ligand, and a sphere of radius 7 Å centered at DFHBI's C1 atom (Figure 3.1a). The survival probability function in MDAnalysis⁵⁸ was used to calculate water survival probability within the same defined spheres in the last 100 ns of the apo simulations.

Convergence analysis

To assess the influence of simulation length and number of replicas on the computed features, we used the protocol described by Knapp, Ospina and Deane⁵⁹, computing the average difference between the features for 100 rounds of random selection without replacement.

Machine learning

Python's scikit-learn library was used to perform unsupervised and supervised learning on the features extracted from 500 ns of simulations. For the designs that have been simulated for 1000 ns and 5 replicas, we used the first 500 ns of the simulation and the first three replicas (generated using random seeds) for the calculations. Feature scaling was performed among the designs of a particular scaffold to prevent dominance of larger-valued features. Logistic regression was performed with a tolerance of 10^{-4} and liblinear solver. K-nearest neighbors used $k=2$ or 5 , uniform weights and Euclidian distance metric.

3.4 Results and Discussion

Binding determinants for DIG designs

To investigate the dynamics of designed small-molecule binding proteins and understand the determinants affecting binding ability and affinity, we first investigated 4 designed proteins of the Nuclear Transport Factor 2 scaffold, which have been engineered to bind to the small molecule digoxigenin⁵ (Figure 3.1a). We conducted extensive simulations in both the apo and holo states of two successful, tight binder designs (DIG10.2 and DIG10.3) and two designs that failed to bind to the ligand despite positive predictions by the computational methodology (DIG12 and DIG16). DIG10.2 and DIG10.3 are third- and fourth-generation designs, respectively, generated following design optimization, and exhibit binding constants in the nano-molar to pico-molar range (Table 3.1). The structures of DIG10.2 and DIG10.3 have been solved by X-ray crystallography and were the starting structures of the simulations. DIG12 and DIG16, being considered failed designs, did not have their structures solved experimentally and the starting structures for the simulations were

modeled by Rosetta based on these designs sequences and the original scaffold from which they were engineered⁵.

We first focused on the dynamics of the proteins in the simulations. The projection of the RMSF values on the protein structures evidences that the highly fluctuating regions are located in the structural motifs lining up the cavity entrance, with DIG12 in particular exhibiting a larger flexible region than the tight binders and thus suggesting a possible negative effect of the protein flexibility in the accessibility of the cavity for ligand binding (Supplementary Figure 3.S1). We also looked at solvent accessibility of hydrophobic residues, since this is an important factor affecting protein stability. Figure 3.2a shows the average SASA/hydrophobic residue calculated for the designs. In line with what would be expected, tight binders show smaller SASA both in the apo and holo simulations, although DIG16 values are not as distinct from the tight binders as DIG12. This indicates that the successful designs not only tend to have a better organized hydrophobic core, but also confirms the importance of solvent shielding of nonpolar residues and promotion of hydrophobic interactions for adequate structural stability.

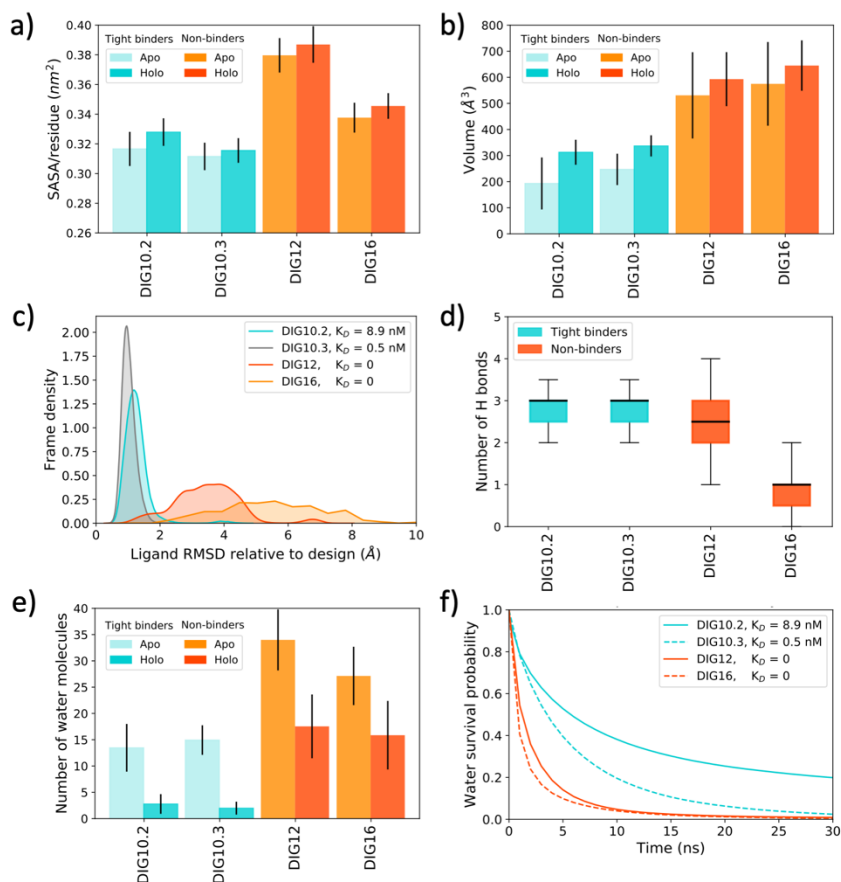


Figure 3.2. Evaluation of binding determinants for DIG designs. (a) Solvent accessible surface area (SASA) of hydrophobic residues. Tight binders are colored in turquoise and non-binders in orange. Results from apo simulations are shown in lighter colors, and holo simulations in darker colors. Error bars represent standard deviation across the five replicates. (b) Average cavity volumes for apo and holo simulations. Color scheme is the same as (a). (c) Ligand RMSD distribution for all replicates. (d) Box plot of the number of hydrogen bonds established between protein and ligand. Black line represents the mean value. The box extends to the lower and upper quartile and whiskers show the top 5 percentile and 95 bottom percentile of the data. (e) Average water count inside the cavity for apo and holo simulations. Coloring scheme is the same as (a). (f) Water survival probability in apo simulations for water molecules located within the cavity. Results are shown for one of the monomers only.

A key unanswered question in the design process of proteins functionalized for ligand binding is how stable and pre-organized the pocket remains in the absence of the ligand³⁵. We set out to explore the survival of the organized pocket through the calculation of cavity volume throughout the simulations. All designs, regardless of binding affinity, showed large variations of

cavity volume in the apo state, indicating that the pocket deviates from its designed conformation but that does not necessarily preclude ligand binding (Supplementary Figure 3.S2a). The non-binders DIG12 and DIG16 have occasional complete closure of the cavity, but the same also happens for the successful binder DIG10.2. The volume density of DIG10.2 at 50% of the frames, for example, shows a partially collapsed pocket (Supplementary Figure 3.S2b). Of the designs, DIG10.3 is the only one that shows apo pocket volume density that completely encompasses the volume occupied by the ligand in the bound state, demonstrating the pocket pre-organization achieved in the last round of design optimization. The C-alpha RMSF values (Supplementary Figure 3.1) from the MD simulations indicate that the lack of cavity pre-organization among the other designs is not only due to sidechain flexibility but also reflects backbone-level dynamics, which overcomes a major limitation of protein design protocols of not accounting for backbone flexibility.

On average, we observe much larger cavity volumes for the non-binders DIG12 and DIG16 than for the tight binders, both in the apo and holo simulations (Figure 3.2b). These designs' scaffolds have a large cavity opening, while the scaffold of DIG10, from which DIG10.2 and 10.3 were generated, presents a more enclosed cavity, such that the cavity volume results are a reflection of this. The simulations of DIG10.2 with ligand bound sampled a high number of cavity conformations with volumes smaller than what was originally designed, suggesting side chain rearrangements that result in a tighter interface around the ligand (Supplementary Figure 3.S2a). The fact that the designs with the sterically most accessible cavities resulted in the lowest affinities with the ligand sheds light on an interesting question: cavity accessibility may not be as important a factor for ligand affinity given these designs innate flexibility, and a "close-fitting" pocket may play a bigger role as it allows for stronger interactions with the ligand when in the bound state.

We furthered our study of pocket pre-organization by looking at the dihedral angles sampled by the residues side chains specifically designed to hydrogen bond to the ligand. The DIG designs present 3 interacting residues at the interior of the cavity in each of the monomer chains (Y34, Y101 and Y115 for DIG10.2 and DIG10.3, W57, H60 and H67 for DIG12 and Y39, H41 and N89 for DIG16), and we found that these remain in their designed conformer for a larger fraction of frames in the successful design simulations (Supplementary Figure 3.S3).

Besides probing protein dynamics, the simulations provide interesting insights into the dynamics of the ligand as well. In the holo simulations of the tight binders DIG10.2 and DIG10.3, very small ligand fluctuations are seen, with the ligands remaining very tightly bound in their original conformation in the cavity (Figure 3.2c). In the simulations of the non-binder examples, on the other hand, the ligand showed a large degree of displacement from its starting position, probably influenced by the larger size of the cavity, including complete dissociation from one of the monomers in 2 out of the 5 holo DIG16 trajectories (Supplementary Figure 3.S4). Somewhat surprisingly, the MD simulations were thus able to distinguish between tight binders and non-binder designs without requiring any ligand steering or information on the design's binding affinity.

We further investigated ligand-protein interaction by counting the number of hydrogen bonds established in the holo simulations. While there's a lot of fluctuation in the number of hydrogen bonds due to the dynamics of the ligands and side chains, DIG10.2 and DIG10.3 show a larger average number of hydrogen bonds than DIG12 and DIG16 (Figure 3.2d). For these successful designs, 3 stable hydrogen bonds are maintained with the designed interacting side chains located at the binding site interior, while additional transient hydrogen bonds are occasionally established with the ligand moiety located at the more solvent exposed opening of

the cavity. Interestingly, DIG12 also establishes a large number of transient hydrogen bonds besides those modeled in the design structure due to its larger ligand dynamics.

Binding is not only highly influenced by the direct interactions between the protein and ligand, but also by the dynamics of the water molecules surrounding them⁶⁰. Consequently, we also looked at the water molecules present in the apo and holo cavity interiors to investigate if there were any differences in water organization between the designs. The protein preparation steps preceding MD production of the apo state involved using Dowser⁴¹ to incorporate water molecules into the void left by the removal of the ligand from the model structure. We accompanied the presence of waters in the binding site by counting the number of molecules within a sphere delimiting the protein cavity. The holo simulations showed a smaller degree of water insertion in the cavity than the apo simulations due to the presence of the ligand (Figure 3.2e). In both states, the non-successful designs allowed for a greater degree of water insertion, promoted by the larger pocket volumes sampled during these simulations (Figure 3.2b).

Finally, as the absolute water count inside the cavity likely does not provide the full picture of the energetics of interactions, we calculated the survival probability of the waters inside the binding pocket in the apo simulations to get information on the stability of these molecules. As seen in Figure 3.2f, survival probability for the tight DIG binders decay more slowly than that for the non-binders, indicating presence of longer-lived waters inside these cavities and that successful design strategies involve promotion of favorable protein-water interactions.

Dynamics of modeled versus resolved crystal structures

A question might arise in the application of MD simulations for design screening regarding the accuracy and reliability of the results obtained from possibly inaccurately modeled starting

structures. This is a particularly valid concern since the prospective application of the simulations involves using structures predicted by Rosetta or some other protein design software that are not experimentally validated, as it antecedes experimental assays. In the above section, we described the dynamics and results obtained from the simulations of the crystal structures of DIG10.2 and DIG10.3. To try to address this question, we also performed simulations starting from the corresponding Rosetta-modelled structures for these designs, here named DIG10.2a and DIG10.3a. Simulations were run for 500 ns both in apo and holo states, and here we compare results from equivalent simulation lengths of the crystallographic structures. Overall, the dynamics obtained from the modeled structures showed similar distribution profiles to those derived from the simulations of the crystal structures (Figure 3.3). Hydrophobic SASA average values of modeled and resolved structure simulations are very similar to each other, as well as pocket volume distributions (Figures 3.3a and 3.3b). Water count inside the cavity is also comparable between modeled and resolved structure simulations (Figure 3.3c). In terms of ligand RMSD, DIG10.3a showed a significant tail of higher ligand RMSD values (Figure 3.3d), due to a large transient ligand displacement in one of the runs, but nonetheless the distributions sampled are still very distinct from that observed for the non-binders DIG12 and DIG16 (Figure 3.2c). Our results suggest that simulations starting from modeled structures are therefore accurate enough in sampling protein dynamics to be used with confidence in the assessment of these designs.

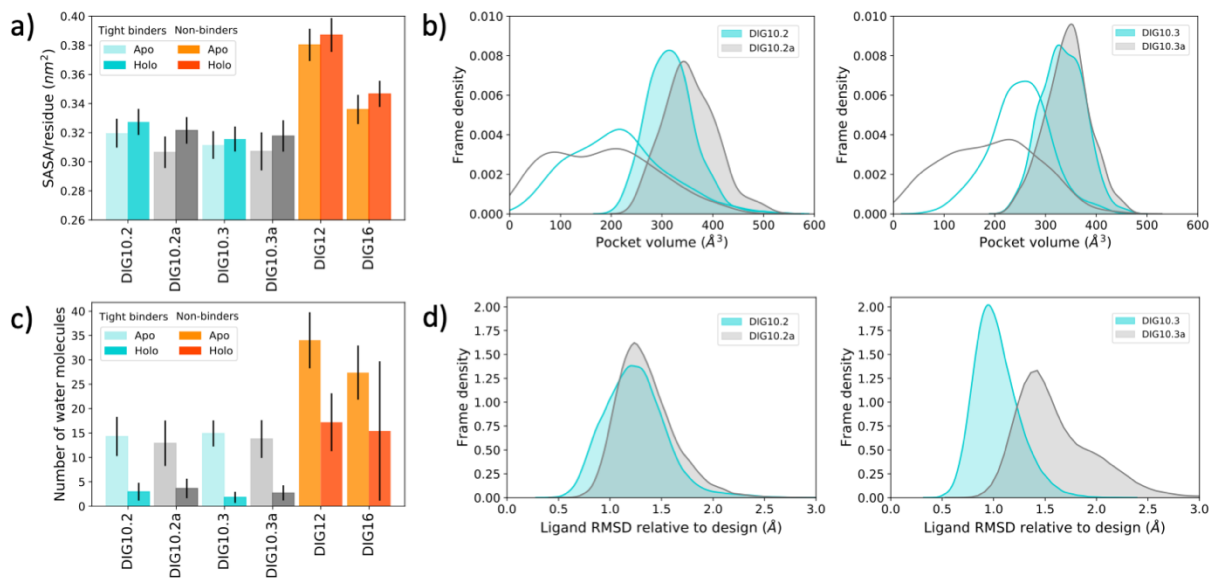


Figure 3.3. Comparison of results from the crystal (DIG10.2 and 10.3) and modeled structure (DIG10.2a and 10.3a) simulations (a) Solvent accessible surface area (SASA) of hydrophobic residues. Tight binders are colored in turquoise (for simulations starting from the crystal structure) or gray (for simulations starting from the modeled structures) and non-binders in orange. Results from apo simulations are shown in lighter colors, and holo simulations in darker. Error bars represent standard deviation across the five replicates (b) Pocket volume distributions. DIG10.2 and DIG10.3 are represented in turquoise, DIG10.2a and DIG10.3a are shown in silver. Holo simulation results are shown with filled curves, and apo simulations with just the curve outline. (c) Cavity water count. Coloring scheme is the same as in (a) (d) Ligand RMSD distributions. Coloring scheme is the same as in (b).

Validation on a distinct scaffold

The analysis of the DIG designs suggests the existence of energetic factors influencing binding ability which manifest themselves in key dynamical properties exhibited by successful binders. To validate these findings in a larger dataset and test the universality of the properties, we performed simulations of 33 experimentally-validated designs of a β -barrel scaffold⁷. These *de-novo* designed proteins not only represent a completely different protein scaffold than the DIG binders, but have also been designed to bind to a distinct small-molecule ligand, DFHBI⁷. Our data set includes 24 first-generation designs, which were all predicted to be tight binders by the

computational design methodology even though the majority was experimentally found to not be so: two were verified to be structurally unstable and thus not fold in the predicted β -barrel structure (HBI_38 and HBI_41), 20 fold properly but fail to bind to the ligand, and only two are successful tight binders, with ligand affinity values in the micro-molar range (Table 3.1). In the work of Dou *et al* these successful initial designs were further optimized, resulting in 9 second and third-generation designs with higher ligand affinity which have also been simulated and included in our analysis here (Table 3.1).

To first verify the necessary sampling time required for appropriate distinction between the designs, we performed convergence analysis of the initial DIG design results as well as a small set of the β -barrel designs (containing the 2 first-generation tight binders and 3 non-binders) which were run for five 1 μ s-long replicates in the apo and holo states, the same sampling length used for the previous designs. Analysis of the identified dynamical features indicated that the simulations do not need to be run so extensively, with results either reaching approximately constant values or maintaining constant relation among each other at around 500 ns (Supplementary Figure 3.S5). Moreover, estimates of the reliability and reproducibility achieved using different number and combination of replicates⁵⁹ indicates that three or four replicas yield property mean values satisfactorily converged and independent, within small variations, of the identity of the replicate simulations (Supplementary Figure 3.S6 shows results for HBI_10 and HBI_11). A key point in our exploration is that we do not intend to perform an exhaustive investigation of the design dynamics, as this would likely require extremely long simulations and defeat the purpose of using molecular dynamics to increase the efficiency of the design process. Instead, we aimed at obtaining sufficient sampling for insightful discrimination between the large number of design candidates. Therefore, in the interest of time efficiency, we performed subsequent simulations of the remaining

β -barrel designs as 500 ns triplicate runs in the apo and holo states, and the following results will be discussed for equivalent sampling times for all simulations.

In the analysis of this larger dataset, it became evident the need for an additional feature that would describe the conformational flexibility of the different designs. While RMSF is useful to investigate structural fluctuation incurred during the simulations, we turned to RMSD clustering of the C α coordinates to provide a single value to represent each design's flexibility and thus allow for a more direct comparison across the different proteins. As in the work of Demir *et al.*, the number of clusters (NOCs) thus obtained was used as a representation of structural flexibility since at least in principle more flexible proteins sample a larger conformational ensemble in the simulations, resulting in a higher number of clusters to represent the variation of the C α positions⁶¹. Figure 3.4 shows the 33 designs in terms of the structural and dynamical properties discussed above: structural stability (represented by the number of clusters and SASA of hydrophobic residues), cavity pre-organization (probed by number of frames in the apo simulations with volumes smaller than a cutoff which would prevent ligand binding, and holo average volume), insertion of waters in the cavity in holo simulations, and ligand dynamics (in terms of ligand RMSD and average number of protein-ligand hydrogen bonds per frame).

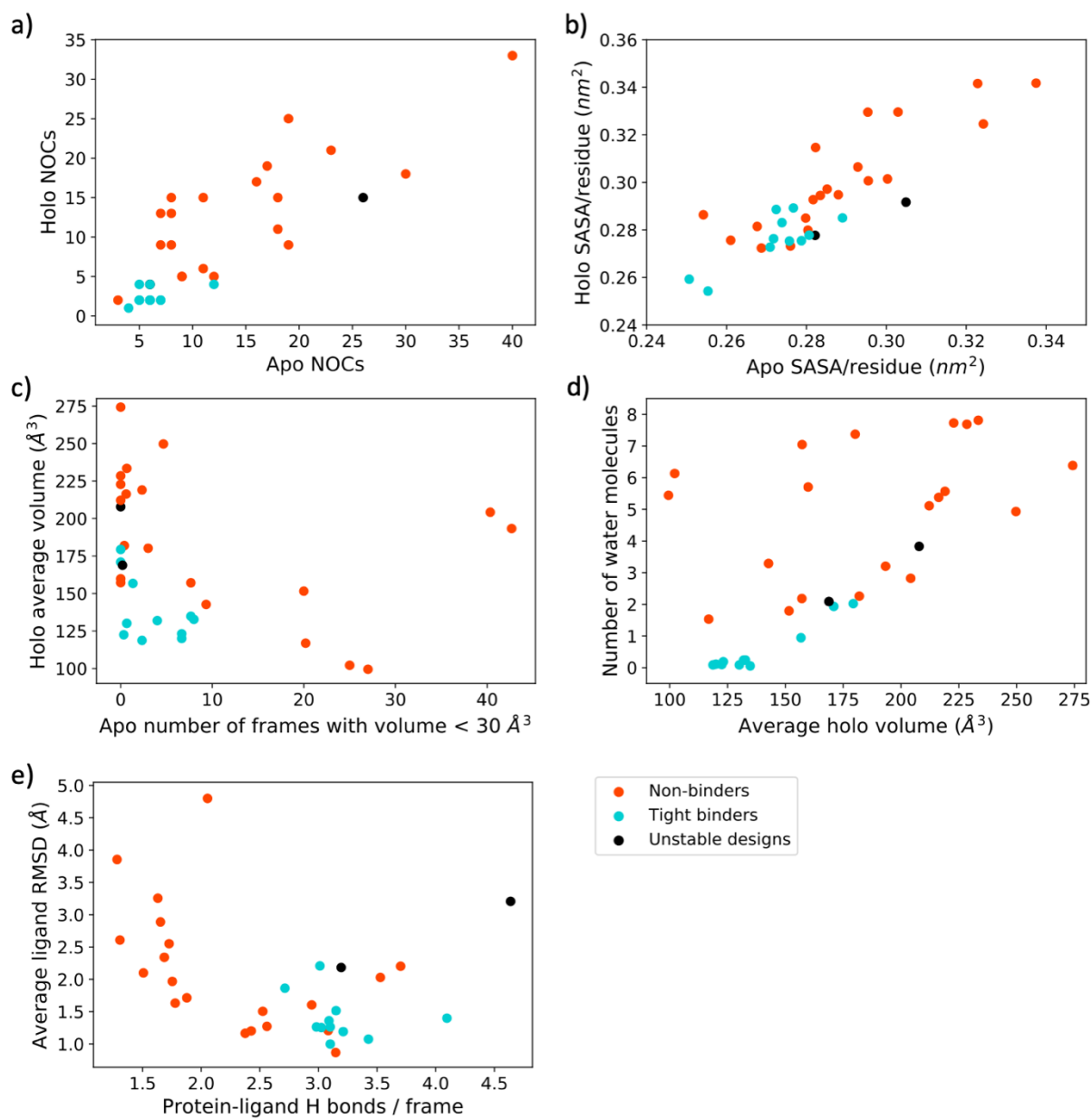


Figure 3.4. β -barrel designs distribution in terms of the identified discriminative features for design screening. (a) Number of clusters (NOC) analysis, (b) SASA of hydrophobic residues, (c) Number of frames with volume below a cutoff of 30 \AA^3 versus average holo cavity volume, (d) the same average holo cavity volume versus number of water molecules inserted in the pocket in the holo simulations and (e) average number of protein-ligand hydrogen bonds versus ligand RMSD. Non-binders are shown in orange, successful designs are shown in turquoise and structurally unstable designs in black.

The profiles in Figure 3.4 support and accentuate the trends observed from our initial reduced data set and evidence that these structural and dynamical descriptors can be useful in the classification of candidate designs. Importantly, the analysis of the same features computed from the original Rosetta-modelled design structures does not evidence any such distinction between the design categories (Supplementary Figure 3.S7), such that the descriptors generated from the static structures are not sufficient to distinguish successes from failures. The incorporation of dynamics, however, indicates that the non-successful designs in general are much more flexible and explore a wider range of conformations (Figure 3.4a, some designs have equal values of apo and holo NOC and overlap in the graph), suggesting that for this scaffold, failure to bind to the small ligand may arise from the lack of accounting for structure dynamics in the structure prediction methods. We observed that several non-binders were structurally destabilized by the introduction of the ligand in the holo simulations, leading to some dramatic structural deformations in some cases (Supplementary Figure 3.S8b). Tight binders, on the other hand, tended to be stabilized by the ligand in the holo simulations as indicated by the dampening of fluctuations in the RMSF plots (Supplementary Figure 3.S8). The number of clusters analysis is particularly promising in that it may allow for early identification of non-stable designs, since HBI_41, one of the two designs that did not fold in the β -barrel structure in our data set, displayed one of the largest number of cluster pair values. The solvent exposure of hydrophobic residues does not allow for such a clear distinction between the design classes, but it's possible to see in Figure 4b the suggestion of an empirical threshold at around 0.29 nm² apo and holo SASA beyond which only non-successful designs can be found.

The apo simulations of some of these designs showed such a large number of frames with completely collapsed pockets that it became evident that another useful discriminating metric

would be something that could capture this phenomenon. Here, we chose that as the number of frames with cavity volumes below a cutoff of 30 \AA^3 , which represents a pocket volume too small to allow for ligand binding (the smallest pocket volume observed in the simulations with ligand bound was 33 \AA^3). This descriptor, in conjunction with the average cavity volume in the holo simulations, permits successful distinction between most of the designs (Figure 3.4c). Comparison of the same average cavity volume with the number of water molecules that insert into the pocket in the holo simulations evidences that while there's a lot of variability for the non-binders, the successful designs cluster around smaller cavities and a reduced number of inserted water molecules (Figure 3.4d).

Finally, as for the DIG designs, probing ligand dynamics also provides valuable information for design identification (Figure 3.4e). While several outliers can be seen, all successful designs show a higher number of ligand-protein hydrogen bonds and reduced ligand dynamics as indicated by the low ligand RMSD values. The incorporation of protein and ligand dynamics into these designed scaffolds provide important additional information that can thus aid design selection, since all generated designs had been originally intended to form at least four hydrogen-bonding interactions with the ligand⁷.

Discriminative models for design screening

For some of the features in Figure 3.4 it is possible to imagine cutoffs of acceptable or promising values exhibited by proteins with favorable ligand interaction that could be used for prospective design predictions. However, as would be imagined from the complexity of the process investigated, each of these descriptors is not perfect in its discernment of binding ability, and we can see the likelihood of both false positive and false negative assignments. We hypothesized that

taking the features jointly into account would result in a more accurate design classification, given the multi-dimensionality of the problem. We performed Principal Component Analysis (PCA) on the scaled features and projected the β -barrel dataset into three principal components (PC) which describe 80% of the data variance (Figure 3.5a). Confirming our hypothesis, the successful binders cluster together in regions of smaller PC1 and PC2, while the non-binders are more spread along the principal components. This is in line with the general notion that protein-ligand binding can be negatively impacted by several causes, and that only a specific (almost serendipitous) combination of the properties result in a tight interaction.

The contributions of each of the features to the principal components can be analyzed to try to rationalize the energetic causes most highly affecting ligand-binding (Figure 3.5b). Entropy seems to play the most pronounced role in determining binding, as properties such as water dynamics in the cavity, protein conformational flexibility, cavity volume and ligand dynamics show the highest contributions in the first principal component. Ligand-protein induced fit comes in as a second determining factor, with the number of frames with too-small cavity volumes to allow for ligand insertion showing negative correlation with design binding ability. Finally, enthalpic components appear in the third PC, encoded by the SASA and number of hydrogen bonds established with the ligand.

Even though the successful and non-binder designs concentrated in different areas of the PC map, the separation is not absolute and there are overlaps or outliers among the two classes. Interestingly, the two first generation successful binders are the ones located closer to the area occupied by the non-binders, while the second and third generation higher affinity binders cluster more closely together, evidencing the successful enhancement of the energetic properties by the experimental optimization. While by visual inspection it can be hard to define a separating line

between the classes, we turned to unsupervised learning and clustering to see if such regions could be determined in an unbiased manner. Using the k-means algorithm, the designs were not accurately clustered when only two clusters were used, but assigning the data to three distinct clusters yielded interesting results with good clustering quality (average silhouette value of 0.45): One of the clusters was enriched in designs from the tight binding class, while the others contained only examples from the non-binding designs (Figure 3.5c).

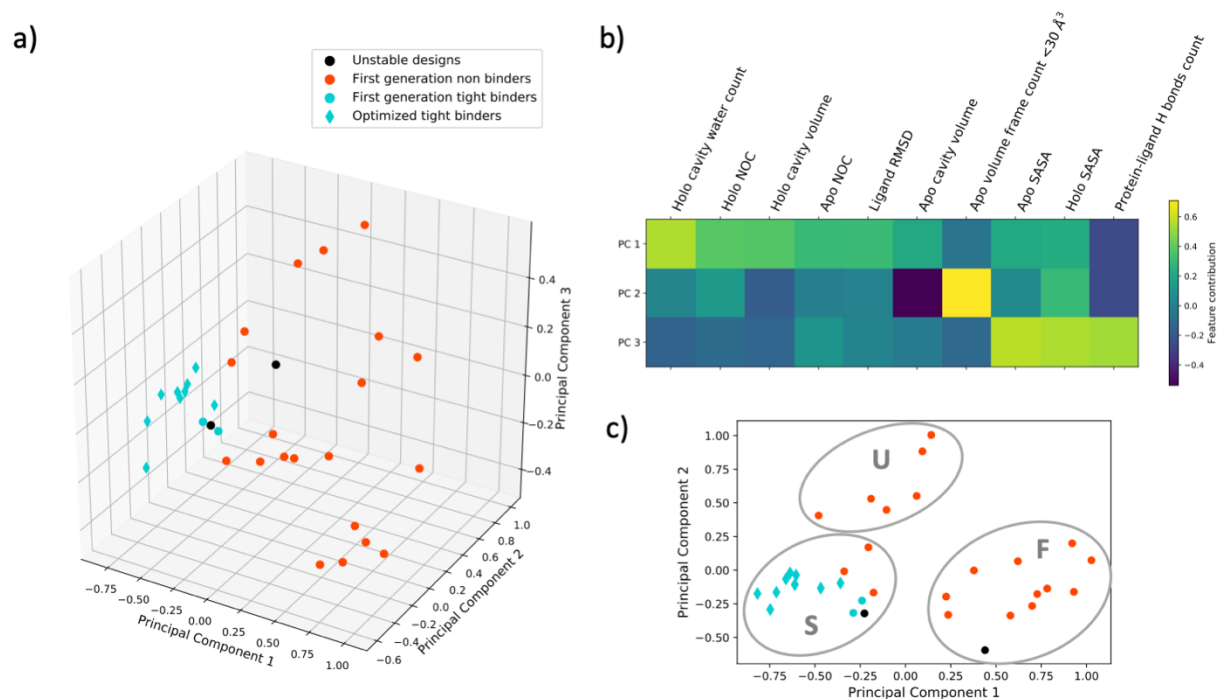


Figure 3.5. Unsupervised learning model for design classification. (a) Three-dimensional plot of β -barrel simulations distribution in terms of the principal components of the discriminative features. Non-binders are shown in orange, structurally unstable designs in black, first-generation tight binders shown as turquoise circles and optimized tight binders as turquoise diamonds. (b) Color representation of feature contribution to the principal components. (c) Representation of cluster assignment on the 2-dimensional plot in terms of principal components 1 and 2. Clusters are named Successful (S), Uncertain (U) and Failed (F) clusters.

One of the clusters, located in the area of higher PC1 values, included designs that showed clearly unstable dynamics from the simulations (such as the designs shown in Supplementary Figure 3.S8a), and can be interpreted as the Failed (F) cluster. The second cluster of only non-binders contained members in the boundary region with the tight binders and exhibited dynamics that would be hard to be accurately classified by visual inspection of the simulations. For this reason, we termed this the Uncertain (U) cluster. Their main distinction from the successful binders is captured by the collapse of the cavity in apo simulations incorporated into Principal Component 2, as all of these designs show small pocket volumes or completely closed cavities for significant portions of the simulations. Finally, the cluster to which all of the tight binders were assigned, here termed Successful (S) cluster, contains only 4 incorrectly classified designs. One of the false positives in this classification is HBI_38, the non-stable design that did not display as different feature profiles as HBI_41 in Figure 3.4. HBI_38 and the tight binder HBI_11 differ by only two mutations in the N terminal that lead to the formation of a stabilizing intramolecular disulfide bridge in HBI_11, and thus the misclassification of HBI_38 is not surprising given the likely much longer timescale that would be required to properly sample the difference between these designs structural stability.

Remarkably for such a complex problem, the unsupervised learning approach here employed on the features measured from the simulations was thus able to identify the high affinity binders with only 4 inaccurate classifications and no false negative assignments. We estimate that the early identification of the 12 unsuccessful designs from the F cluster and the 6 designs from the U cluster could have saved about 6 weeks of work, including protein expression, purification, folding and binding assays. However, this is likely to be a lower estimate as the Baker lab is very well equipped for protein characterization and the entire process could probably take 2 to 3 times

longer in a different lab. On the other hand, the MD system preparation, simulation and analysis workflow greatly automates the required steps such that the whole set of proteins can be simulated and analyzed in less than 2 weeks, using parallel GPUs and requiring minimal human intervention.

This unsupervised approach is useful to identify inherent differences between designs of the same structural scaffold, but lacks transferability with the DIG design results (Supplementary Figure 3.S9). However, taking advantage of the availability of experimentally-validated labels, we explored the use of supervised learning for the classification of the joint design scaffold^{62,63}. K-nearest neighbors and logistic regression classifiers were trained using 5-fold cross validation on the 10 dynamical fingerprints identified in the joint, 37 β -barrel and DIG design simulations, and showed good classification performance (Table 3.2). The precision values, the rate of true positive classifications over all positive assignments (including false positives), indicate the presence of misclassified non-binders. However, the recall metric at 1.0 for both algorithms, given as the ratio of true positive assignments over all assignments of the real positive class (including false negatives), indicates a complete absence of tight binders being classified as unsuccessful designs. In the same way, the high accuracy of the classifications and the Matthew correlation coefficient (MCC) and F_1 scores, all used as measures of a classifier performance and with a maximum value of 1.0 for a perfect classification, evidence the generality of the proposed approach. Moreover, the feature weights of the logistic regression model indicate that pocket dynamics plays the most determinant role for identifying non-binders, quantified by the insertion of water molecules in the cavity when in the ligand-bound state and the collapse of the cavity when in the absence of the ligand (Supplementary Figure 3.S10). Logistic regression, in particular, resulted in good classifiers even when trained on small sets (50% or even 33% of the dataset, Supplementary Table

3.S1), suggesting that not many designs need to be experimentally validated in order to yield accurate predictions in a prospective study.

Table 3.2. Evaluation of the supervised learning classifiers using 5-fold cross validation^a.

Validation metric	Classification algorithm	
	k-nn (k = 5)	Logistic regression
Accuracy	0.84 ± 0.16	0.93 ± 0.10
Precision	0.79 ± 0.21	0.87 ± 0.17
Recall	1.0 ± 0.0	1.0 ± 0.0
MCC	0.74 ± 0.25	0.87 ± 0.16
F1S	0.87 ± 0.13	0.92 ± 0.10

^a Values correspond to average and standard deviation of the 5 cross validations.

Finally, to further test the universality of this approach, we constructed models solely on the β -barrel designs and checked the predictions on the DIG dataset. With a large set of β -barrel designs available, we further split the data into training and validation sets to verify absence of overfitting. Using logistic regression, training the model on 70% of the β -barrel designs yields perfect classification of the designs of the distinct DIG scaffold (Table 3.3). Conversely, models trained solely on the 4 DIG designs display lower accuracy and precision due to the much smaller training set in this case, but the recall still indicates a perfect absence of false negative classification (Table 3.4). Interestingly, the 12 non-binders correctly identified correspond exactly to the designs classified in the F cluster using unsupervised learning. Regardless of the classification approach employed, the computation of dynamical fingerprints⁶⁴ from molecular dynamics simulations of designed proteins, thus, emerges as a potential general and scaffold-independent screening methodology to aid the challenging protein design process (Figure 3.1b).

Table 3.3. Evaluation of the generality of the classifiers, with models trained exclusively on the β -barrel designs^a.

Validation metric	Training + validation set: 33 β -barrel designs (70:30 split)			
	Test set: 4 NTF2 designs			
	k-nn (k = 5)		Logistic regression	
	Validation set	Test set	Validation set	Test set
Accuracy	0.83 \pm 0.09	1.0 \pm 0.0	0.91 \pm 0.08	1.0 \pm 0.0
Precision	0.70 \pm 0.15	1.0 \pm 0.0	0.80 \pm 0.19	1.0 \pm 0.0
Recall	0.98 \pm 0.05	0.95 \pm 0.15	1.0 \pm 0.0	1.0 \pm 0.0
MCC	0.71 \pm 0.13	0.96 \pm 0.13	0.83 \pm 0.16	1.0 \pm 0.0
F1S	0.81 \pm 0.09	1.0 \pm 0.0	0.87 \pm 0.13	1.0 \pm 0.0

^a Values correspond to average and standard deviation of 10 rounds of random splits of the data set according to the 70%:30% training:validation ratio.

Table 3.4. Evaluation of the generality of the classifiers, with models trained exclusively on the DIG designs.

Validation metric	Training set: 4 DIG designs	
	Validation set: 33 β -barrel designs	
	k-nn (k = 2)	Logistic regression
Accuracy	0.70	0.70
Precision	0.52	0.52
Recall	1.0	1.0
MCC	0.53	0.53
F1S	0.69	0.69

3.5 Conclusions

In this work, we used MD simulations to investigate the dynamics of designed ligand-binding proteins as a source of insight into the failure of some of these designs to bind to the ligand with high affinity. It became evident that the design model generated by the protein design protocol may differ from the ensemble of structures accessed by the simulations, such that the modeled structural descriptions can be further enriched by the incorporation of dynamic fingerprints.

The results obtained here suggest that successful and non-successful designs differ in their dynamical properties. Entropic components play a significant role in determining ligand affinity, which are complex and often very challenging to incorporate in the empirical models of protein design. Easily measured MD-realized descriptors (including number of clusters, cavity volume, hydrophobic solvent-accessible surface area, water count in cavity and number of protein-ligand hydrogen bonds) allow for the investigation of multiple design candidates, and analysis of these enthalpic and entropic feature profiles in a data set of 33 β -barrel designs resulted in a 88% accuracy of binding ability classification using unsupervised learning. This data set included 24 first-generation designs, among which only two were found to bind with high affinity, and 9 optimized second and third-generation designs⁷. The application of the unsupervised learning method in the screening of the first generation designs would result in the identification of the two successful binders and 4 false positive non-binders, which constitutes a 4-fold enrichment $((2/6)/(2/24))$ over the initial candidate design data set and a minimum net time and effort “savings” of one month of work. Moreover, the application of supervised learning in the form of k-nearest neighbors or logistic regression classifiers on the full dataset consisting of two different protein scaffolds resulted in accurate classification with no false negatives, suggesting the generality of this approach. The results here described emphasize how MD can act as a promising

screening step in the protein design process, avoiding the experimental testing of non-stable and low affinity designs and increasing the efficiency of the pipeline.

3.6 Acknowledgements

Chapter 3, in full, is a modified reprint of the material as it appears in “Barros, E. P., Schiffer, J.M., Vorobieva, A., Dou, J., Baker, D., Amaro, R. E., Improving the Efficiency of Ligand-Binding Protein Design with Molecular Dynamics Simulations, *Journal of Chemical Theory and Computation*, vol. 15, 2019. The dissertation author was the primary investigator and author of this paper.

3.7 Supplementary Information

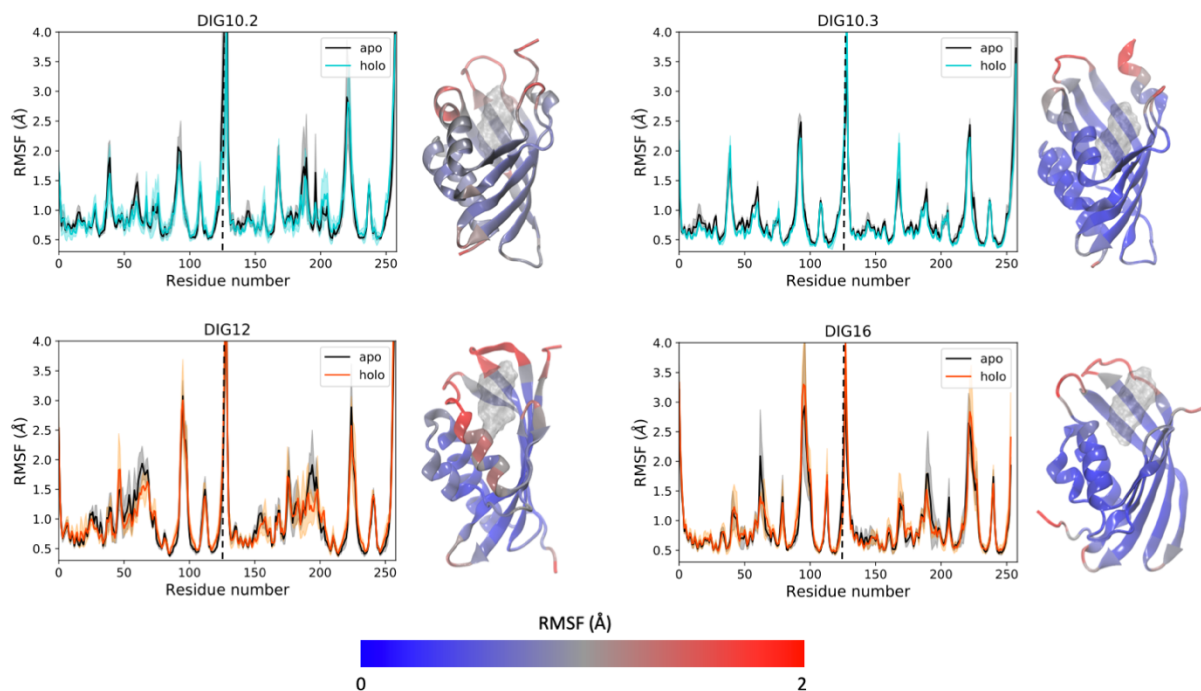


Figure 3.S1. RMSF values for the DIG designs (left panels). Apo results are shown in black, holo results are shown in color according to design classification: non-binders are colored orange, high-affinity binders are colored turquoise. Shaded areas represent standard deviation among the five replicates. Representation of high fluctuation areas on the protein structure, with residues colored according to RMSF values of the apo simulations (right panels). Only one of the dimers are shown for simplicity, and ligand's position in the designed model is shown as density in the binding pocket for reference.

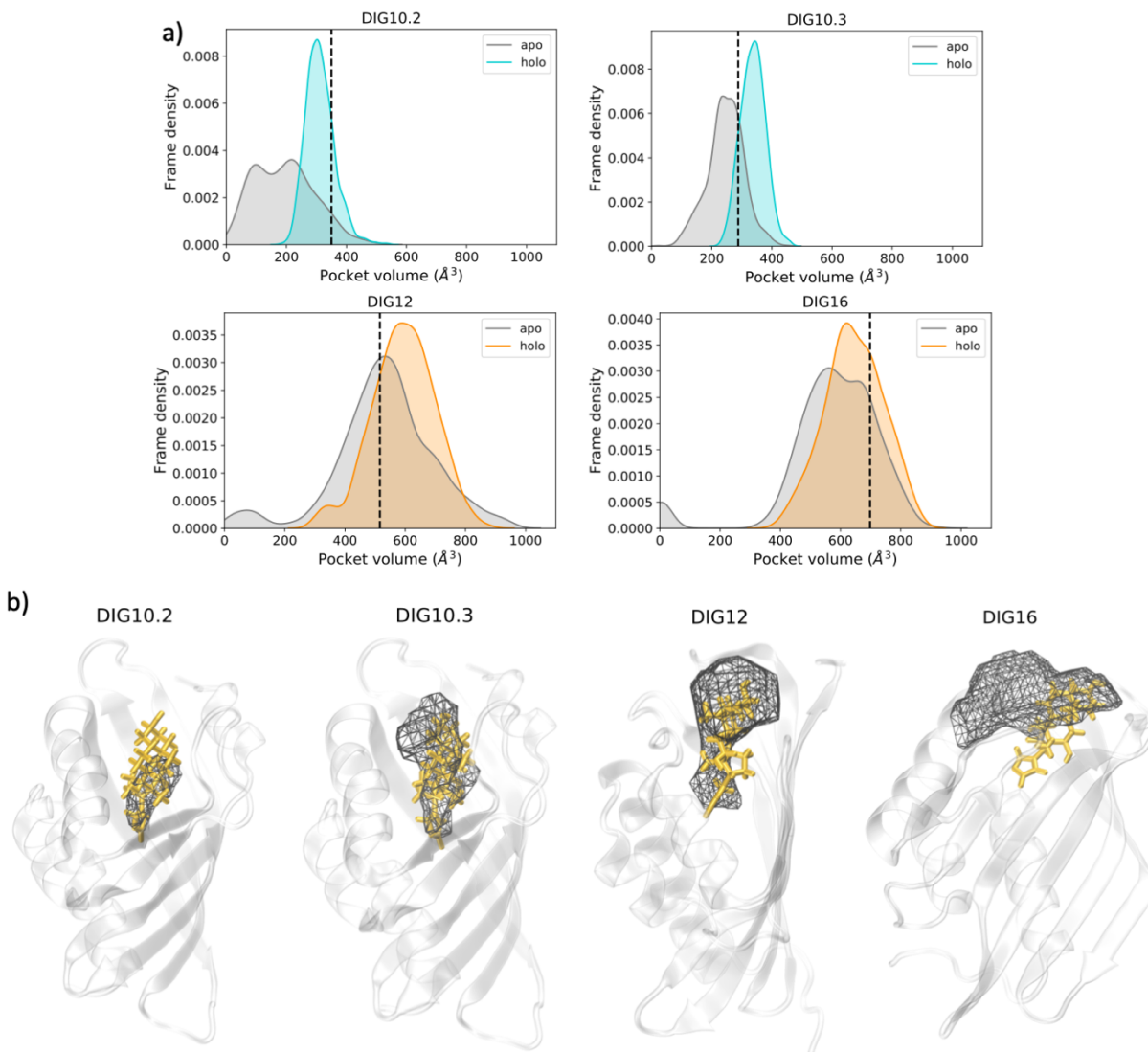


Figure 3.S2. POVME results for DIG designs (a) Distribution of pocket volume. Apo results are shown in gray and holo results are shown in color. The volume from the initial modeled structure is represented by the dashed line. (b) Cavity volume density at 50% of the simulation time represented as a black mesh. Starting protein structure is shown in white and initial ligand orientation shown in gold for reference.

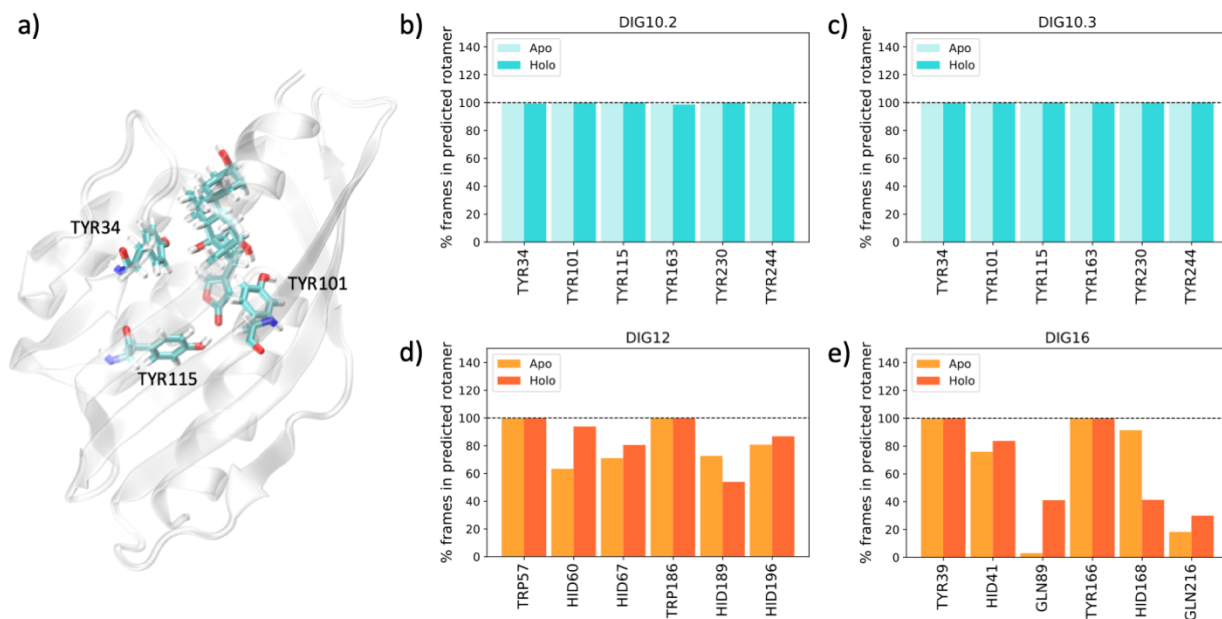


Figure 3.S3. Dihedral angle analysis of ligand-interacting side chains. (a) Representation of the designed hydrogen-bonding interacting residues located in the interior of the protein cavity, shown for the resolved crystal structure of DIG10.3 chain A. The interacting residues in the other designs occupy similar positions in the protein structure. (b) Sampling of side-chain orientation in the rotameric state as the designed structure in terms of χ_1 . High affinity designs are colored in turquoise and non-binders in orange. Results from apo simulations are shown in lighter colors, and holo simulations in darker.

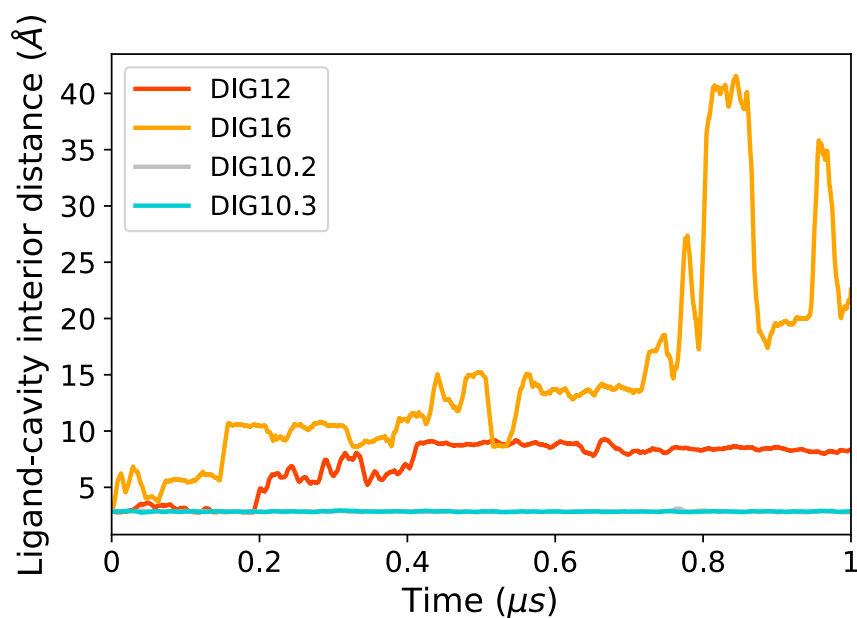


Figure 3.S4. Protein-ligand distance for select replicates. DIG16 shows ligand dissociation.

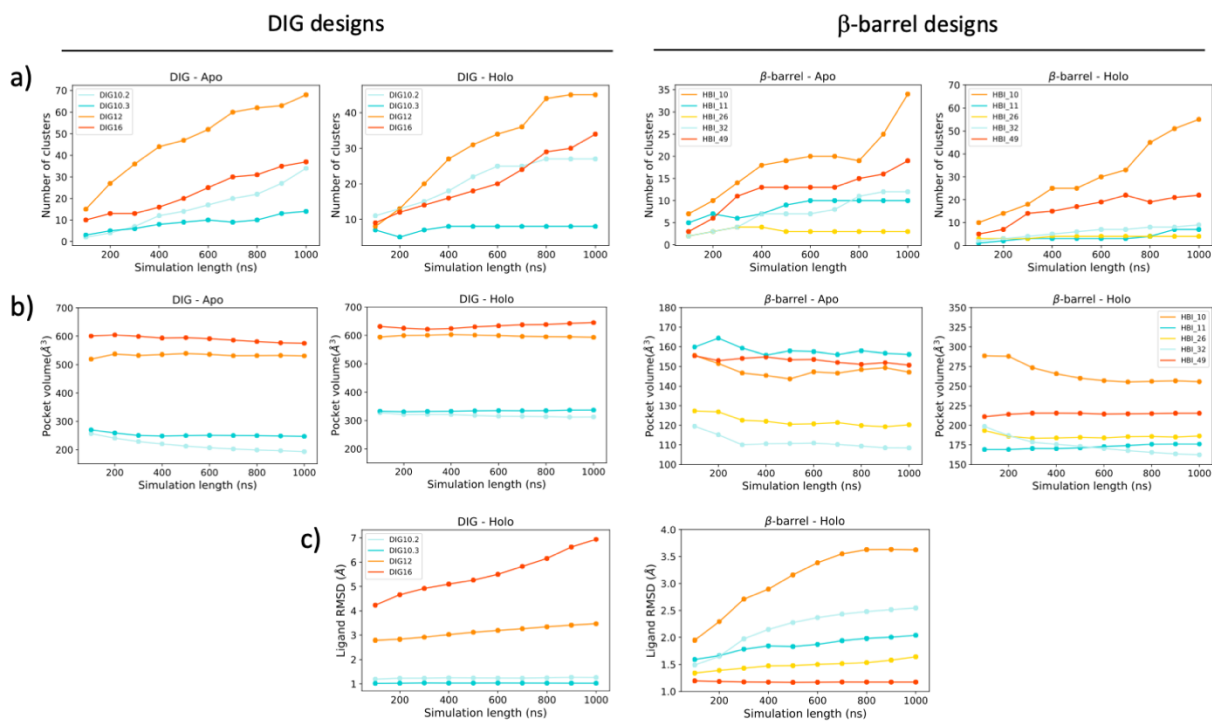


Figure 3.S5. Convergence analysis for the DIG and β -barrel simulations. Values computed from the 5 replicas are shown. (a) Number of clusters (NOC), (b) average pocket volume and (c) average ligand RMSD.

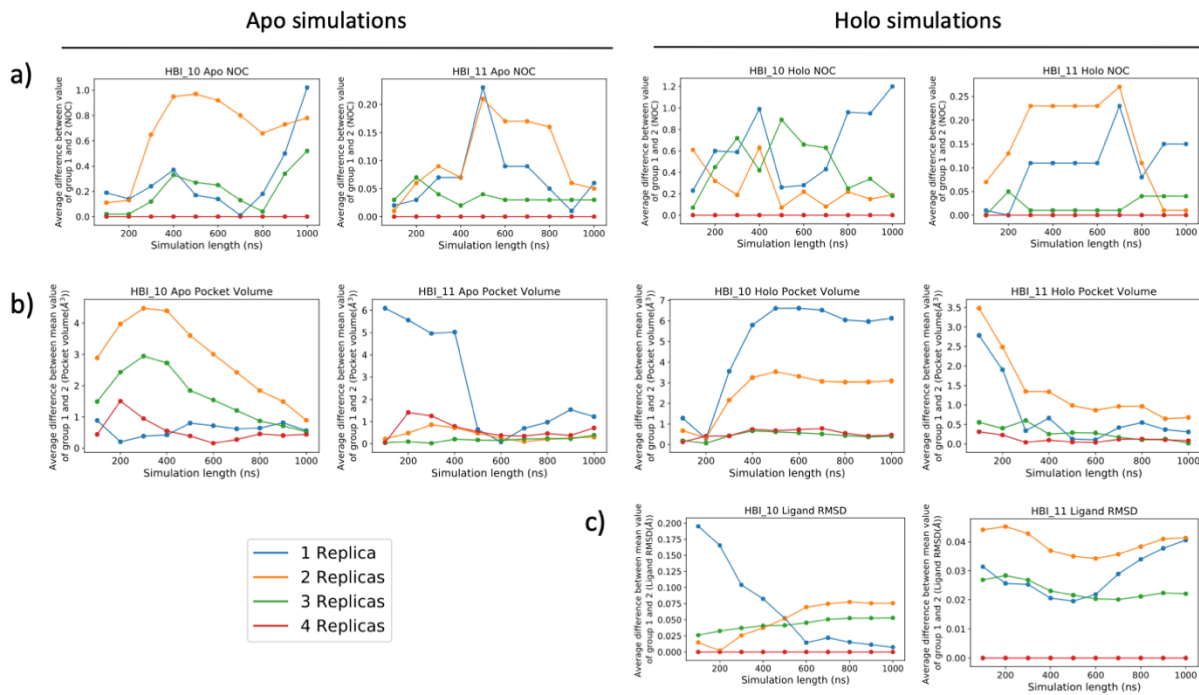


Figure 3.S6. Replicate convergence analysis for the HBI_10 and HBI_11 β -barrel simulations. (a) Number of clusters (NOC), (b) average pocket volume and (c) average ligand RMSD.

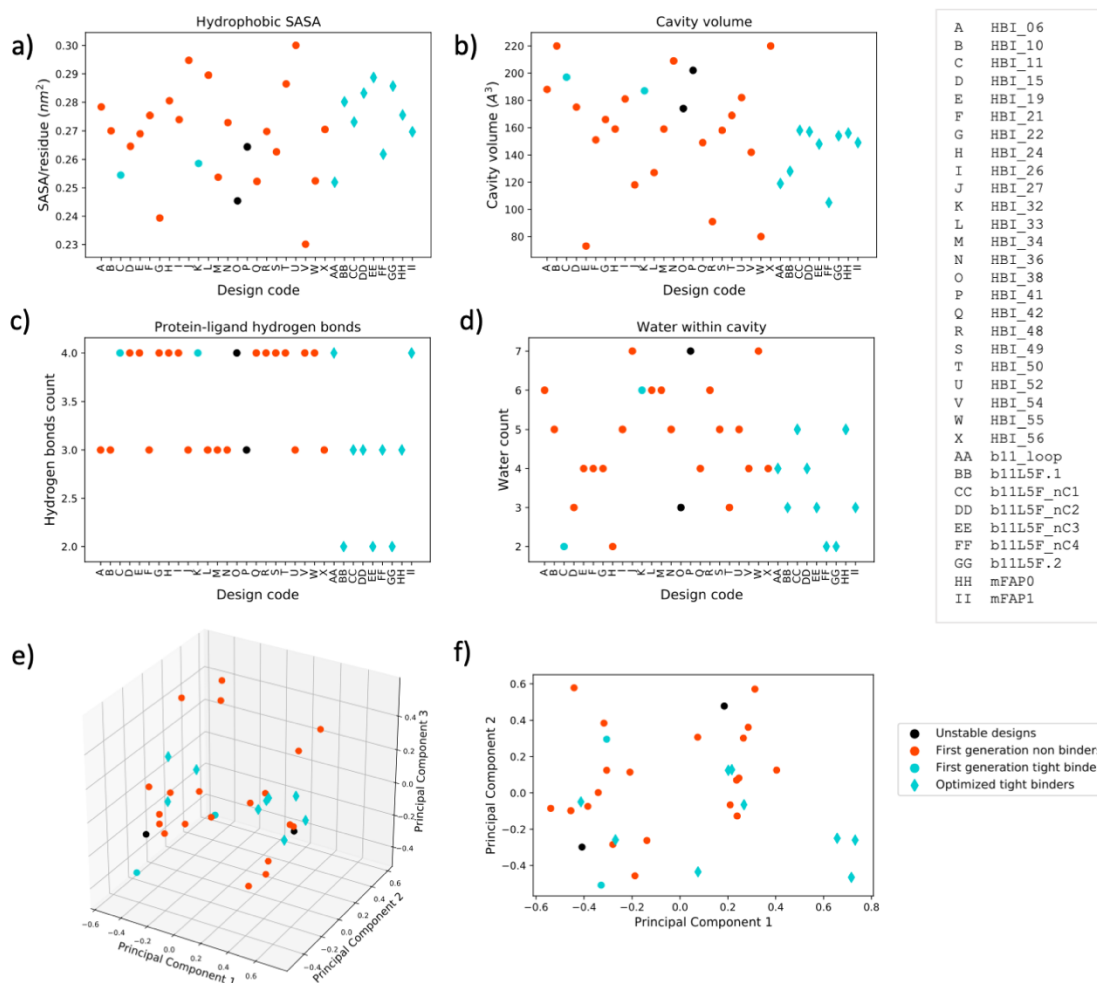


Figure 3.S7. Properties of the static, Rosetta-designed protein structures used as starting conformations for the simulations. (a) Hydrophobic SASA, (b) cavity volume, (c) count of protein-ligand hydrogen bonds and (d) cavity water count after solvation of apo structures. Design name key is provided on the right. (e) Data distribution in terms of the first three principal components and (f) projected on the first two principal components. Non-binders are shown in orange, successful designs are shown in turquoise and structurally unstable designs in black.

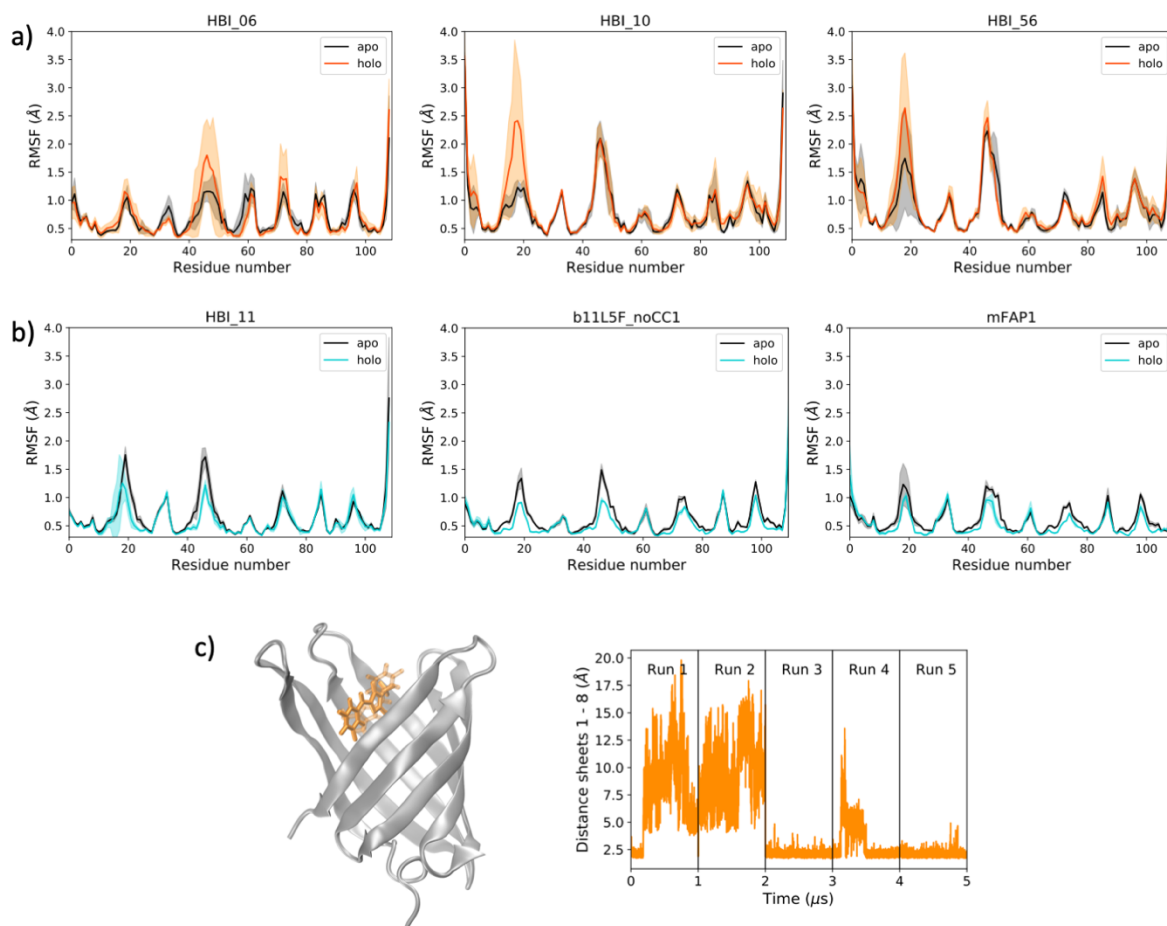


Figure 3.S8. (a) RMSF values for select non-binders and (b) tight binders of the β -barrel design dataset. Apo results are shown in black, while holo results are shown in color according to design classification: non-binders are colored orange, high-affinity binders are colored turquoise. Shaded areas represent standard deviation among the replicates. (c) Example of β -sheet unzipping deformation seen in the HBI_10 simulations. The ligand's initial coordinates are shown in light orange while its final coordinates after the deformation are shown in darker orange. The panel on the right shows the distance between the sheets throughout the five replicates.

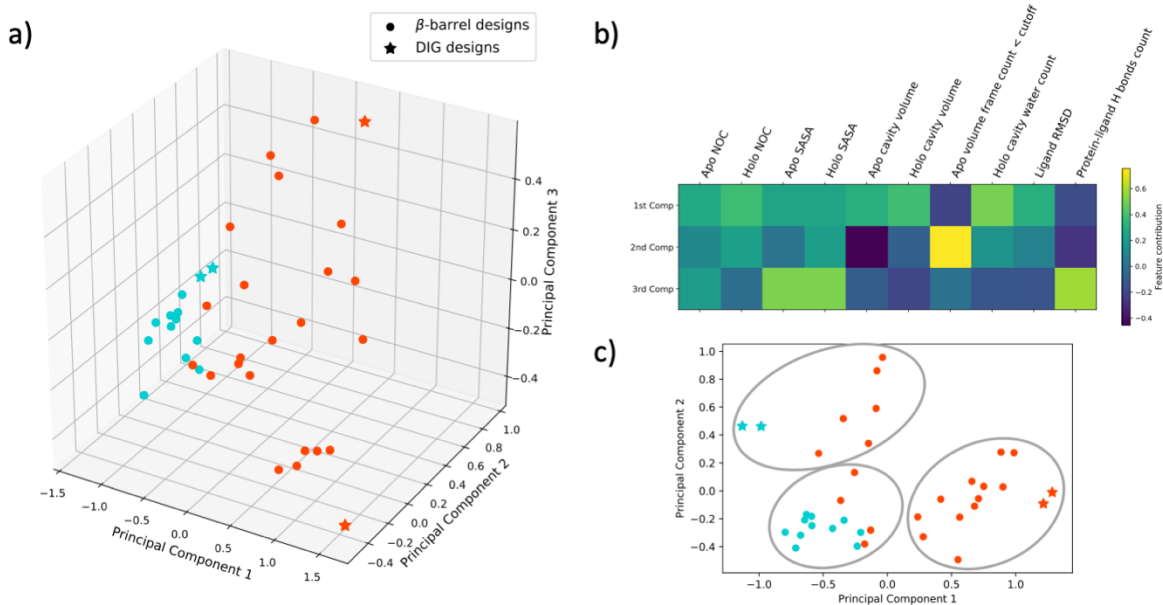


Figure 3.S9. Joint unsupervised classification model for DIG and β -barrel designs. (a) Three-dimensional plot of DIG (shown as stars) and β -barrel simulations (shown as circles) distribution in terms of the principal components of the discriminative features. Non-binders are shown in orange, tight binders in turquoise. (b) Color representation of feature contribution to the principal components. (c) Representation of cluster assignment on the 2-dimensional plot in terms of principal components 1 and 2.

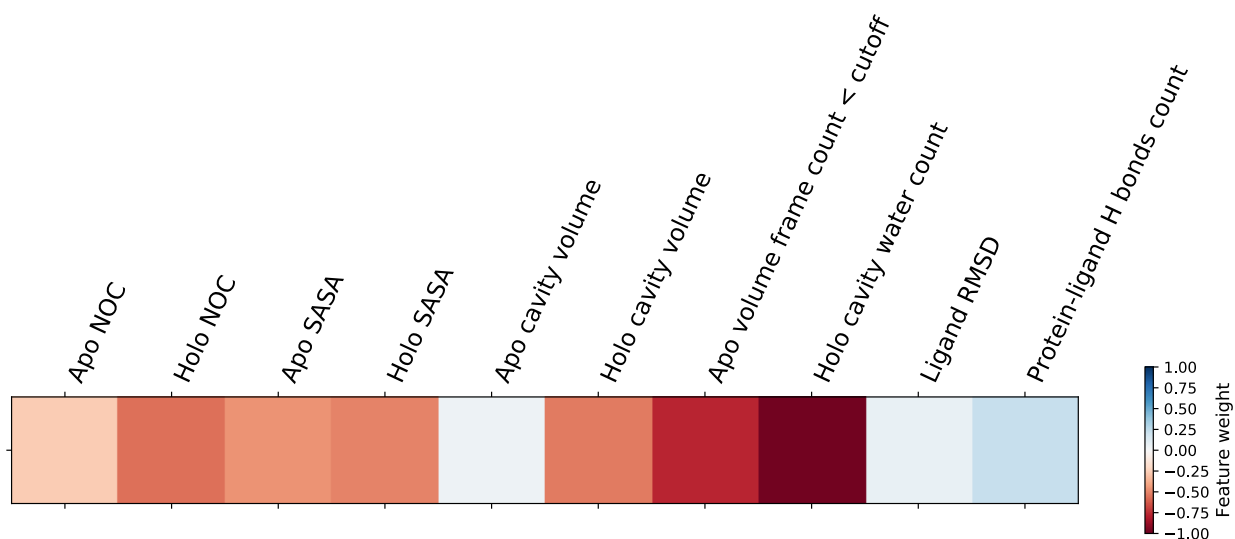


Figure 3.S10. Logistic regression feature weights from a model trained on 80% of the DIG and β -barrel design data.

Table 3.S1. Evaluation of the supervised learning classifiers, using 33% or 50% of the data in the training set.

Validation metric	Logistic regression		k-nearest neighbors (k=5)	
	training = 33%	training = 50%	training = 33%	training = 50%
Accuracy	0.84 ± 0.09	0.85 ± 0.08	0.73 ± 0.11	0.81 ± 0.11
Precision	0.72 ± 0.13	0.73 ± 0.15	0.58 ± 0.24	0.68 ± 0.17
Recall	0.94 ± 0.14	0.97 ± 0.09	0.79 ± 0.33	0.96 ± 0.11
MCC	0.70 ± 0.17	0.74 ± 0.12	0.52 ± 0.22	0.68 ± 0.15
F1S	0.80 ± 0.11	0.82 ± 0.09	0.63 ± 0.24	0.78 ± 0.13

^a Values correspond to average and standard deviation of 100 rounds of random splits of the data set according to the training/test set membership ratio (33% or 50% of the data in the training set).

3.8 References

- (1) Huang, P.; Boyken, S. E.; Baker, D. The Coming of Age of de Novo Protein Design. *Nature* **2016**, *537*, 320–327. <https://doi.org/10.1038/nature19946>.
- (2) Woolfson, D. N.; Bartlett, G. J.; Burton, A. J.; Heal, J. W.; Niitsu, A.; Thomson, A. R.; Wood, C. W. De Novo Protein Design: How Do We Expand into the Universe of Possible Protein Structures? *Curr. Opin. Struct. Biol.* **2015**, *33*, 16–26.
- (3) Huang, P.-S.; Feldmeier, K.; Parmeggiani, F.; Fernandez Velasco, D. A.; Höcker, B.; Baker, D. De Novo Design of a Four-Fold Symmetric TIM-Barrel Protein with Atomic-Level Accuracy. *Nat. Chem. Biol.* **2016**, *12* (November), 29–34. <https://doi.org/10.1038/nchembio.1966>.
- (4) Lin, Y.-R.; Koga, N.; Tatsumi-Koga, R.; Liu, G.; Clouser, A. F.; Montelione, G. T.; Baker, D. Control over Overall Shape and Size in de Novo Designed Proteins. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E5478–E5485. <https://doi.org/10.1073/pnas.1509508112>.
- (5) Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.; Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard, B. L.; et al. Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity. *Nature* **2013**, *501* (7466), 212–216. <https://doi.org/10.1038/nature12443>.

- (6) Thomas, F.; Dawson, W. M.; Lang, E. J. M.; Burton, A. J.; Bartlett, G. J.; Rhys, G. G.; Mulholland, A. J.; Woolfson, D. N. De Novo-Designed A-helical Barrels as Receptors for Small Molecules. *ACS Synth. Biol.* **2018**, *7*, 1808–1816. <https://doi.org/10.1021/acssynbio.8b00225>.
- (7) Dou, J.; Vorobieva, A. A.; Sheffler, W.; Doyle, L. A.; Park, H.; Bick, M. J.; Lee, M. Y.; Gagnon, L. A.; Carter, L.; Sankaran, B.; et al. De Novo Design of a Fluorescence-Activating β -Barrel. *Nature* **2018**, *561*, 485–491. <https://doi.org/10.1038/s41586-018-0509-0>.
- (8) Bjelic, S.; Nivon, L. G.; Celebi-Olcum, N.; Kiss, G.; Rosewall, C. F.; Lovick, H. M.; Ingalls, E. L.; Gallaher, J. L.; Seetharaman, J.; Lew, S.; et al. Computational Design of Enone-Binding Proteins with Catalytic Activity for the Morita-Baylis-Hillman Reaction. *ACS Chem. Biol.* **2013**, *8*, 749–757. <https://doi.org/10.1021/cb3006227>.
- (9) Burton, A. J.; Thomson, A. R.; Dawson, W. M.; Brady, R. L.; Woolfson, D. N. Installing Hydrolytic Activity into a Completely de Novo Protein Framework. *Nat. Chem.* **2016**, *8*, 837–844. <https://doi.org/10.1038/nchem.2555>.
- (10) Strauch, E.-M.; Bernard, S. M.; La, D.; Bohn, A. J.; Lee, P. S.; Anderson, C. E.; Nieuwma, T.; Holstein, C. A.; Garcia, N. K.; Hooper, K. A.; et al. Computational Design of Trimeric Influenza-Neutralizing Proteins Targeting the Hemagglutinin Receptor Binding Site. *Nat. Biotechnol.* **2017**, *35*, 667–671. <https://doi.org/10.1038/nbt.3907>.
- (11) Chevalier, A.; Silva, D.; Rocklin, G. J.; Hicks, D. R.; Vergara, R.; Murapa, P.; Bernard, S. M.; Zhang, L.; Lam, K.; Yao, G.; et al. Massively Parallel de Novo Protein Design for Targeted Therapeutics. *Nature* **2017**, *550*, 74–79. <https://doi.org/10.1038/nature23912>.
- (12) Koday, M. T.; Nelson, J.; Chevalier, A.; Koday, M.; Kalinoski, H.; Stewart, L.; Carter, L.; Nieuwma, T.; Lee, P. S.; Ward, A. B.; et al. A Computationally Designed Hemagglutinin Stem-Binding Protein Provides in Vivo Protection from Influenza Independent of a Host Immune Response. *PLOS Pathog* **2016**, *12*, e1005409. <https://doi.org/10.1371/journal.ppat.1005409>.
- (13) King, N. P.; Sheffler, W.; Sawaya, M. R.; Vollmar, B. S.; Sumida, J. P.; André, I.; Gonen, T.; Yeates, T. O.; Baker, D. Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science (80-.)*. **2012**, *336*, 1171–1174.
- (14) Fletcher, J. M.; Harniman, R. L.; Barnes, F. R. H.; Boyle, A. L.; Collins, A.; Mantell, J.; Sharp, T. H.; Antognozzi, M.; Booth, P. J.; Linden, N.; et al. Self-Assembling Cages from Coiled-Coil Peptide Modules. *Science (80-.)*. **2013**, *340*, 595–599.
- (15) Hsia, Y.; Bale, J. B.; Gonen, S.; Shi, D.; Sheffler, W.; Fong, K. K.; Nattermann, U.; Xu, C.; Huang, P.-S.; Ravichandran, R.; et al. Design of a Hyperstable 60-Subunit Protein Icosahedron. *Nature* **2016**, *535*, 136–139. <https://doi.org/10.1038/nature18010>.
- (16) Bale, J. B.; Gonen, S.; Liu, Y.; Sheffler, W.; Ellis, D.; Thomas, C.; Cascio, D.; Yeates, T. O.; Gonen, T.; King, N. P.; et al. Accurate Design of a Megadalton-Scale Two-Component

- Icosahedral Protein Complexes. *Science* (80-.). **2016**, 353, 389–394. <https://doi.org/10.5061/dryad.8c65s>.
- (17) Regan, L.; Caballero, D.; Hinrichsen, M. R.; Virrueta, A.; Williams, D. M.; Hern, C. S. O. Protein Design: Past, Present, and Future. *Pept. Sci.* **2015**, 104 (4), 334–350. <https://doi.org/10.1002/bip.22639>.
- (18) Martin, S. F.; Clements, J. H. Correlating Structure and Energetics in Protein-Ligand Interactions: Paradigms and Paradoxes. *Annu. Rev. Biochem.* **2013**, 82, 267–293. <https://doi.org/10.1146/annurev-biochem-060410-105819>.
- (19) Zanghellini, A. De Novo Computational Enzyme Design. *Curr. Opin. Biotechnol.* **2014**, 29, 132–138. <https://doi.org/10.1016/j.copbio.2014.03.002>.
- (20) Feng, J.; Jester, B. W.; Tinberg, C. E.; Mandell, D. J.; Antunes, M. S.; Chari, R.; Morey, K. J.; Rios, X.; Medford, J. I.; Church, G. M.; et al. A General Strategy to Construct Small Molecule Biosensors in Eukaryotes. *Elife* **2015**, 4 (2015), e10606. <https://doi.org/10.7554/eLife.10606>.
- (21) Bick, M. J.; Greisen, P. J.; Morey, K. J.; Antunes, M. S.; La, D.; Sankaran, B.; Reymond, L.; Johnsson, K.; Medford, J. I.; Baker, D. Computational Design of Environmental Sensors for the Potent Opioid Fentanyl. *Elife* **2017**, 6, e28909.
- (22) Roy, A.; Nair, S.; Sen, N.; Soni, N.; Madhusudhan, M. S. In Silico Methods for Design of Biological Therapeutics. *Methods* **2017**, 131, 33–65. <https://doi.org/10.1016/j.ymeth.2017.09.008>.
- (23) Entzminger, K. C.; Hyun, J.; Pantazes, R. J.; Patterson-Orazem, A. C.; Qerqez, A. N.; Frye, Z. P.; Hughes, R. A.; Ellington, A. D.; Liebermanl, R. L.; Maranas, C. D.; et al. De Novo Design of Antibody Complementarity Determining Regions Binding a FLAG Tetrapeptide. *Sci. Rep.* **2017**, No. 7, 10295. <https://doi.org/10.1038/s41598-017-10737-9>.
- (24) Mondal, J.; Friesner, R. A.; Berne, B. J. Role of Desolvation in Thermodynamics and Kinetics of Ligand Binding to a Kinase. *J. Chem. Theory Comput.* **2014**, 10, 5696–5705.
- (25) Dou, J.; Doyle, L.; Greisen, P. J.; Schena, A.; Park, H.; Johnsson, K.; Stoddard, B. L.; Baker, D. Sampling and Energy Evaluation Challenges in Ligand Binding Protein Design. *Protein Sci.* **2017**, 26, 2426–2437. <https://doi.org/10.1002/pro.3317>.
- (26) Kiss, G.; Pande, V. S.; Houk, K. N. Molecular Dynamics Simulations for the Ranking, Evaluation, and Refinement of Computationally Designed Proteins. In *Methods Enzymol.*; 2013; Vol. 523, pp 145–170.
- (27) Childers, M. C.; Daggett, V. Insights from Molecular Dynamics Simulations for Computational Protein Design. *Mol. Syst. Des. Eng.* **2017**, 2, 9–33. <https://doi.org/10.1039/C6ME00083E>.

- (28) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative Approach to Computational Enzyme Design. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 3790–3795. <https://doi.org/10.1073/pnas.1118082108>.
- (29) Lindert, S.; Mccammon, J. A. Improved CryoEM-Guided Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction. *J. Chem. Theory Comput.* **2015**, *11*, 1337–1346. <https://doi.org/10.1021/ct500995d>.
- (30) Leelananda, S. P.; Lindert, S. Iterative Molecular Dynamics-Rosetta Membrane Protein Structure Refinement Guided by Cryo-EM Densities. *J. Chem. Theory Comput.* **2017**, *13*, 5131–5145. <https://doi.org/10.1021/acs.jctc.7b00464>.
- (31) Yu, H.; Huang, H. Engineering Proteins for Thermostability through Rigidifying Flexible Sites. *Biotechnol. Adv.* **2014**, *32*, 308–315.
- (32) Joo, J. C.; Pack, S. P.; Kim, Y. H.; Yoo, Y. J. Thermostabilization of Bacillus Circulans Xylanase: Computational Optimization of Unstable Residues Based on Thermal Fluctuation Analysis. *J. Biotechnol.* **2011**, *151*, 56–65. <https://doi.org/10.1016/j.jbiotec.2010.10.002>.
- (33) Liu, J.; Yu, H.; Shen, Z. Insights into Thermal Stability of Thermophilic Nitrile Hydratases by Molecular Dynamics Simulation. *J. Mol. Graph. Model.* **2008**, *27*, 529–535. <https://doi.org/10.1016/j.jmglm.2008.09.004>.
- (34) Chen, J.; Yu, H.; Liu, C.; Liu, J.; Shen, Z. Improving Stability of Nitrile Hydratase by Bridging the Salt-Bridges in Specific Thermal-Sensitive Regions. *J. Biotechnol.* **2012**, *164*, 354–362.
- (35) Kiss, G.; Rothlisberger, D.; Baker, D.; Houk, K. N. Evaluation and Ranking of Enzyme Designs. *Protein Sci.* **2010**, *19*, 1760–1773. <https://doi.org/10.1002/pro.462>.
- (36) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A.; Friesner, R. A. A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins* **2004**, *55*, 351–367. <https://doi.org/10.1002/prot.10613>.
- (37) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations. *J. Mol. Biol.* **2002**, *320*, 597–608. [https://doi.org/10.1016/S0022-2836\(02\)00470-9](https://doi.org/10.1016/S0022-2836(02)00470-9).
- (38) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (39) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J Mol Graph Model* **2006**, *25*, 247–260. <https://doi.org/10.1016/j.jmglm.2005.12.005>.

- (40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. Gaussian 09, Revision C.01. Gaussian, Inc.: Wallingford CT 2010.
- (41) Gumbart, J.; Trabuco, L. G.; Schreiner, E.; Villa, E.; Schulten, K. Regulation of the Protein-Conducting Channel by a Bound Ribosome. *Structure* **2009**, *17*, 1453–1464. <https://doi.org/10.1016/j.str.2009.09.010>.
- (42) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; T.E. Cheatham, I.; Darden, T. A.; Duke, R. E.; Gohlke, H.; et al. AMBER 14. University of California, San Francisco 2014.
- (43) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713. <https://doi.org/10.1021/acs.jctc.5b00255>.
- (44) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys* **1983**, *79*, 926–932.
- (45) Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B. Constant-PH Molecular Dynamics Simulations for Large Biomolecular Systems. *J. Chem. Theory Comput.* **2017**, *13*, 5933–5944. <https://doi.org/10.1021/acs.jctc.7b00875>.
- (46) Purawat, S.; Jeong, P. U.; Malmstrom, R. D.; Chan, G. J.; Yeung, A. K.; Walker, R. C.; Altintas, I.; Amaro, R. E. A Kepler Workflow Tool for Reproducible AMBER GPU Molecular Dynamics. *Biophys. J.* **2017**, *112*, 2469–2474. <https://doi.org/10.1016/j.bpj.2017.04.055>.
- (47) Salomon-Ferrer, R.; Go, A. W.; Poole, D.; Grand, S. Le; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888. <https://doi.org/10.1021/ct400314y>.
- (48) Darden, T. A.; York, D.; Pedersen, L. Particle-Mesh Ewald: An N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (49) Humphery, W.; Dalke, A.; Schulten, K. VMD-Visual Molecular Dynamics. *J. Molec. Graph.* **1996**, *14*, 33–38.
- (50) Kluyver, T.; Ragan-kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; et al. Jupyter Notebooks - a Publishing Format for Reproducible Computational Workflows. *Position. Power Acad. Publ. Play. Agents Agendas* **2016**, 87–90.
- (51) Mcgibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open

- Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532. <https://doi.org/10.1016/j.bpj.2015.08.015>.
- (52) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095. <https://doi.org/10.1021/ct400341p>.
- (53) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noe, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542. <https://doi.org/10.1021/acs.jctc.5b00743>.
- (54) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAAnalysis : A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem. Chem.* **2011**, *32*, 2319–2327. <https://doi.org/10.1002/jcc>.
- (55) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; et al. MDAAnalysis : A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proc. of the 15th Python in Science Conf.*; 2016; pp 98–105.
- (56) Durrant, J. D.; Votapka, L.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory Comput.* **2014**, *10*, 5047–5056.
- (57) Steiner, T. The Hydrogen Bond in the Solid State. *Angew. Chem. Int. Ed.* **2002**, *41*, 48–76.
- (58) Liu, P.; Harder, E.; Berne, B. J. On the Calculation of Diffusion Coefficients in Confined Fluids and Interfaces with an Application to the Liquid - Vapor Interface of Water. *J. Phys. Chem. B* **2004**, *108*, 6595–6602.
- (59) Knapp, B.; Ospina, L.; Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulations: The Importance of Replicas. *J. Chem. Theory Comput.* **2018**, *14*, 6127–6138. <https://doi.org/10.1021/acs.jctc.8b00391>.
- (60) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: ‘What You See’ Is Not Always ‘What You Get.’” *Structure* **2009**, *17*, 489–498. <https://doi.org/10.1016/j.str.2009.02.010>.
- (61) Demir, Ö.; Baronio, R.; Salehi, F.; Wassman, C. D.; Hall, L.; Hatfield, G. W.; Chamberlin, R.; Lathrop, R. H.; Amaro, R. E. Ensemble-Based Computational Approach Discriminates Functional Activity of P53 Cancer and Rescue Mutants. *PLoS Comput. Biol.* **2011**, *7*, e1002238. <https://doi.org/10.1371/journal.pcbi.1002238>.
- (62) Domingos, P. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **2012**, *55*, 78–87.
- (63) Chicco, D. Ten Quick Tips for Machine Learning in Computational Biology. *BioData Min.* **2017**, *10*, 1–17. <https://doi.org/10.1186/s13040-017-0155-3>.

- (64) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data to Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726–741. <https://doi.org/10.1021/acs.jcim.6b00778>.

On the conformational diversity within protein crystals

Emilia P. Barros¹, David Wych², Michael E. Wall³, David Mobley^{2,4}, Rommie E. Amaro^{1,5}

Author affiliations:

1 – Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA

2 – Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA, USA

3 – Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, USA

4 – Department of Chemistry, University of California, Irvine, Irvine, CA, USA

5 – National Biomedical Computation Resource, University of California, San Diego, La Jolla, CA, USA

4.1 Abstract

Even in the crystal lattice, proteins retain a significant degree of flexibility and can adopt multiple conformations. Diffuse scattering is an experimental technique that accounts for half of the signal captured in X-ray crystallography experiments and that carry information on the correlated motions in crystals. However, they are typically ignored in the elucidation of macromolecular structures, due to challenges in the acquisition of high quality signal compared to the stronger Bragg peaks and interpretation of the results in terms of intra- and intermolecular motions in the crystal. As improvements to the data acquisition promise to make diffuse scattering more accessible in biophysical experiments, attention has been turned to the validation of models for interpretation of protein dynamics in crystals. Here, we characterize the extent of conformational variability in experimentally-validated molecular dynamics simulations of a supercell of *staphylococcal* nuclease using the Markov state model (MSM) methodology. Our results evidence not only the degree of protein flexibility and absence of symmetry across the unit cells but also the existence of significant chain cross-correlation effects and long-range communication in these crowded environments. This work sets the stage for applications of MSMs in the interpretation of correlated motions in protein crystals probed by diffuse scattering experiments and suggests the use of this often overlooked experimental technique in the validation of MSM parameters.

4.2 Introduction

Diffuse scattering, the streaked and cloudy features present in diffraction patterns, are a result of imperfections in crystal structures and correlated motions between atoms¹. These patterns constitute half of the data collected in X-ray experiments but are traditionally ignored in favor of

the stronger Bragg peaks, which report on the mean electron density and result in an averaged structure model². In the past few decades, however, increased efforts in the acquisition of high-resolution diffuse scattering signal and interpretation of the results in terms of protein dynamics models placed diffuse scattering as a unique and newly accessible biophysical probe for reporting on atomic spatial correlation on structure and dynamics of macromolecules³.

While the relationship between diffuse scattering and correlated motions is well established, the interpretation of the scatter results in terms of the actual dynamics in the crystal remains challenging. Several models have been proposed^{4–10}, and more recently molecular dynamics simulations of crystalline proteins have been used to enrich the interpretation of the results, by allowing the calculation of diffuse scattering profiles from the protein models and comparison with the experimental results^{2,11–16}. In addition to providing information on the protein dynamics and correlations at the atomic level, these two methodologies can also complement each other in another way: validation of MD simulations by experimentally-characterized diffuse scattering maps have been suggested as a way to improve the development of force fields and models of protein dynamics^{1,17} (Figure 4.1). Some key limitations in the use of MD simulations for this purpose include adequate sampling of these large systems¹⁸ and especially accurate modeling of the anisotropic component of the diffuse scatter, related to correlated motions in the protein components of the crystal².

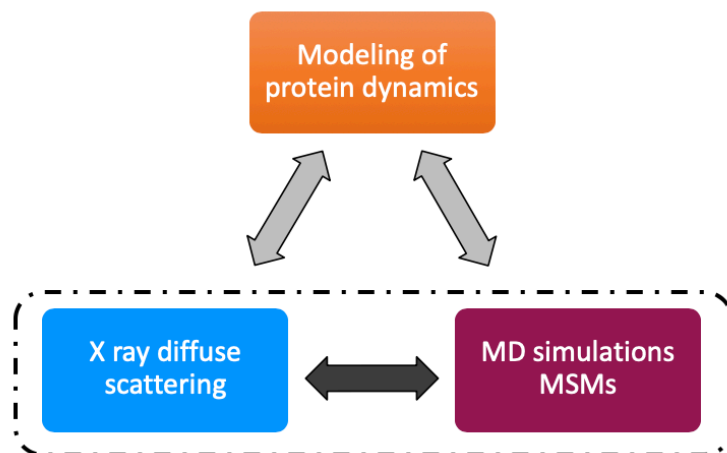


Figure 4.1. Representation of the reciprocal relationship between diffuse scattering and MD simulations in modeling protein motions.

In an effort to aid the interpretability of the diffuse scattering signal, we have developed Markov state models (MSMs) of a long-timescale *staphylococcal* nuclease 2x2x2 supercell simulation conducted by Wall et al², represented in Figure 4.2. This configures the first application of MSMs to diffuse scattering experiments and crystalline systems, and is motivated by the mathematically-rigorous discretization of protein conformational dynamics, coupled with thermodynamic and kinetics information, that is provided by the analysis of MD simulations in the MSM framework¹⁹⁻²⁴. Moreover, since the crystalline simulations include multiple copies of the protein, the use of MSMs to unify the conformational ensemble explored by all copies is a natural follow up step in the interpretation of the protein conformational dynamics observed in the crystalline model. Additionally, taking advantage of the reciprocal relationship between MD simulations and diffuse scatter, we also propose here the application of diffuse scattering as an experimental observable to optimize Markov state models, by providing external validation for the many parameters and features that need to be decided upon for model construction.

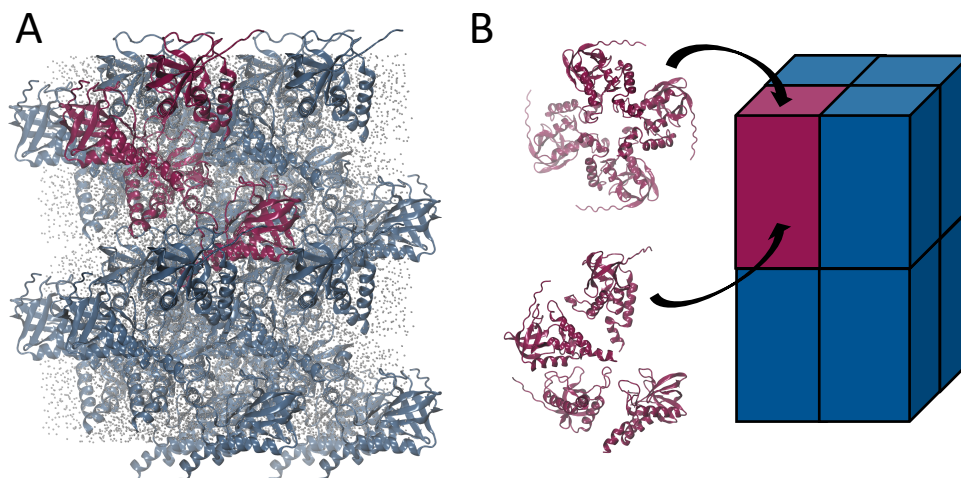


Figure 4.2. Representation of the staphylococcal nuclease supercell model used in this study.

Despite constructing models based on individual proteins internal motions, the observed metastable states indicate the extent of intermolecular interactions enabled by each chain's intrinsic dynamics. Inter-chain correlations are observed at large distances beyond the confines of the unit cell, and add to the evidence on the importance of long range motions in diffuse scattering. The application to crystalline systems is a substantial innovation in MSM MD simulations, and suggests the use of diffuse scattering as observables for the experimental validation of MSMs.

4.3 Methods

Simulations and Markov State Models

Simulations were taken from an experimentally-validated crystalline 2x2x2 supercell of staphylococcal nuclease, containing a total of 32 protein chains and explicit water molecules². The protein structure was taken from PDB 1SNC, and missing N and C termini residues modelled using *UCSF Chimera*. Details of the system preparation and simulation can be found in M. E. Wall, *IUCrJ* (2018). The supercell was simulated for a total of 5 μ s of production.

For MSM construction, the coordinates of each chain were extracted from the supercell simulation and saved as individual protein trajectories. The ensemble of 32 individual 5 μ s trajectories was processed and models built using PyEMMA²⁵, version 3.5.6. The flexible regions model (details on the selection of features provided in the Results section) was built using time-independent component analysis (tICA)²⁶ with a lag time of 1 ns and MSM lag time of 20 ns. Discretization was performed with k-means clustering, $k = 333$. Model accuracy was verified by implied timescale (ITS) plots and Chapman-Kolmogorov tests (Supplementary Figures 4.S1 and 4.S2).

Additional models were constructed using different feature selections. The active site MSM, using all combinations of pairwise distances between the active site residues (Arg35, Glu43, Tyr85, Arg87, Tyr113 and Tyr115), was constructed with $k = 376$ and tICA and MSM lag times of 5 ns and 30 ns, respectively. An unit cell MSM used internal features from the initial model in addition to all pairwise distances between Ile92 in each chain in the unit cell, $k = 400$ and tICA and MSM lag times of 5 ns. The MSM lag time was chosen based on the respective model ITS plot.

State cross-correlation

Chain cross-correlations in terms of the state distribution as defined by the MSM metastable states during the simulation was calculated using Pearson correlation between each pair of chains. Correlation dependence on inter-chain distances was verified through the calculation of the symmetry-corrected distances between each pair of chains in the supercell.

Calculation of diffuse scattering

For calculation of the diffuse scattering produced from the metastable state conformations identified in the MSMs, snapshots of the supercell were reconstructed by randomly placing protein

conformations extracted from the metastable states, with the overall state distribution in a single snapshot following the MSM state's equilibrium population. A total of 10,000 frames were constructed for diffuse scattering calculations. Simulated diffuse scattering was calculated following the procedure outlined in M. E. Wall, IUCrJ (2018) using the “Lunus” diffuse scattering data processing software suite (<https://github.com/mewall/lunus>).

4.4 Results and Discussion

Metastable description of protein conformational flexibility

The *staphylococcal* nuclease was chosen as our model system as it constitutes one of the few crystalline protein systems for which high-resolution diffuse scattering has been obtained and that has been extensively studied in the context of MD crystalline simulations^{2,13,27}. The supercell simulation used as starting point for the analysis was experimentally validated against X-ray B factors and diffuse scattering², suggesting a sufficient degree of accuracy in the modeling of the protein ensemble by the simulation.

RMSF analysis of the protein chains during the simulation evidences that even in these crystalline environments, the protein shows some degree of flexibility, particularly in the termini and in its central loop located close to the active site, residues 42 to 54 (Supplementary Figure 4.S3). To characterize the global conformational flexibility of the protein chains, we selected a number of pairs distributed along the protein structure and computed their pairwise distances during the simulations (Supplementary Table 4.S1). The pairs' influence on the description of slow transitions was analyzed using time-lagged Independent Components Analysis (tICA)²⁶, which identifies the linear combination of features that describes the slowest degrees of freedom of the system. After an iterative process of eliminating features with low tICA coordinate correlations

we arrived at a final set of 10 feature pairs, among which one pair is located in the N terminal, two in the C terminal and the remaining 7 involve one residue in the above identified central flexible loop and another in nearby secondary structure motifs (Figure 4.3a, pairs highlighted in Supplementary Table 4.S1). Their identification in this tICA-directed procedure suggests that these motifs are involved in the slowest motions in the protein system. The feature contributions are distributed among the tICA components, such that this novel coordinate space is a combination of each of these regions' influence on the dynamics (Figure 4.3b and Supplementary Figure 4.S4).

For a more human-interpretable description of the conformational space sampled by the chains during the simulation, we constructed Markov state models on the coordinate space defined by these 10 pairs. Six interconnected metastable states are identified (Figure 4.3c). Structures randomly-selected from these states are shown in Figure 4d and evidence the degree of flexibility of the N and C termini, as well as the flexible loop. The other loops and all secondary structure elements are more dynamically restricted, in accordance with NMR studies of other protein crystals²⁸.

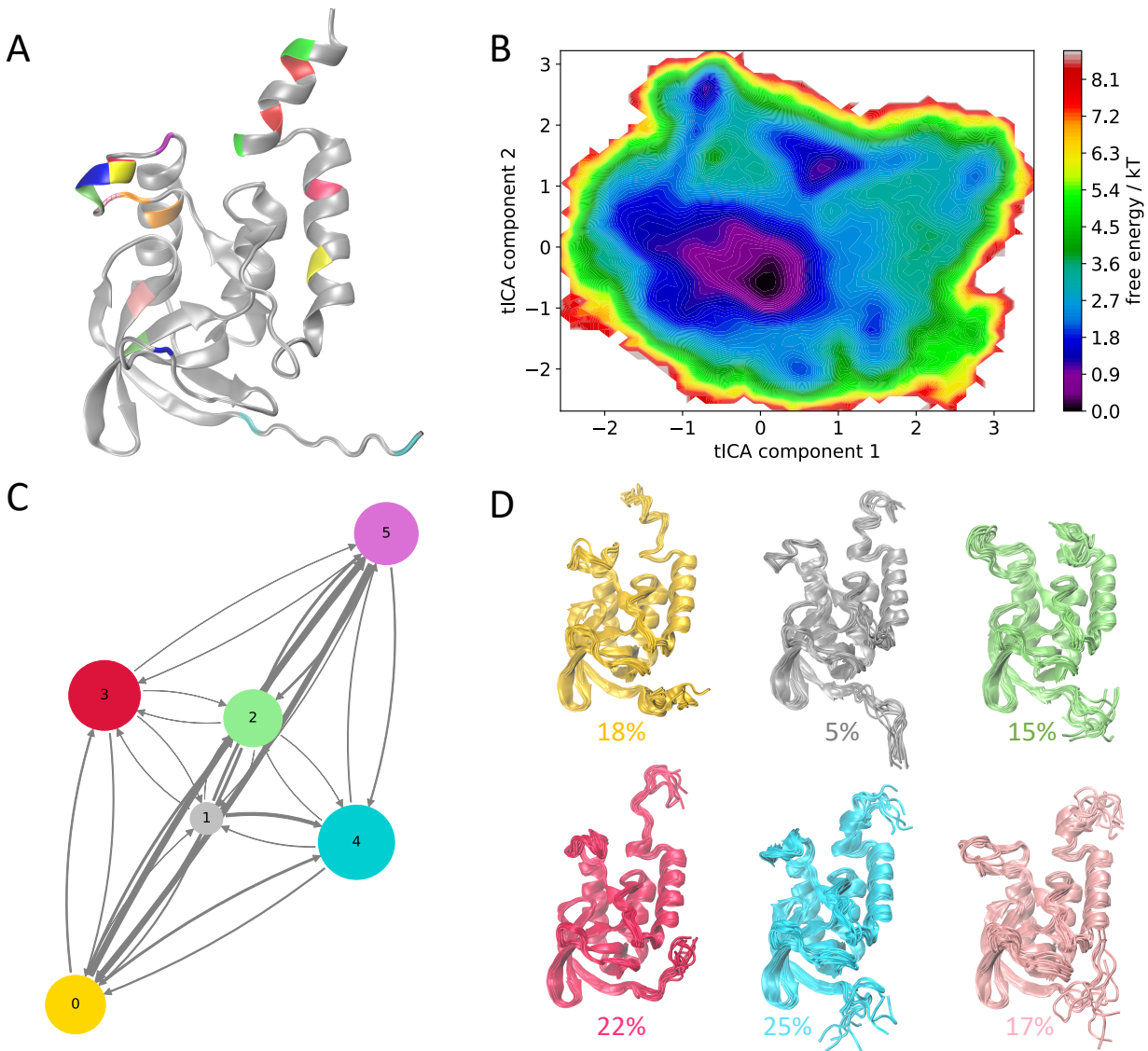


Figure 4.3. Flexible regions MSM. (a) Pairs used as features for MSM construction (in the form of pairwise C α distances). Each pair's location on the protein secondary structure are indicated by a different color. (b) Free energy landscape in terms of the tICA-transformed feature space. (c) Metastable states identified by Hidden Markov models. (d) Representation of 10 randomly-selected conformations from each metastable state shown in (c).

Figure 5 shows these flexible motifs conformations in more detail. The N terminal, being a flexible tail, shows dramatically different conformations among the metastable states, suggesting the existence of not only intra-chain interactions with helix 3 (in states 3 and 5) but also possible inter-chain interactions when in the downwards-extended conformations observed in states 1 and

4 with chains that were not originally among their interacting partners in the initial supercell structure. State 4 is predicted to be the most populated state in equilibrium, accounting for 28% of the equilibrium population, suggesting the weight that these unexpected inter-chain interactions may have in the crowded environments of the protein crystal.

The C terminal contains a small helix that partially unwinds in all but state 1. States 0 and 3, particularly, have dramatically-extended C termini and are also found to be interacting with novel nearby protein chains (Supplementary Figure 4.S5). State 3 is the second most populated state at equilibrium, at 22%. Importantly, both the N and C termini have been modeled computationally as the crystal structure was missing atom coordinates, such that the motions here observed could be artifacts from the modeled starting conformation. However, the high correlation obtained for the simulated diffuse scattering compared to experimental validates at least in part the accuracy of the protein motions observed in the simulations.

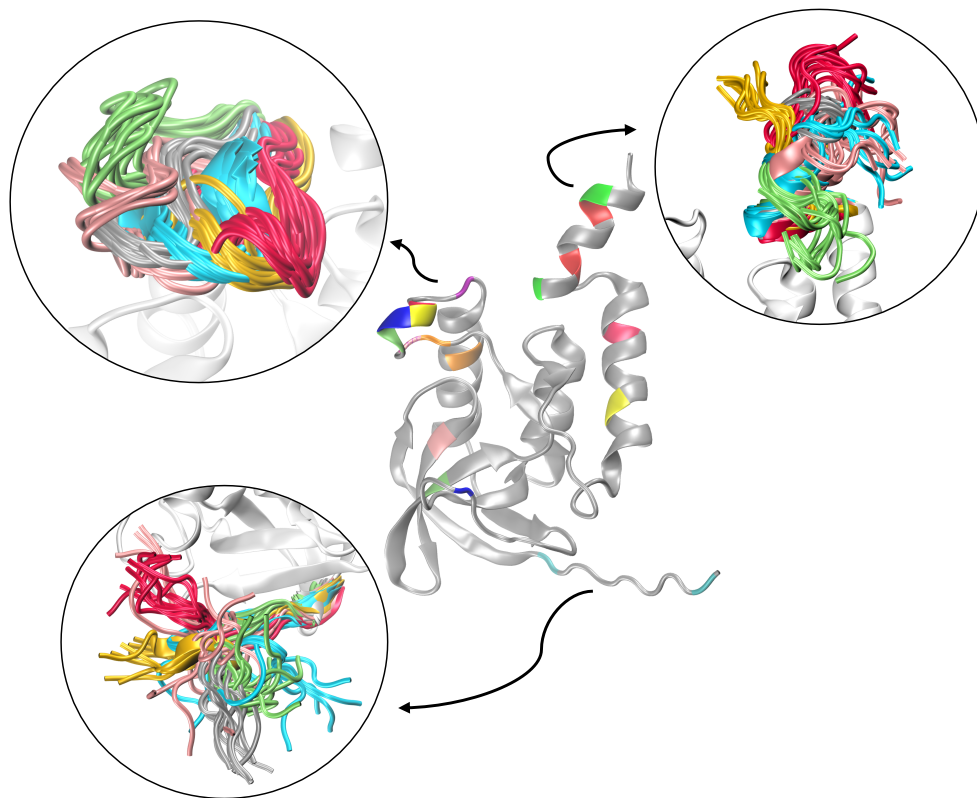


Figure 4.4. Details of the different flexible regions identified in the metastable states. States' colors are the same as in Figure 4.3.

The central loop also explores a significant range of conformations, moving both closer to the protein core from the starting conformations (in states 0, 3 and 4) as well as further away (most strikingly in state 2), in a hinging movement spanning 15.5 Å. Interestingly, the two most populated states at equilibrium, 3 and 4, exhibit the folded, closer-to-protein-core conformations, suggesting this as the most relevant loop conformation in the crystalline environment. In these states the loop is sufficiently close to the unfolded C terminal that a salt bridge between these two motifs can be observed.

State distributions and cross-correlations in the crystal

The description of the protein ensemble in terms of these metastable states has the advantage of allowing the interrogation of whether there are preferred conformations accessed by each protein chain in the supercell and in providing information on the presence (or lack thereof) of crystal symmetry at the dynamic level. In order to investigate that, we computed each chain's metastable state distribution, that is, the relative frequency with which each state is visited by each chain in the supercell simulation (Figure 4.5a). About a third of the protein copies, such as chains 4, 6, 7 and 8, have clear metastable state memberships. The majority of the chains, however, can be split between two or more preferred states.

Even though every 4th chain is in symmetry-equivalent positions in each unit cell making up the supercell, the metastable distribution does not follow this four-fold symmetry, suggesting that symmetry is not conserved in terms of these proteins' conformational ensemble. This becomes clearer as we look at each unit cell state distribution individually (Figure 4.5b) or reconstruct the supercell in terms of each chain's most probable state (Figure 4.5c). Interestingly, there appears to be a dominance of states 0 and 5 at each respective sides of the supercell. It is thus evident that even in these constrained environments there is still considerable protein motion that moves the crystalline system away from a symmetric distribution and highlights the importance of considering deviations from the perfect crystal in structure prediction based on X-ray experiments²⁹.

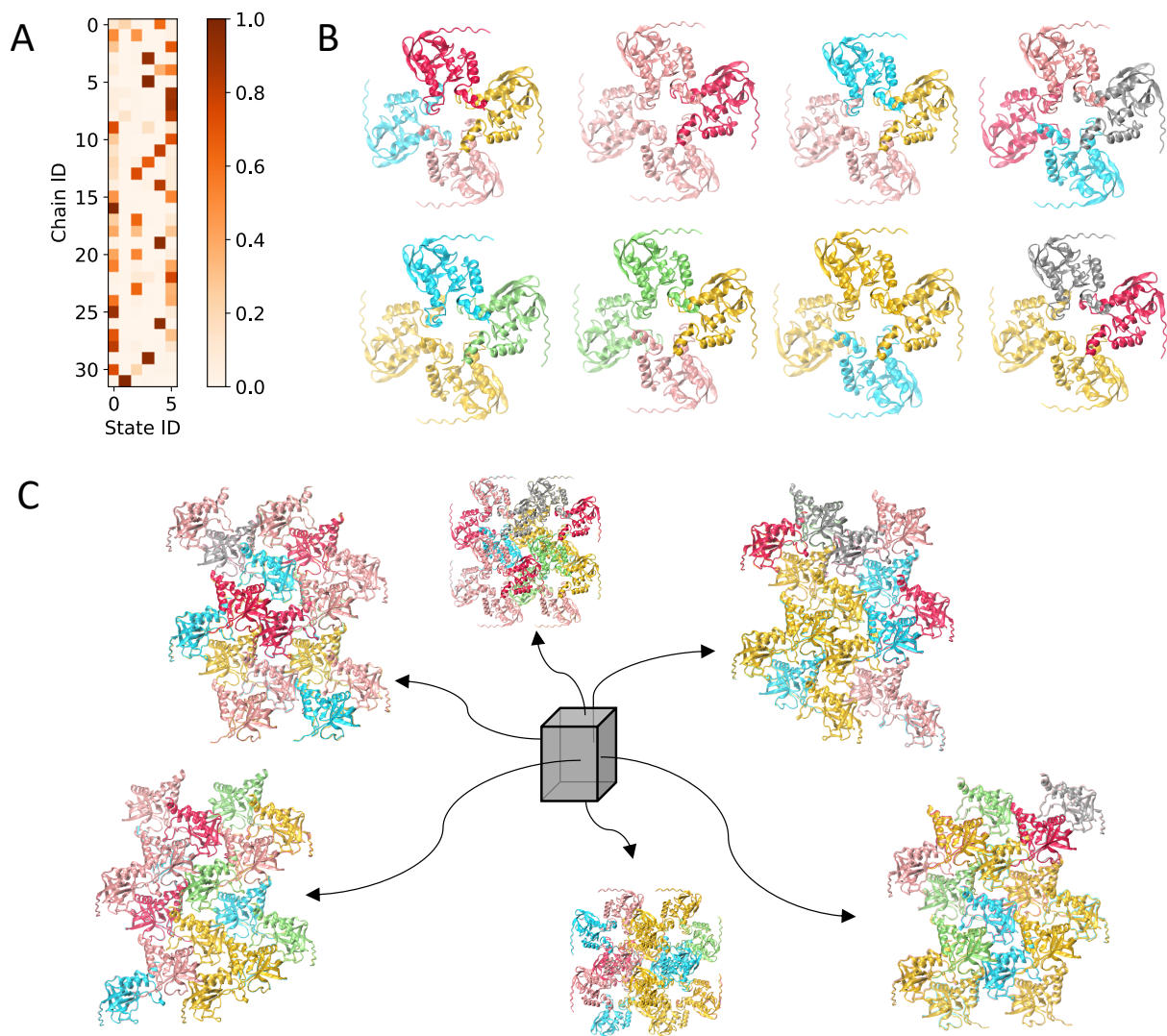


Figure 4.5. Chain state distributions in the supercell. (a) Per-chain metastable state sampling across the simulation. (b) Unit cell preferred state distribution for the eight asymmetric units in the supercell. (c) Views of the supercell according to each chain preferred state.

Analysis of the time evolution of the state distributions in the simulation indicate a significant degree of state cross-correlation among the chains in the supercell, as shown in Figure 4.6a. Interestingly, such correlation found between states is not random, as reconstructed crystal trajectories in which each chain's conformational states were drawn at random, or weighted by the equilibrium population obtained in the MSM, do not show any significant chain cross-correlations

(Figure 4.6b). This indicates that the chains' state dependencies found here are due to causal relationships that have to be brought about by inter-chain interactions.

Almost half of the 32 chains in the supercell show high cross-correlations with at least one other chain (according to a cutoff of absolute Pearson correlation of above 0.5). These are distributed across the supercell, as shown in Figure 4.6c, and the correlations surpass the unit cell boundaries and involve proteins located at large distances even when chain distances are corrected for crystal symmetry (Figure 4.6d), indicating the existence of long-range communication between the chains. These findings agree with the emerging understanding that long-range correlations play an important role in the origin of diffuse scattering^{10,17}.

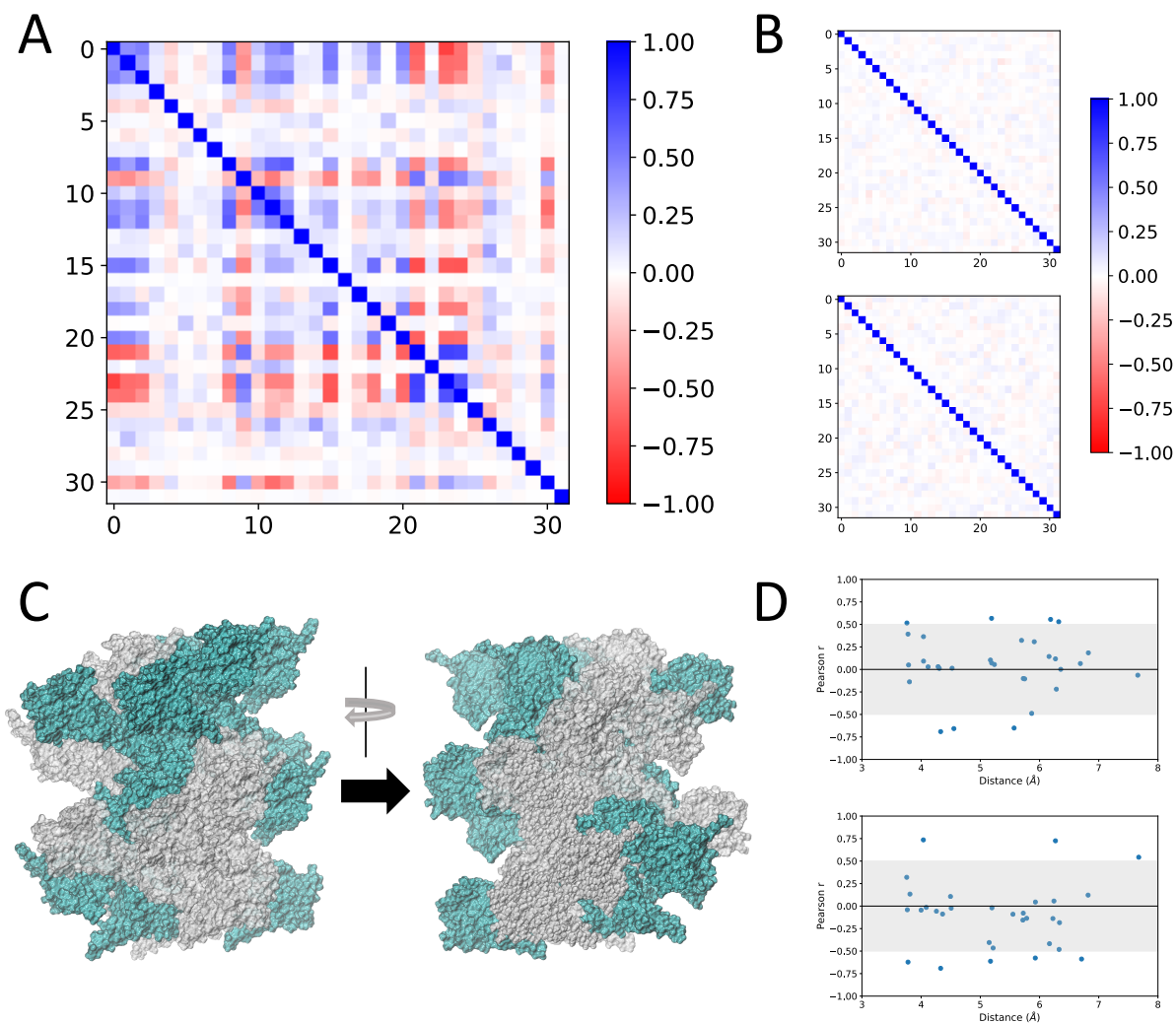


Figure 4.6. Chain cross-correlations. (a) Pearson correlations between states during the simulation for every pair of chains in the supercell. (b) Pearson correlations computed from reconstructed crystal trajectories in states drawn at random (top) or weighted by the equilibrium population obtained in the MSM (bottom). (c) Representation of the highly correlated chains in the supercell, shown in cyan. (d) Symmetry-corrected distance dependence of the inter-chain correlations for two representative chains in the supercell.

Prediction of diffuse scattering from reconstructed crystals based on MSM states

A current limitation of MSMs is that the conformations explored by the MD simulations are clustered into discrete states based on a researcher-selected state definition (e.g., RMSD of the backbone carbons, torsion angles, or residue contact maps). This first step in model construction

essentially defines the protein conformations that are taken into the following steps of data discretization (clustering), transition probability calculation and coarse graining for human-interpretation. In this way, the feature choice affects all MSM outcome and an appropriate selection for the problem being investigated is a determinant step for accurate model construction. Other important parameters that may affect the accuracy of MSMs are the lag time, the number of clusters used in the discretization step and the number of metastable states used for coarse graining and construction of hidden MSMs. There are some established methods and tests used to verify the accuracy of the constructed model, such as the implied timescale plots and Chapman-Kolmogorov tests. However, despite improvements in the methodology and code, MSM construction remains very much an art greatly led by intuition. However, diffuse scattering could be used to overcome this limitation, by monitoring different selection criteria for increased agreement with the experimental data.

We decided to test that by reconstructing trajectories of the supercell based on the state equilibrium populations predicted by the MSM (Figure 4.3d). A varying number of crystal frames were reconstructed using the *CrystalBuilder* program being developed by our collaborators. However, the correlations with the experimental diffuse scattering obtained thus far remain very low, maxing at 0.45 in the total intensity and < 0.2 in the anisotropic scattering (Figure 4.7), which constitute poorer results than those computed directly from the original MD simulations².

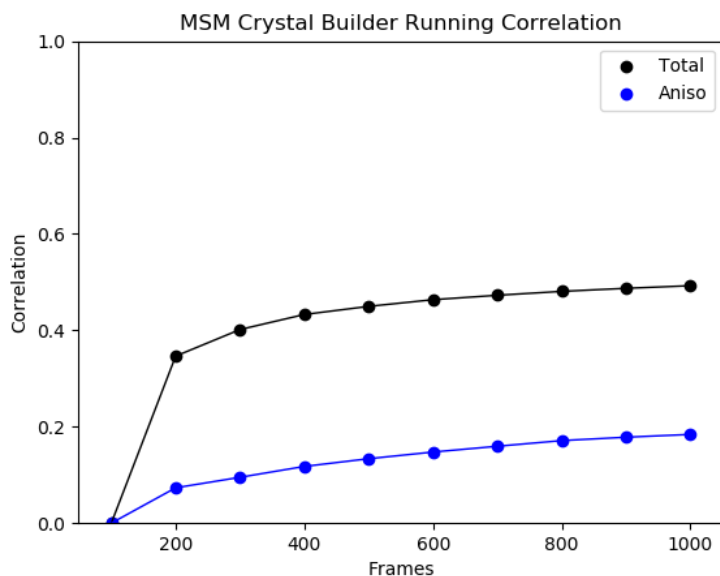


Figure 4.7. Correlation of experimental and MSM-computed diffuse scattering.

This poor initial result is not completely surprising considering that this naïve reconstruction of the crystal based on the single chain MSM state populations did not incorporate information on the correlated motions between the chains observed above, and did not account for likely atomic clashes between neighboring chains. To try to address this weakness, a new version of *CrystalBuilder* is currently being developed to incorporate the information on chains cross correlations and conditional probabilities (that is, the probability that a chain in a particular location in the crystal will be in a determined state given the state of another) in the reconstruction of the crystals, and we expect to be able to capture some improvement in the modeling of the experimental diffuse scattering.

Exploring the influence of different feature definitions

The comparison of MSM-predicted and experimental diffuse scattering can help us explore the influence of different feature sets on the obtained models of dynamics and the predicted

diffuse scatterings. Thus, in addition to the above constructed MSM, we decided to verify if additional models based on different features could be constructed from the supercell simulations. Previously, interesting dynamics were observed in the active site residues of crystalline staphylococcal nuclease simulations¹³. Calculating all pairwise distances between the same set of residues yielded another successful MSM, this one containing 3 states based on the timescale separation in the implied timescale plots (Figure 4.8).

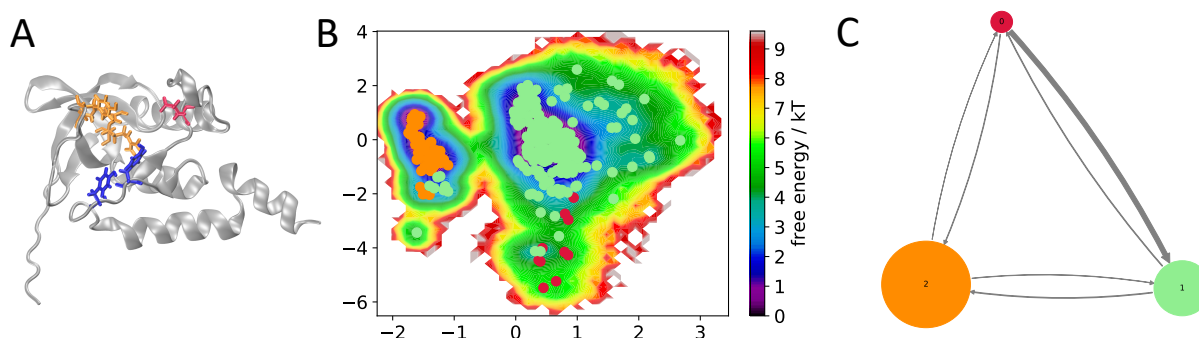


Figure 4.8. Active site MSM. (a) Active site residues. Features consisted of all combination of pairwise distances between them. (b) Free energy landscape with metastable states distribution overlaid. (c) Metastable states identified by Hidden Markov models.

An alternative avenue that could potentially better account for the chain's intermolecular interactions could be the use of features that integrate descriptions of the unit cell, besides just the intramolecular distances so far being considered. A model defined in terms of a few inter-chain distances (to account for global protein motions in the unit cell) in addition to the internal features from the initial model could be successfully constructed (Figure 4.9), although model statistics is worsened by the significantly smaller sampling of the full unit cell conformations from the supercell simulation (8 unit cells x 5 μ s = 40 μ s). In this model the unit cell dynamics can be coarse-grained into four metastable states. Importantly, both of these models showed satisfactory

profiles in the MSM validation tests (Supplementary Figures 4.S6 and 4.S7) despite describing completely different protein motions. This underscores the challenges involved in selecting the features for MSM construction. Additionally, validation of the predicted diffuse scattering profiles based the respective models' state populations and correlations has the potential of providing a means for directed investigation of the roles of different scales and ranges of motion in the diffuse scattering measured experimentally.

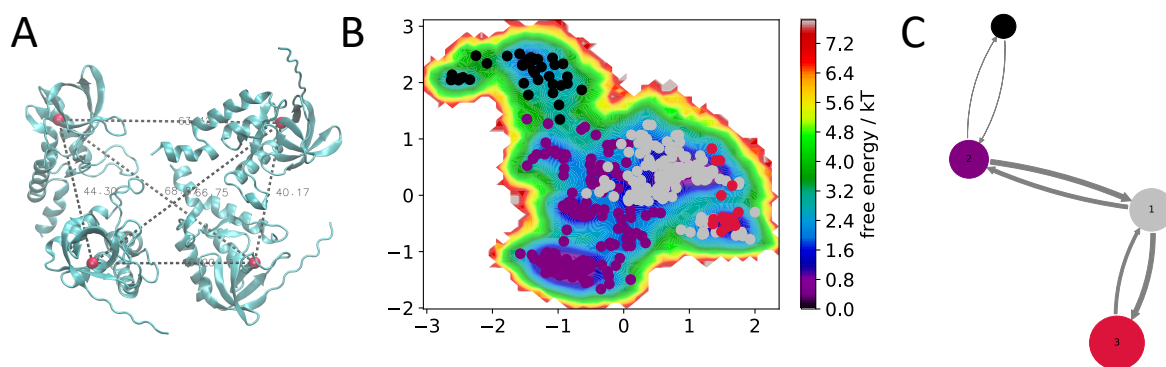


Figure 4.9. Unit cell MSM. (a) Representation of inter-chain distances used as features. (b) Free energy landscape with metastable states distribution overlaid. (c) Metastable states identified by Hidden Markov models.

4.5 Conclusions

Based on an MSM-directed metastable state investigation of the conformational ensemble of chains of a supercell of *staphylococcal* nuclease, it becomes clear how the proteins in these crystalline environments retain a significant degree of flexibility, which affect the diversity of interactions formed with neighboring chains. Our models indicate that these interactions are not constrained to local pockets but surpass the unit cell to attain much larger distances involving several protein chains. This highlights the importance of considering many unit cells in the simulations of protein crystals¹⁷.

Additionally, our work evidences how, given enough sampling, accurate MSM models can be constructed from a variety of feature selections. We have remained within the realm of pairwise distances, but even more models could theoretically be made based on RMSD to reference structures or torsion angles, to name a few possibilities. How does one choose the best set of features to build a model, and the other several parameters that are necessary during MSM construction? Here, we propose the use of diffuse scattering as an experimental observable against which to tune the MSM parameters. Additionally, the comparison of predicted diffuse scattering by the different MSM models against the experimental measurements provides a mechanism for investigating the weights that distinct areas or scale of protein correlated motions play in the origin of diffuse scattering.

4.6 Acknowledgements

Chapter 4, in full, is currently being prepared for submission for publication of the material. “Barros, E. P., Wych, D., Wall, M. E., Mobley, D., Amaro, R.E., On the conformational diversity within protein crystals”. The dissertation author is the primary investigator and author of this paper.

4.7 Supplementary Information

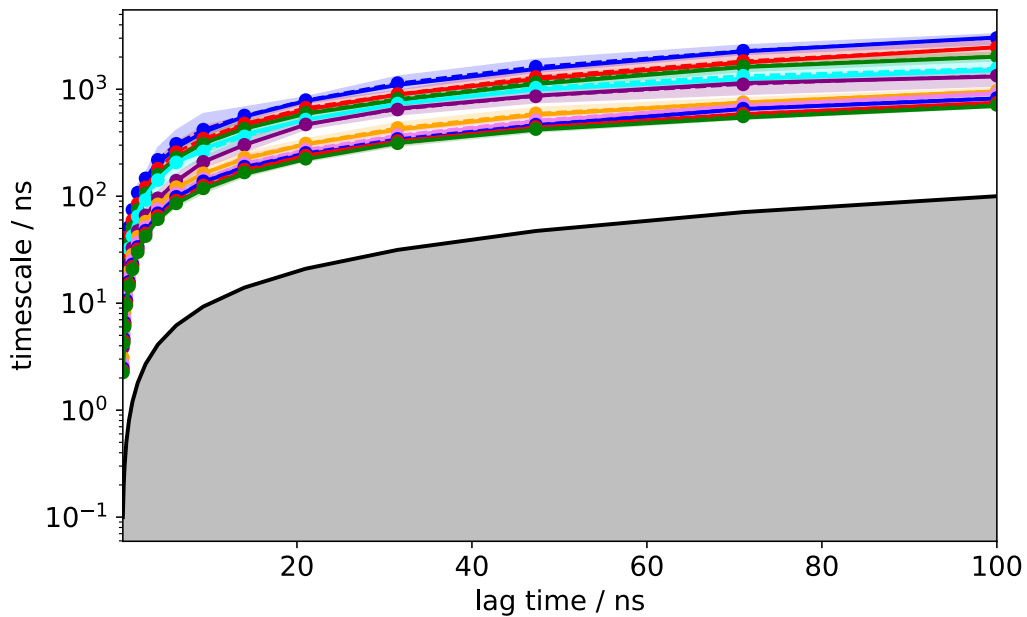


Figure 4.S1. Implied timescale plot for the flexible regions MSM.

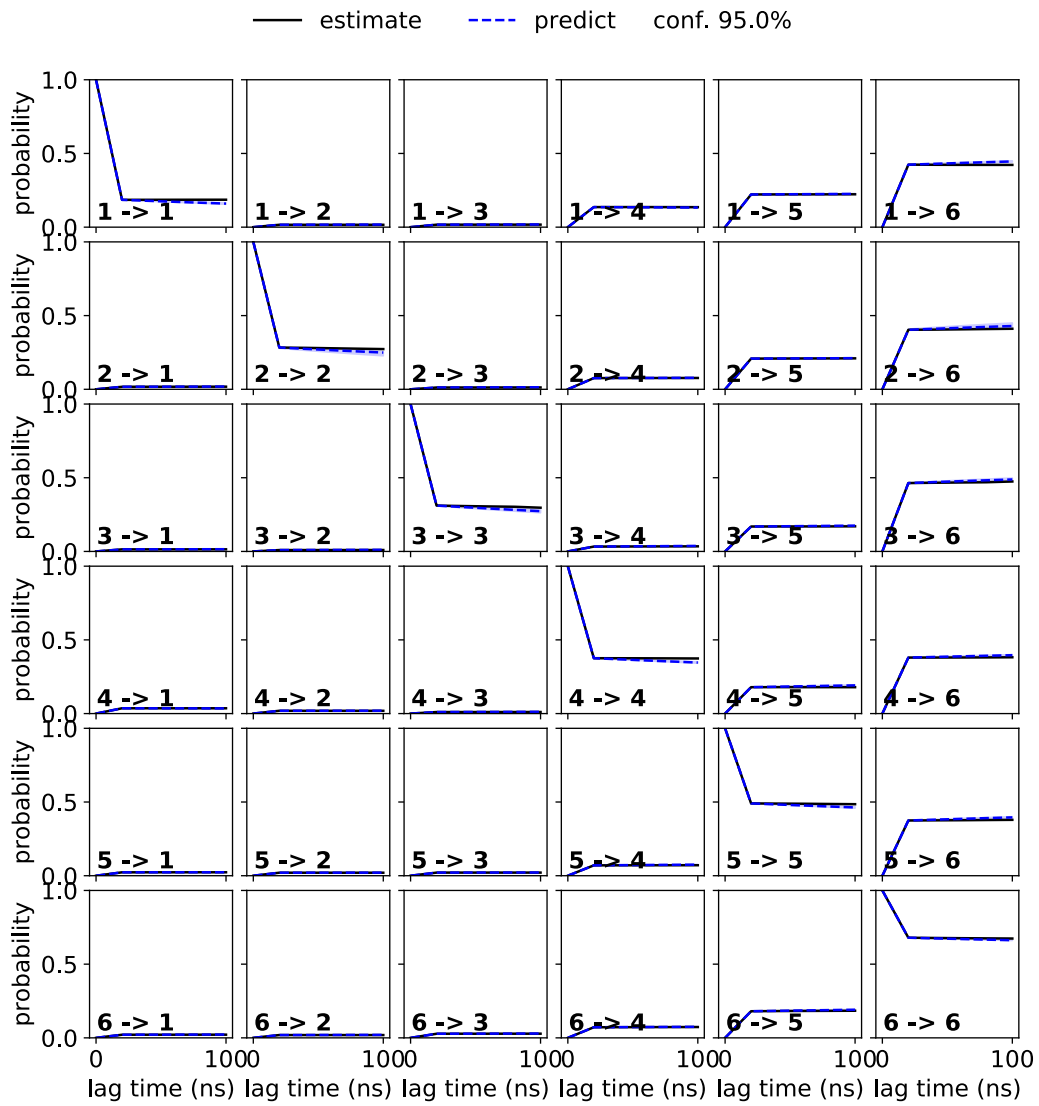


Figure 4.S2. Chapman-Komolgorov test for the flexible regions MSM.

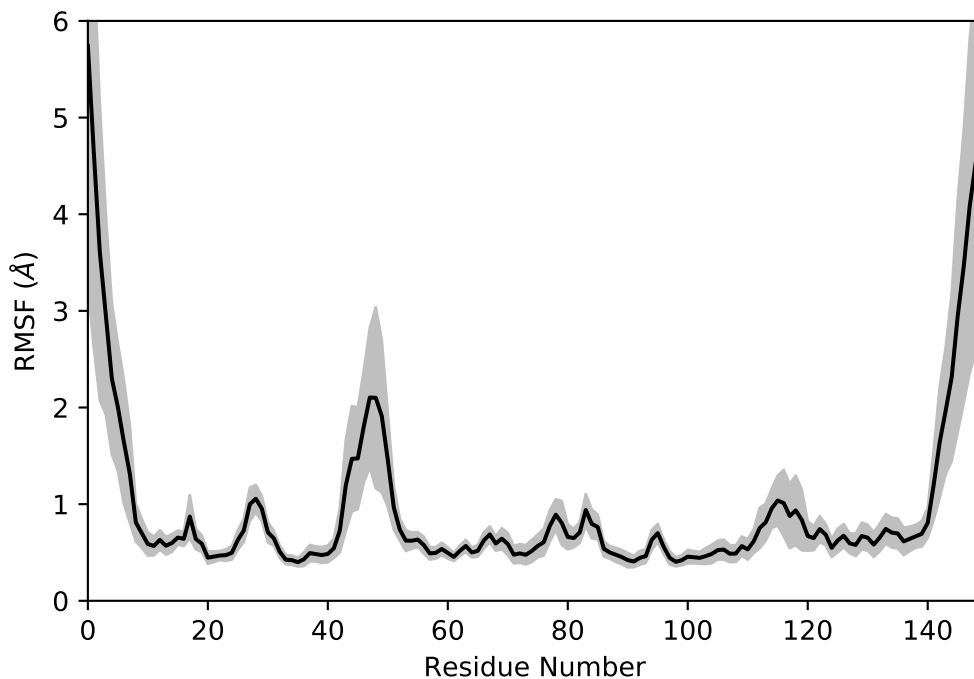


Figure 4.S3. Protein RMSF. Average of all chains is shown in black, and standard deviation as gray area.

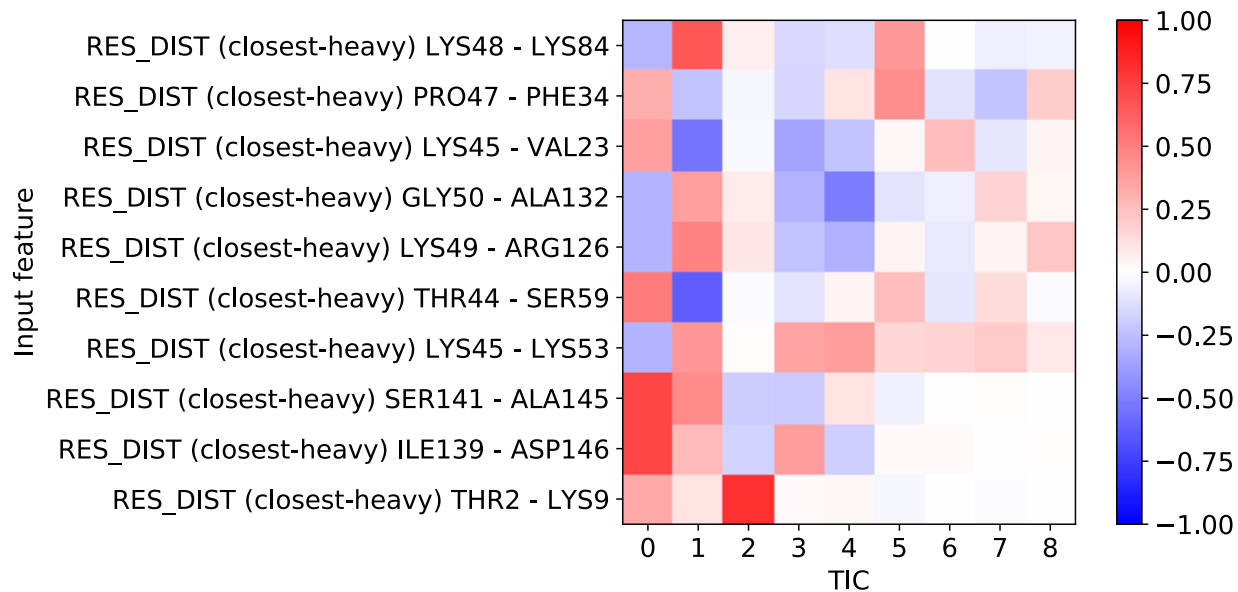


Figure 4.S4. Correlation of flexible regions features identified by the tICA-directed procedure with tICA coordinates.

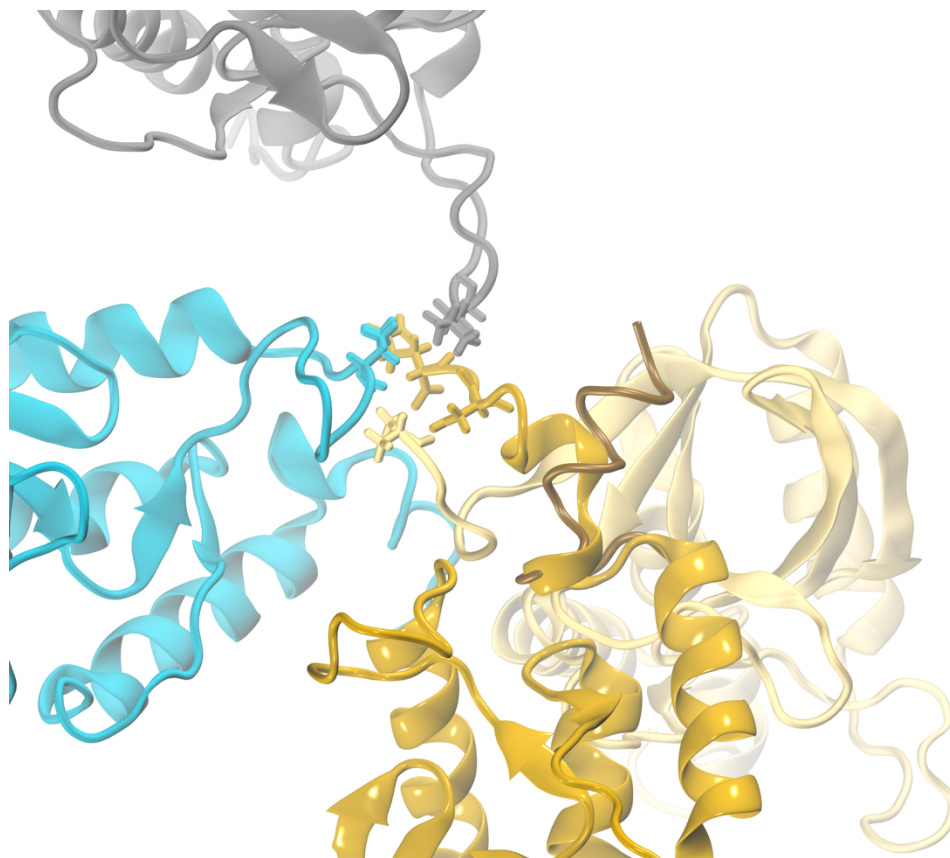


Figure 4.S5. Inter-chain interactions observed for a protein chain that exhibits the extended C-terminal conformation seen in metastable state 0 (yellow, center). The conformation of the C terminal in the starting structure is represented in brown. Nearby chains are colored according to their metastable membership at that exact frame. Residues involved in inter-chain hydrogen bonding interactions are represented in licorice.

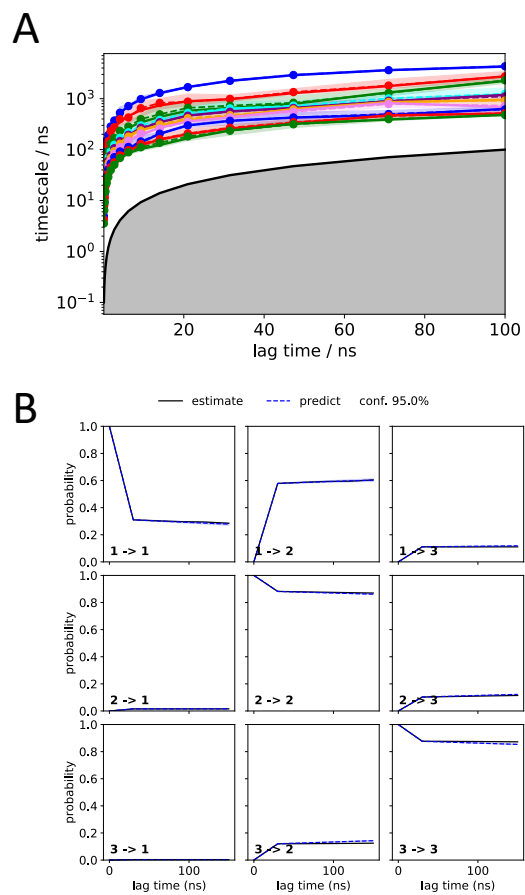


Figure 4.S6. Validation metrics of the active site MSM (a) Implied timescale plot, (b) CK test.

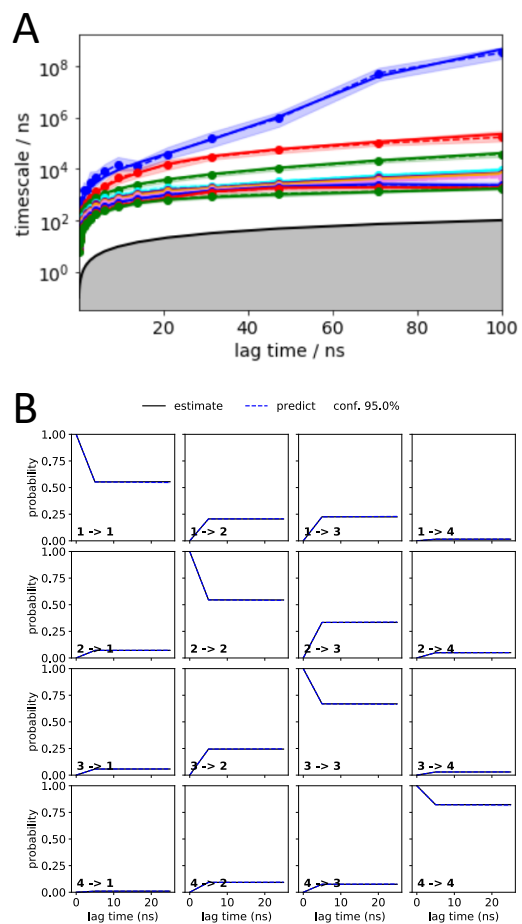


Figure 4.S7. Validation metrics of the unit cell MSM (a) Implied timescale plot, (b) CK test.

Table 4.S1. Pairs selected as initial features in the tICA analysis.

Location in protein	Residue 1	Residue 2
N terminal	Thr2*	Lys9*
	Lys5	Leu7
β-barrel	Glu10	Val74
	Ala12	Leu89
	Leu14	Leu25
	Ile18	Ile92
	Val23	Phe34
	Ile72	Tyr93
	Lys16	Glu73
β -barrel - loop	Gln30	Lys16
	Met32	Gly86
β -sheets	Asp21	Tyr113
	Val39	Lys110
Loops	Leu37	Asp40
	Glu75	Gln80
Intra - helices	Leu38	Lys78
	Lys116	Thr120
	Lys63	Gly67
	Arg126	Lys133
Inter - helices	Gln123	Arg126
	Met65	Val99
	Ala58	Gln106
	Pro56	Ala132
Helix - sheets	Ala102	Ser128
	Lys24	Ala58
C-terminal	Val111	Arg126
	Ile139*	Asp146*
Internal flexible loop	Ser141*	Ala145*
	Lys45*	Lys53*
Flexible loop - helix	Pro47	Gly50
	Thr44*	Ser59*
	Lys49*	Arg126*
	Gly50*	Ala132*

Table 4.S1. Pairs selected as initial features in the tICA analysis (continued).

Location in protein	Residue 1	Residue 2
Flexible loop - sheet	Lys45*	Val23*
	Pro47*	Phe34*
	Lys48*	Lys84*

* Indicates features carried over for MSM construction because of high correlation with tICA components

4.8 References

- (1) Wall, M. E.; Wolff, A. M.; Fraser, J. S. Bringing Diffuse X-Ray Scattering into Focus. *Curr. Opin. Struct. Biol.* **2018**, *50*, 109–116. <https://doi.org/10.1016/j.sbi.2018.01.009>.
- (2) Wall, M. E. Internal Protein Motions in Molecular-Dynamics Simulations of Bragg and Diffuse X-Ray Scattering. *IUCrJ* **2018**, 172–181.
- (3) Wall, M. E.; Adams, P. D.; Fraser, J. S.; Sauter, N. K. Diffuse X-Ray Scattering to Model Protein Motions. *Structure* **2014**, *22*, 182–184.
- (4) Riccardi, D.; Cui, Q.; Phillips Jr., G. N. Evaluating Elastic Network Models of Crystalline Biological Molecules with Temperature Factors, Correlated Motions, and Diffuse X-Ray Scattering. *Biophys. J.* **2010**, *99* (8), 2616–2625. <https://doi.org/10.1016/j.bpj.2010.08.013>.
- (5) Polikanov, Y. S.; Moore, P. B. Acoustic Vibrations Contribute to the Diffuse Scatter Produced by Ribosome Crystals. *Acta Crystallogr., Sect. D Biol. Crystallogr* **2015**, *71*, 2021–2031. <https://doi.org/10.1107/S1399004715013838>.
- (6) Van Benschoten, A. H.; Liu, L.; Gonzalez, A.; Brewster, A. S.; Sauter, N. K.; Fraser, J. S.; Wall, M. E. Measuring and Modeling Diffuse Scattering in Protein X-Ray Crystallography. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (15), 4069–4074. <https://doi.org/10.1073/pnas.1524048113>.
- (7) Cleary, M. Molecular Dynamics Studied by Analysis of the X-Ray Diffuse Scattering from Lysozyme Crystals. *Doucet, J. Benoit, J. P.* **1987**, *325*, 643–646. <https://doi.org/10.1017/CBO9781107415324.004>.
- (8) Clarage, J. B.; Clarage, M. S.; Phillips, W. C.; Sweet, R. M.; Caspar, D. L. D. Correlations of Atomic Movements in Lysozyme Crystals. *Proteins* **1992**, *12*, 145–157. <https://doi.org/10.1002/prot.340120208>.
- (9) Moore, P. B. On the Relationship between Diffraction Patterns and Motions in Macromolecular Crystals. *Structure* **2009**, *17*, 1307–1315. <https://doi.org/10.1016/j.str.2009.08.015>.

- (10) Peck, A.; Poitevin, F.; Lane, T. J. Intermolecular Correlations Are Necessary to Explain Diffuse Scattering from Protein Crystals. *IUCrJ* **2018**, *5*, 211–222. <https://doi.org/10.1107/S2052252518001124>.
- (11) Héry, S.; Genest, D.; Smith, J. C. X-Ray Diffuse Scattering and Rigid-Body Motion in Crystalline Lysozyme Probed by Molecular Dynamics Simulation. *J. Mol. Biol.* **1998**, *279*, 303–319. <https://doi.org/10.1006/jmbi.1998.1754>.
- (12) Meinhold, L.; Smith, J. C. Protein Dynamics from X-Ray Crystallography: Anisotropic, Global Motion in Diffuse Scattering Patterns. *Proteins* **2007**, *66*, 941–953. <https://doi.org/10.1002/prot>.
- (13) Wall, M. E.; Van Benschoten, A. H.; Sauter, N. K.; Adams, P. D.; Fraser, J. S.; Terwilliger, T. C. Conformational Dynamics of a Crystalline Protein from Microsecond-Scale Molecular Dynamics Simulations and Diffuse X-Ray Scattering. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (50), 17887–17892. <https://doi.org/10.1073/pnas.1416744111>.
- (14) Chan, E. J. On the Use of Molecular Dynamics Simulation to Calculate X-Ray Thermal Diffuse Scattering from Molecular Crystals. *J. Appl. Cryst.* **2015**, *48*, 1420–1428.
- (15) Wych, D. C.; Fraser, J. S.; Mobley, D. L.; Wall, M. E. Liquid-like and Rigid-Body Motions in Molecular-Dynamics Simulations of a Crystalline Protein. *Struct. Dyn* **2019**, *6*, 064704. <https://doi.org/10.1063/1.5132692>.
- (16) Cerutti, D. S.; Case, D. A. Molecular Dynamics Simulations of Macromolecular Crystals. *Wires Comput. Mol. Sci.* **2019**, *9*, e1402. <https://doi.org/10.1002/wcms.1402>.
- (17) Meisburger, S. P.; Case, D. A.; Ando, N. Diffuse X-Ray Scattering from Correlated Motions in a Protein Crystal. *bioRxiv* **2019**. <https://doi.org/10.1101/805424>.
- (18) Meinhold, L.; Smith, J. C. Fluctuations and Correlations in Crystalline Protein Dynamics: A Simulation Analysis of Staphylococcal Nuclease. *Biophys. J.* **2005**, *88* (4), 2554–2563. <https://doi.org/10.1529/biophysj.104.056101>.
- (19) Wagner, R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. **2016**. <https://doi.org/10.1021/acs.chemrev.5b00631>.
- (20) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48* (2), 414–422. <https://doi.org/10.1021/ar5002999>.
- (21) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99–105. <https://doi.org/10.1016/j.ymeth.2010.06.002>.
- (22) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schtte, C.; No??, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.*

- 2011, 134, 174105. <https://doi.org/10.1063/1.3565032>.
- (23) Bowman, G. R.; Pande, V. S. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*.
- (24) Chodera, J. D.; Noe, F. Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Bstructural Biol.* **2014**, 25, 135–144. <https://doi.org/10.1016/j.sbi.2014.04.002>.
- (25) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noe, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, 11, 5525–5542. <https://doi.org/10.1021/acs.jctc.5b00743>.
- (26) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, 139, 015102. <https://doi.org/10.1063/1.4811489>.
- (27) Wall, M. E.; Ealick, S. E.; Gruner, S. M. Three-Dimensional Diffuse x-Ray Scattering from Crystals of Staphylococcal Nuclease. *Proc. Natl. Acad. Sci.* **1997**, 94, 6180–6184.
- (28) Ma, P.; Xue, Y.; Coquelle, N.; Haller, J. D.; Yuwen, T.; Ayala, I.; Mikhailovskii, O.; Willbold, D.; Colletier, J.-P.; Skrynnikov, N. R.; et al. Observing the Overall Rocking Motion of a Protein in a Crystal. *Nat. Commun.* **2015**, 6, 8361. <https://doi.org/10.1038/ncomms9361>.
- (29) Woldeyes, R. A.; Sivak, D. A.; Fraser, J. S. E Pluribus Unum , No More: From One Crystal, Many Conformations. *Curr. Opin. Struct. Biol.* **2014**, 28, 56–62. <https://doi.org/10.1016/j.sbi.2014.07.005>.

Towards mutant-specific therapies: Uncovering the dynamical landscape and druggability of p53 DNA binding domain with Markov State Models

Emilia P. Barros¹, Özlem Demir¹, Rommie E. Amaro^{1,2}

Author affiliations:

1 – Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA

2 – National Biomedical Computation Resource, University of California, San Diego, La Jolla, CA, USA

5.1 Abstract

The transcription factor p53 functions as a tumor suppressor and is the most frequently mutated gene in human cancer. Its inactivation by single point mutation is associated with progression of about 50% of cancers, and therefore reactivation of mutated p53 is emerging as an exciting possibility for cancer therapy. More than 90% of the cancer mutations are found in the DNA-binding domain of p53, but the mechanism through which a single mutation leads to change in function remains elusive and hinders the rational development of mutant-specific drug leads. Analysis of long-timescale molecular dynamics simulations of monomeric wildtype and the Y220C cancer mutant through the Markov state model framework has uncovered the involvement of loop 6 (L6), where the mutation is located, in the slowest dynamics in the protein. Due to its location far from the DNA binding surface, the conformational dynamics of this loop has so far remained largely unexplored. However, our simulations indicate the existence of allosteric communication between L6 and the functionally-important loop L1 as the mutation affects not only the conformational ensemble of the former but also of the latter. We observe the stabilization of alternate L6 conformations, distinct from all available X-ray crystal and NMR structures, in which the loop is extended and located further away from L1. As L6 can form hydrogen-bonding interactions with L1 when in the recessed conformation, our simulations suggest an allosteric mechanism for the inactivation effect of the Y220C mutation and evidence the existence of several novel protein conformations that can be targeted for p53 rescue efforts. Our approach exemplifies the power of the differential dynamics MSM methodology for uncovering intrinsic dynamical and kinetic differences among distinct protein ensembles.

5.2 Introduction

The transcription factor p53, known as the “guardian of the genome”, is the most important tumor suppressor in humans due to its regulation of a wide range of cellular activities such as cell cycle arrest, apoptosis, senescence and promotion of anti-tumor microenvironments^{1,2}. Because of its role preventing tumor initiation and maintenance, p53 is found to be the most frequently mutated gene in human cancers^{3,4}. Loss of its function through missense mutations is associated with progression of about half of human cancers^{5,6}, and therefore reactivation of mutated p53 is emerging as an exciting possibility in cancer treatment as it has been found to lead to tumor regression⁷⁻¹¹.

More than 90% of the cancer mutations are found in the DNA-binding domain (DBD) of p53¹² (Figure 5.1a), but the mechanism through which a single mutation affects function is far from resolved. Moreover, the current paradigm is that p53 mutants are not equivalent proteins, but rather have distinct individual profiles in terms of loss of wildtype activity and acquisition of unique tumor-promoting gain of functions^{13,14}. Generally, the oncogenic variations can be classified as contact mutations, which lead to loss of function due to disruption of the interaction network with DNA¹⁵, or structural mutations, which cause perturbations to the DBD and lead to inactivation due to destabilization of the protein structure, unfolding and aggregation¹⁶⁻¹⁹.

A strategy currently pursued for reactivation of structural mutants is the development of small molecules that bind to the folded but not the unfolded state of the protein and restore wildtype p53 conformation and function, with promising results achieved by several groups²⁰⁻³⁴. Even in proof-of-concept studies, the success of small molecules in reactivating one or a few specific mutants but not others points to the unique behavior of each p53 cancer mutant. In this way,

exploring and characterizing the dynamic behavior of different p53 cancer mutants as individual entities promises to open up novel therapeutic opportunities for mutant-specific p53 reactivation.

One such mutant being targeted for reactivation through small molecules is Y220C, a structural mutant responsible for about 100,000 new cancer cases every year¹⁶ and the most frequent p53 cancer mutation observed outside the DNA-binding interface of the protein. The mutation of the bulky tyrosine to the smaller cysteine induces the formation of a crevice in the protein surface that is amenable to small molecule binding³⁵⁻³⁷, but so far current efforts have failed to yield very high affinity binders³⁸⁻⁴⁰.

While use of molecular dynamics (MD) simulations has allowed the successful identification of druggable pockets on the protein surface of the p53 core domain^{22,39,41}, our understanding of the protein conformational ensemble and dynamics is restricted by sampling limitations. This leaves large regions of the energy landscape unexplored which may include many of the functionally important slower motions. Already, relatively short-scale MD simulations of Y220C have evidenced the flexibility of the protein and the Y220C pocket³⁹. However, a comprehensive model of p53's conformational ensemble and the underlying free energy landscape is desirable as it will allow the understanding of the dynamics of key loops and druggable pockets and their role in the overall function and motions of the protein. To help in overcoming this sampling limitation, we employ here the Markov state model (MSM) methodology in conjunction with extensive MD simulations for the investigation of the conformational dynamics of wildtype and the Y220C mutant.

MSMs allow the integration of multiple MD simulations into a single model of the protein conformational ensemble that contains key thermodynamic and kinetic properties in addition to retaining atomic level details of the system⁴²⁻⁴⁷. Because the MSM is built on the transitions

between states, the information from multiple MD simulations of the same system can be combined into a single model and no single simulation has to explore all the states. Importantly, as the equilibrium distribution of states can be derived for the final model, the thermodynamics of the states can be determined, in addition to kinetics, principle motions, and transition pathways of the protein conformational ensemble.

In this study, the combination of MD simulations with MSMs allows for the first time a thorough exploration of the conformational ensemble of p53 DBD and uncovers the involvement of a loop located away from the DNA binding site, L6, in the slowest dynamics of the wildtype protein. This is the site of the Y220C mutation but interestingly our models indicate that the mutation affects the conformational landscape of not only L6 but also of the essential L1 loop, which is involved in key interactions with DNA. The existence of allosteric communication between the two loops is suggested and provides a mechanistic rationalization to the effect of the mutation in the activity of p53. Moreover, analysis of the conformational diversity of loop L6 evidences the existence of very distinct loop conformations than previously observed experimentally, and the identification of a novel pocket nestled in the extended conformation of L6 that could be exploited for mutant-specific drug design efforts.

5.3 Methods

System set up

The DNA binding domain initial coordinates were taken from chain B of PDB 1TSR, which include p53 amino acids 96 – 289. For the mutant simulations, the tyrosine in position 220 (125 in the clipped domain) was mutated to a cysteine using tleap module in Amber14⁴⁸. The crystallographic water molecules were retained and each system was solvated in an 8Å TIP3P

water box⁴⁹. The zinc ion and its coordinating residues were modeled using the cationic dummy atom model⁵⁰. Each system was brought to 0.12 M salt concentration by adding K⁺ and Cl⁻ atoms. The structure file of each system consisted about 27,220 atoms, which were prepared using Amber FF14SB force field^{48,51}.

Molecular dynamics simulations

The solvated proteins were minimized and equilibrated using common protocol⁵². To increase the conformational sampling, a round of accelerated MD simulations (aMD)⁵³ was performed from the equilibrated structure using Amber14 program. Each system was simulated for 100 ns and 10 structures were selected for each system by clustering the conformations based on RMSD of the center of mass of each residue using a k-means algorithm in MSMBuilder2⁵⁴ and using the cluster centroids. These 10 structures were used as seeds for short unbiased MD simulations, each performed in triplicate with new starting velocities. After each round of simulation, the joint trajectories were processed for MSM model construction, and new starting coordinates were selected, prioritizing the exploration of new areas in the conformational space, until converged models were obtained based on MSM validation metrics (see below). Individual simulation lengths ranged from 10 to 300 ns. In total, the wildtype system was simulated for 89 μ s, while Y220C required 63 μ s for appropriate model construction.

Markov state model construction

Simulation data was processed and models were built using PyEMMA⁵⁵, version 3.5.6. Features consisted of pairwise distances, with pairs being selected after a tICA-based iterative process that eliminated pairs located consistently close ($< 3\text{\AA}$) or far ($>10\text{\AA}$) in all frames of the

simulations, as well as pairs involving residues located close to the clipped termini, with low variance ($<0.05 \text{ \AA}$) and those that accounted for low correlation with the first tICs (Supplementary Table 5.S1). This is explained in more detail in the Results section. The final feature set consisted of 24 pairs (Supplementary Table 5.S2). Time-independent component analysis (tICA)⁵⁶ with a lag time of 10 ns was used to process the joint wildtype and Y220C featurized data. Distinct loop-centered Markov state models were constructed using the 17 features that are centered in loop L6 and the 7 for L1. Discretization was performed with k-means clustering, $k = 200$, for each system (wildtype and Y220C) separately, and accuracy of the models verified by implied timescale (ITS) plots and Chapman-Kolmogorov tests (Supplementary Figures 5.S1 and 5.S2). The L6 and L1-focused models were constructed with tICA and MSM lag times of 10 ns each.

Pocket characterization

Pocket volume measurements were performed with POVME, version 2.0⁵⁷, and druggability assessments were based on computational solvent mapping of randomly selected conformations from the MSM metastable states using FTMap⁵⁸. Existence of hydrogen bonds across the simulations was probed using MDTraj⁵⁹ (hydrogen bond defined if donor-acceptor distance $< 2.5 \text{ \AA}$ and angle $> 120^\circ$).

5.4 Results and Discussion

L6 is the slowest loop in p53 DBD dynamics

Markov state models provide a framework for exploring protein dynamics with atomic resolution beyond the timescales typically accessed by molecular dynamics simulations. A crucial step when integrating molecular dynamics trajectories for model building is the selection of

features used to discretize the protein conformations sampled, which decreases the dimensionality of the conformational space while still allowing for discrimination between distinct states and appropriate representation of the relevant motions. Depending on the process under investigation, devising the best features for model building can be relatively trivial (such as for ligand binding or protein unfolding), but when aiming for a general understanding of the protein conformational ensemble, the task can become challenging due to the conflict between the large degrees of freedom required to describe the protein ensemble and the need to limit the number of features to a small, tractable number for model building.

To investigate the basal dynamics of wildtype p53, we employed an unbiased method that started from computing all possible pairwise distances (18,336 features), and iteratively performed time lagged Independent Components Analysis (tICA)⁵⁶ to identify the linear combination of features that describe the slowest motions of the system, followed by elimination of the features with low tICA correlation. Using this methodology we arrived at a final number of 24 pairs (Iterative process described in Supplementary Table 5.S1). tICA is useful in the data processing for MSM construction as it maximizes the feature combination to yield kinetically relevant independent components (tICs), which represent the slowest degrees of freedom in the system. Despite starting from all possible pairwise distances and including no directed selection of features besides the elimination of pairs that involve the terminal residues or that are consistently too close ($< 3\text{Å}$) or too far ($>10\text{Å}$) throughout the whole simulations, the final set consisted of interacting pairs centered around loops L1 and L6. Interestingly, all pairs involved at least one residue located in either loop L1 (Ser116) or loop L6 (Pro223, Glu224, Gly226), hereafter referred to as L1 and L6 anchor residues, respectively (Figures 5.1b and c).

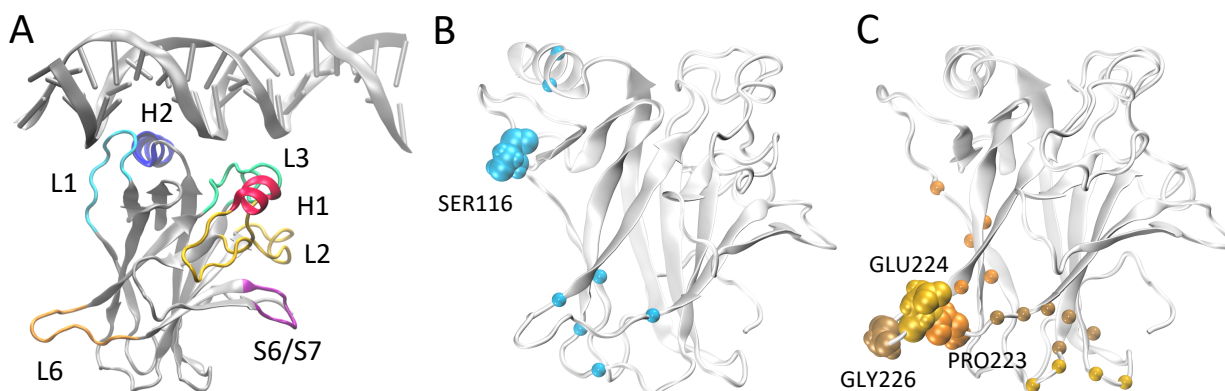


Figure 5.1. (a) Monomeric p53 DNA-binding domain in complex with DNA (from PDB 1TSR) with important functional regions highlighted. (b) and (c) Residues used for MSM construction based on pairwise distances, with L1 (b) and L6 (c) anchor residues highlighted in VDW representation. The C α carbons of the residues that were selected as the second member of the pair with the respective anchor are represented as spheres.

The presence of the repeated anchor residues in the final feature pairs suggests that loops L1 and L6 are involved in the slowest and most significant motions of the protein. Loop L1 is known as a dynamic and biologically important motif for p53 function, having been observed experimentally and computationally in two very distinct conformations^{60–62}, extended (in which it is highly solvent-exposed as represented in Figure 5.1a) and recessed (folded closer to the protein core, with a smaller solvent-accessible surface), both of which are sampled in the simulations. The identification of the relevance of loop L6, however, constitutes novel information in terms of this protein’s conformational dynamics. Not much attention has been given to this structural motif, probably because of its distance from the DNA binding surface. However, elevated B factors in p53 crystal structures points to its intrinsic dynamics, and flexibility in this loop was observed in an early short simulation of wildtype p53 starting from the same crystal structure as in here, but not much was explored in terms of its implication for functionality as it was deemed to stem from a lack of crystal packing⁶³.

For a comparison of the conformational landscapes of wildtype and Y220C, the conformations explored by each of the simulations and represented by the 24 features were jointly used as input for tICA, and are plotted separately in Figure 5.2a. The tIC independent component space is therefore the same for wildtype and mutant free energy landscapes and allows for a direct comparison of the conformational ensemble explored by each system. The wildtype simulation presents two preferred states corresponding to the minima in the free energy landscape. The main distinction between them are the conformations of L1 and indicate the same recessed and extended L1 conformations that have been previously observed (Figure 5.2b). Interestingly, the pairwise features used for construction of the map align very well with the tICA components in this novel feature space, permitting a direct interpretation in terms of protein conformation: tIC1 is closely correlated with features that include loop L6 anchors, and tIC2 is more closely correlated with features involving the L1 anchor, Ser116 (Figure 5.2c). Visual inspection of the conformations distributed on the free energy landscape evidence that smaller values of each of the tICs describe conformations with extended loops (L6 and L1, respectively), while larger values describe the recessed loop conformations. In this way, the transition from low to large tICA values is related to transitions from extended to recessed conformations.

Since the tICs are ordered in terms of slowest to fastest motions, the correspondence of L6 anchor features with the first of the components indicates that, surprisingly, transitions involving loop L6 are slower than those for loop L1. This suggests that important protein dynamics and potentially druggable conformations have so far remained unexplored. While the extended L6 conformations are not that dominant in the wildtype system, the free energy landscape of the Y220C mutant (in which the mutation is located in L6) indicates a significant effect of the single-point mutation on this loop's dynamics, with a much greater proportion of conformations exploring

the extended loop conformation (lower tIC0 values, Figure 5.2a lower panel). Additionally, in the Y220C mutant the recessed L1 conformations lose importance compared to the wildtype, indicated by the loss of the low energy well at high values of tIC1 and tIC2.

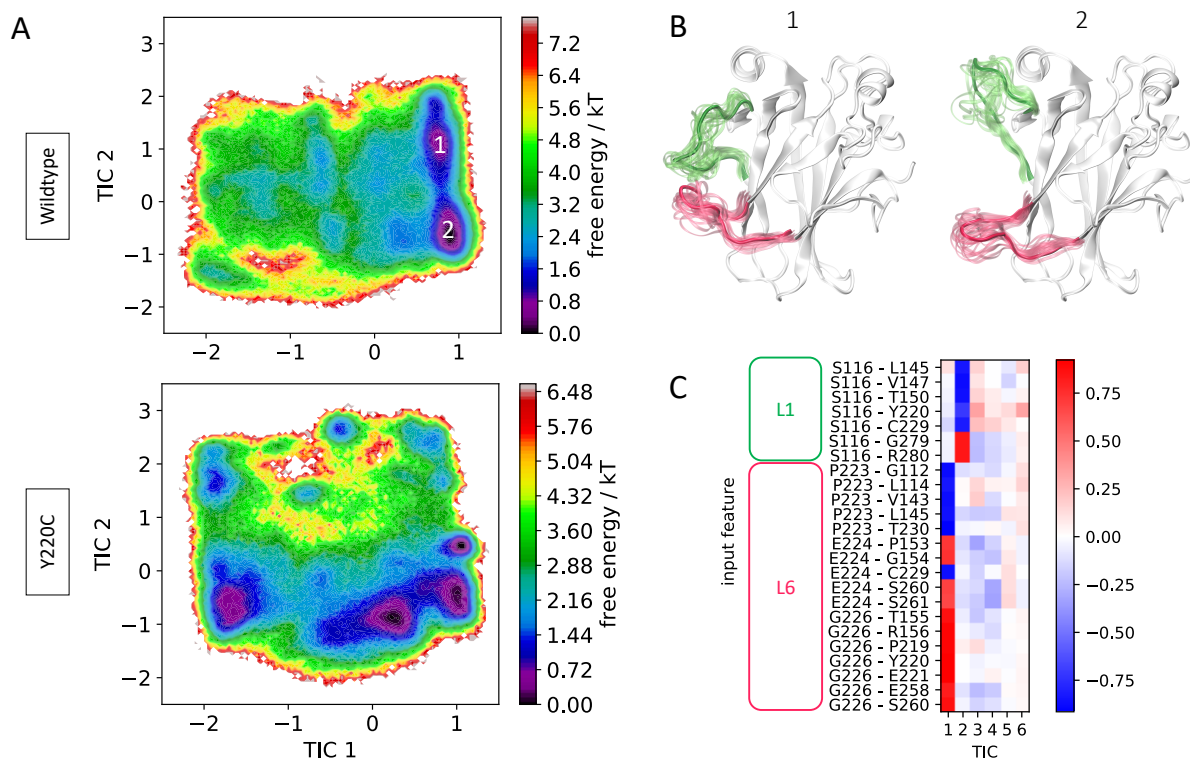


Figure 5.2. (a) Free energy landscape of wildtype (top) and Y220C (bottom) in terms of tICA components (tICs). (b) Representative conformations from the wildtype preferred states. Loops L1 and L6 are highlighted in green and magenta, respectively. (c) Feature correlation with the first five tICA components. Pairwise distances involving L1 or L6 loop anchor residues are indicated.

To further check the importance of these loops in the relevant motions of the protein, we performed additional tICA analysis incorporating other motifs known to play significant roles in p53 function: helices H1 and H2 and loops L2 and L3, which together with L1 make up the DNA interaction surface, and loop S6/7, recently identified as a flexible region in p53 mutants⁶⁴ (Figure 5.1a). Even though several of these loops show pronounced flexibility in the simulations as

indicated by Calpha RMSF (Supplementary Figure 5.S3), loops L1 and particularly L6 still dominate the slowest transitions (Supplementary Figure 5.S4). This suggests that, while other regions such as loops L2 and S6/7 may be highly flexible as evidenced by their high RMSF values, they display fast dynamics and act as further evidence to the important role of loop L6 on the slow dynamics of p53.

Allosteric communication between L1 and L6

In Figure 5.2a it can be seen that the Y220C mutation affects not only the conformational landscape of loop L6, where it is located, but also of loop L1. This loop L1 is essential for p53 activity as it is involved in key interactions with DNA through hydrogen bonds formed by Lys120 and Ser121⁶². Wildtype p53 shows important intrinsic L1 flexibility, but the effect of the mutation on this loop's dynamics indicates the existence of possible long-range communication between L1 and L6.

To look into this in more detail, we constructed MSMs for the wildtype and mutant system using only the above identified features that include the L1 anchor, Ser116. The free energy landscape in terms of these 7 features, following tICA transformation, is shown in Figure 5.3a. Coarse-graining of the structures using Hidden Markov state models identifies the presence of 5 metastable states in each case. Two metastable states, states A and B, are retained in the mutant system. State A is the most populated state in both systems, and shows loop L1 in the most extended-like conformations (average loop L1 alpha carbon RMSD to the extended L1 in chain B of 1TSR is 2.19 Å for wildtype and 3.10 Å for Y220C). In wildtype state B, we see a previously-identified 3-10 helix in the L1 loop, absent in the corresponding Y220C state.

The second, shallower wildtype minima, centered at TIC1 = -1, is absent in the Y220C sampled conformations. Indeed, we find that two wildtype metastable states are abrogated by the mutation (states C and D), being substituted by a single state in the Y220C system (state F). These wildtype states show L1 in recessed conformations, and jointly account for 19% of the equilibrium population. Interestingly, in both cases we find that loop L6 is also organized in a recessed conformation, such that both loops are located in close proximity to each other. Investigation of the loop residues suggests the existence of inter-loop hydrogen bonds formed between the side-chain oxygen of Ser116 in L1 and backbone nitrogen of Asp228 (in state C) or side-chain oxygen of Thr231 (state D) in L6 (Figure 5.3b and Supplementary Figure 5.S5).

Loop L1 in the corresponding Y220C state F, on the other hand, is found to be more collapsed into the protein surface, in a conformation that does not allow for interaction with loop L6. Rather, a salt bridge between loop L1's Lys120 and Glu198 in loop S5/S6 seems to promote the stabilization of this alternate conformation, which accounts for 31% of the Y220C equilibrium population. The sequestering of the DNA-interacting Lys120 in this significant metastable state could provide a mechanistic explanation to the p53 inactivation effect of the mutation. Even more interestingly, the conformation-dependent interaction between loop L1 and L6 identified here suggests the existence of allosteric communication between the loops in functional p53, which is disrupted by the mutation.

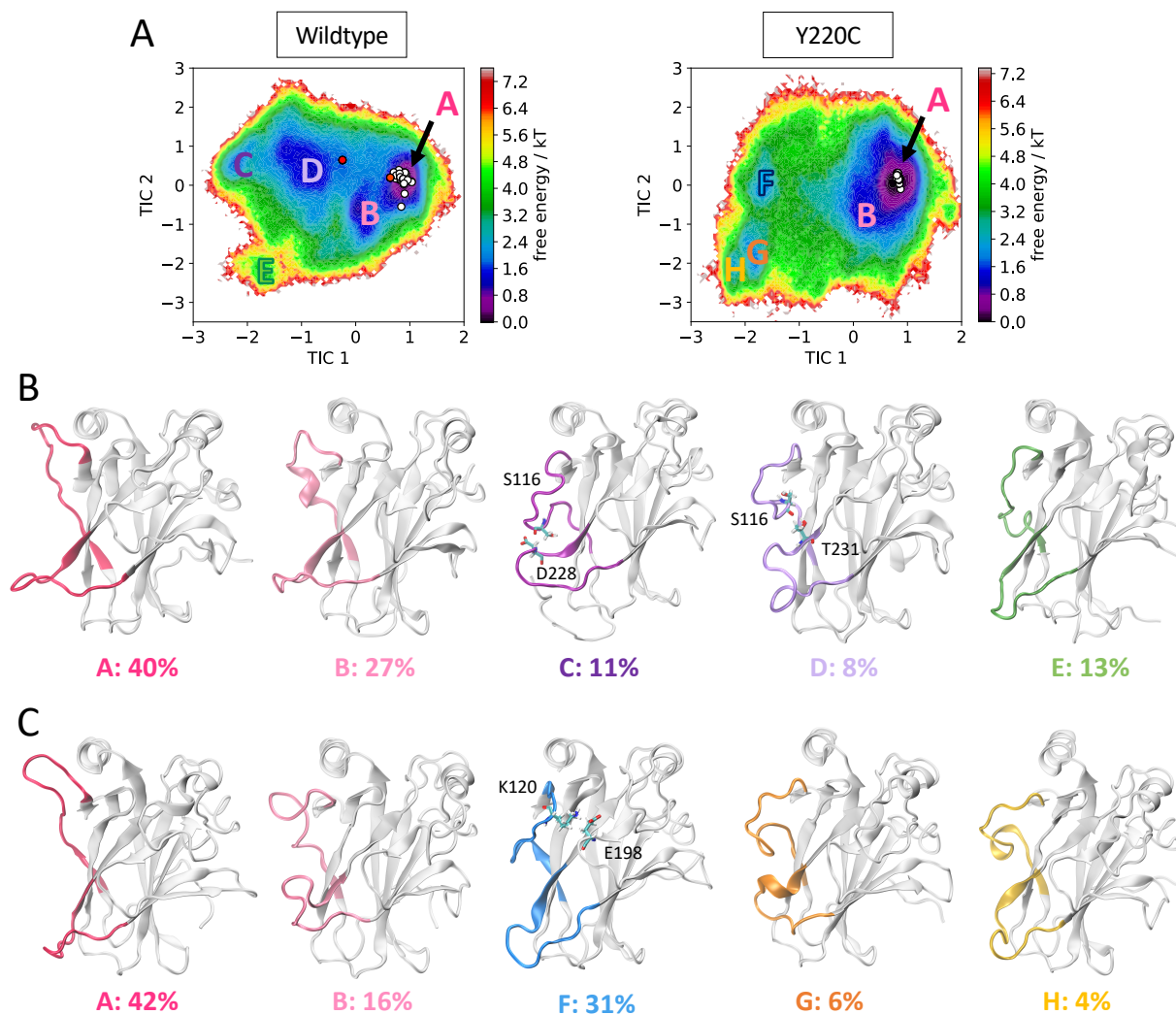


Figure 5.3. L1-centered MSM. (a) Free energy landscape of wildtype (left) and Y220C (right) in terms of the features that describe L1 relative dynamics. Location of metastable states are indicated with letters from A to H. Experimentally resolved DBD structures (X-ray crystallography and NMR) are indicated as white (extended L1 conformation) and red (recessed L1) circles. (b) Conformations from each of the wildtype metastable states. Equilibrium populations are indicated. (c) Y220C metastable states.

Finally, we observe a slight destabilization of states located at low values of TICs 1 and 2 in the Y220C system, which display loop L1 in extremely-recessed conformations (equilibrium population of 13% for wildtype state E and 10% for Y220C states G and H). There are no persistent

L1-L6 interactions in these states. A helical content in loop L6 of Y220C state G seems to be promoted by an inter-L6 hydrogen bond between Ser227 and Thr231.

Dynamics and druggability of loop L6

The significance of loop L6 dynamics suggested by the tICA analysis and its effect on the conformational ensemble of wildtype and Y220C prompted us to consider its conformational plasticity in more detail. Figure 5.4 shows the free energy landscape of the wildtype and Y220C systems now in terms of the tICA components calculated from the 17 previously identified features that include the L6 anchors. For comparison, we also overlay the corresponding coordinates of all experimentally-resolved structures (by X-ray crystallography and NMR) of wildtype and Y220C p53. It is striking how all the previously identified structures are confined to a small area of the graph, and the simulations suggest the existence of novel protein conformations that remain unexplored to date and could be potentially targeted for drug discovery.

All crystal structures align with the wildtype low energy well. The mutation, however, alters the dynamics of this loop and leads to the stabilization of multiple alternative loop L6 conformations, including two mutant-exclusive wells at high values of tIC1. Using Hidden Markov Models to kinetically-coarse grain the microstates results in five metastable states each for the wildtype and Y220C systems (Supplementary Figure 5.S6). The two most populated wildtype metastable states at equilibrium remain significant states in the Y220C ensemble, albeit with changes to their relative equilibrium population and rate of transitions. Three low-populated wildtype states are abrogated by the mutation, while we observe the formation of two Y220C-exclusive metastable states. The conformational differences between the highly populated states,

their implications for rationalizing the mutation effect on p53 function and potential for drug discovery are explored in more detail below.

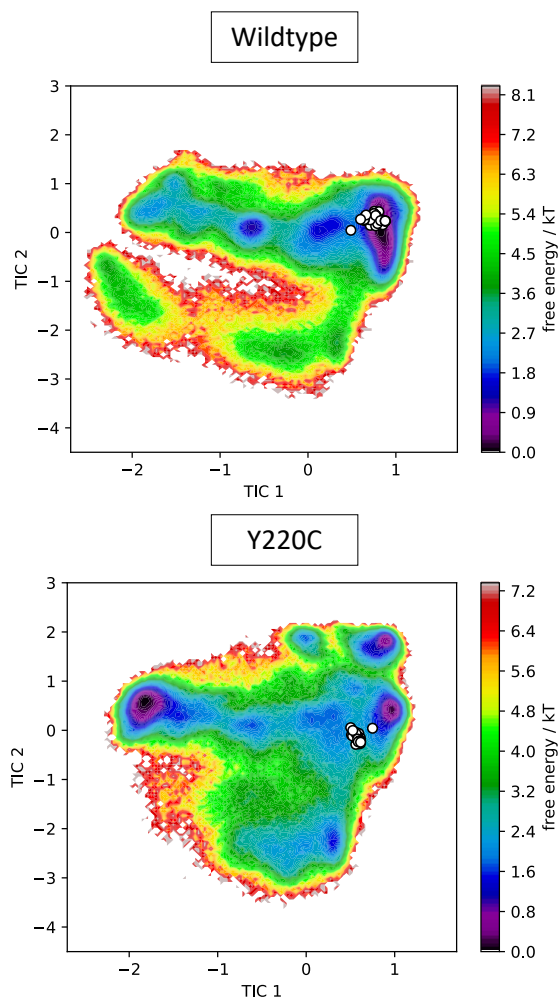


Figure 5.4. Free energy landscape of wildtype and Y220C systems in terms of L6 features. Experimentally resolved DBD structures (X-ray crystallography and NMR) are indicated as white circles.

The mutation induces stabilization of extended L6 conformation

The most populated metastable state in the wildtype ensemble, accounting for over 50% of the population at equilibrium, corresponds to loop L6 in a recessed conformation similar to that observed by NMR and X-ray crystallography (Figure 5.4a). This organization of the loop allows

for the formation of a crevice in between loops L6 and S3/S4 upon the substitution of the bulky tyrosine for the much smaller cysteine residue, which results in the crevice currently being targeted for p53 rescue³⁵⁻⁴⁰.

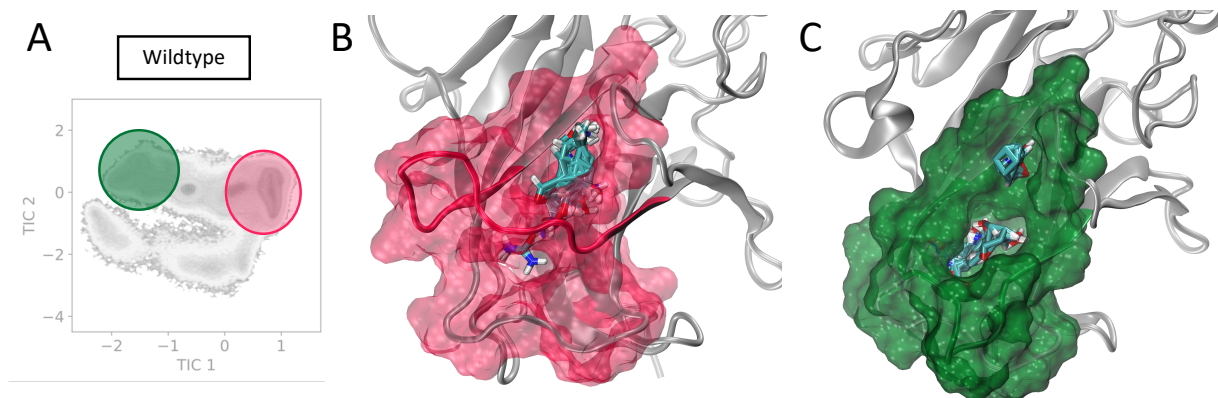


Figure 5.5. L6-centered MSM. (a) Common wildtype and mutant L6 metastable states, which exhibit recessed (pink) and extended (green) loop conformations. Their location on the wildtype free energy landscape is shown. (b) Representation of the cryptic channel spanning loop L6 in the recessed metastable state. FTMap probes indicating hotspots for drug binding are shown in licorice. (c) Representation of the novel L6-extended pocket and solvent mapping results performed by FTMap.

In several of the mutant frames belonging to this metastable state we observed the opening of a transient channel through loop L6, connecting the crevice to another area of the protein surface. This cryptic pocket has been identified previously by Fersht and co-workers using molecular dynamics simulations³⁹, and in agreement with their studies, we find it to exhibit promising druggable characteristics (as suggested by FTMap solvent mapping analysis, Figure 5.5b). Exploitation of this channel by small molecules could improve the potency of rescue drugs and increase specificity towards mutant p53, as the channel is unavailable in the wildtype simulations due to the larger volume occupied by the tyrosine residue.

Besides this well-characterized state, the simulations and MSMs evidence the existence of an additional significantly populated state in the wildtype ensemble at equilibrium. This metastable state, corresponding to 19% of the wildtype population and 24% of the Y220C ensemble, exhibits loop L6 in a previously unknown extended conformation (green state in Figure 5.6). In this conformation, the crevice underneath L6 typically targeted for Y220C rescue is closed. However, visual inspection identified the formation of another cryptic pocket nestled within this loop, promoted by the extended conformation of loop L6. Similar to the mutant-induced crevice, this pocket is only evident in the Y220C simulations due to the presence of the less bulky cysteine in its center. The entrance of the cavity in this case faces “up” relative to the loop, in the direction of the DNA binding surface, and corresponds to a relatively deep hydrophobic pocket with opportunities for hydrogen bonding interaction, as well as other hydrophilic interactions in the more solvent-exposed region above loop L6 (Figure 5.5c).

Several hydrogen bonds between L6 and S3/S4, the loop directly “below” it, are found to be established for longer fractions of the simulation in the mutant state (with increases of up to 100x in persistence time) and suggest possible interactions promoting the extended conformation (Supplementary Table 5.S3). Further indication of the stabilization of the extended conformation promoted by the mutation is given by the calculation of mean first passage times (MFPT) between these metastable states: The mutation decreases the mean first passage time from the recessed to the extended L6 conformation by a factor of more than 2, resulting in a faster transition, while the mean first passage time out of the extended conformation and into the recessed increases by 1.5 (Figure 5.6).

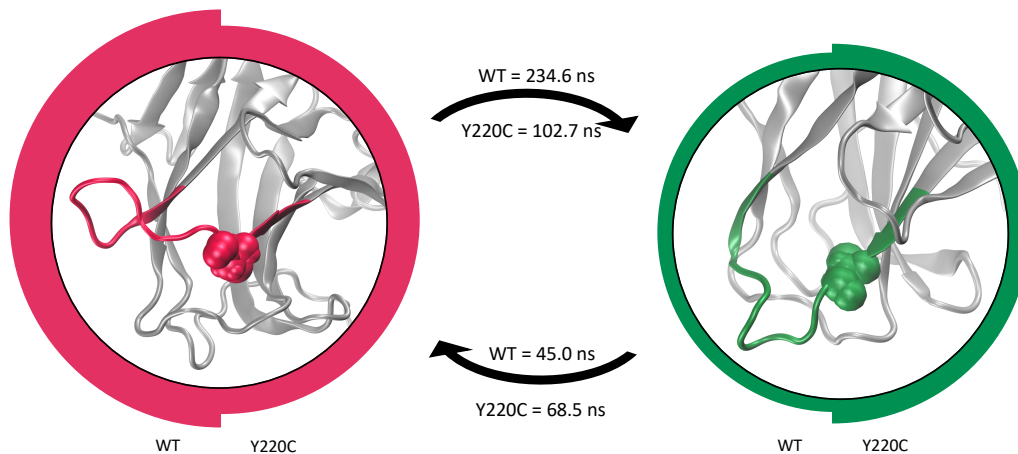


Figure 5.6. Equilibrium population and mean first passage times (MFPT) for the two major wildtype and Y220C metastable states. The images at the center of the circles represent the respective state L6 conformations, with the mutated residues highlighted. Thickness of the circle edge is proportion to the equilibrium population in the respective system (wildtype on the left, Y220C on the right). MFPTs of the transitions are indicated above (for wildtype) and below (for Y220C) the respective arrows.

Characterization of mutant-exclusive metastable states

Finally, our long-timescale exploration of the Y220C mutant dynamics evidenced the sampling of two mutant-exclusive states (Figure 5.7). Jointly, these metastable states account for 33% of the relative Y220C ensemble population, a significant portion of the conformational ensemble that opens up promising avenues for specific therapeutic opportunities. In these states the loop L6 shows a similar extended conformation to the novel metastable state described above, but with a “sideways” bend likely promoted by a Thr54-Pro127 interaction (Supplementary Figure 5.S7). This bend disrupts slightly the cryptic pocket identified in the fully extended L6 conformation, resulting in a smaller and shallower cavity, but also leads to the formation of a channel across loop L6 and underneath the mutation site which reaches across to the protein surface at a different side (Figure 5.S7b). Transitions into or out of these states constitute the slowest process in the Y220C MSM, with a timescale of approximately 1.2 μ s. While the identified

pocket conformations show small pocket volumes and haven't been shown to be druggable by computational solvent mapping, this lays the foundation for further exploration of these mutant-exclusive states.

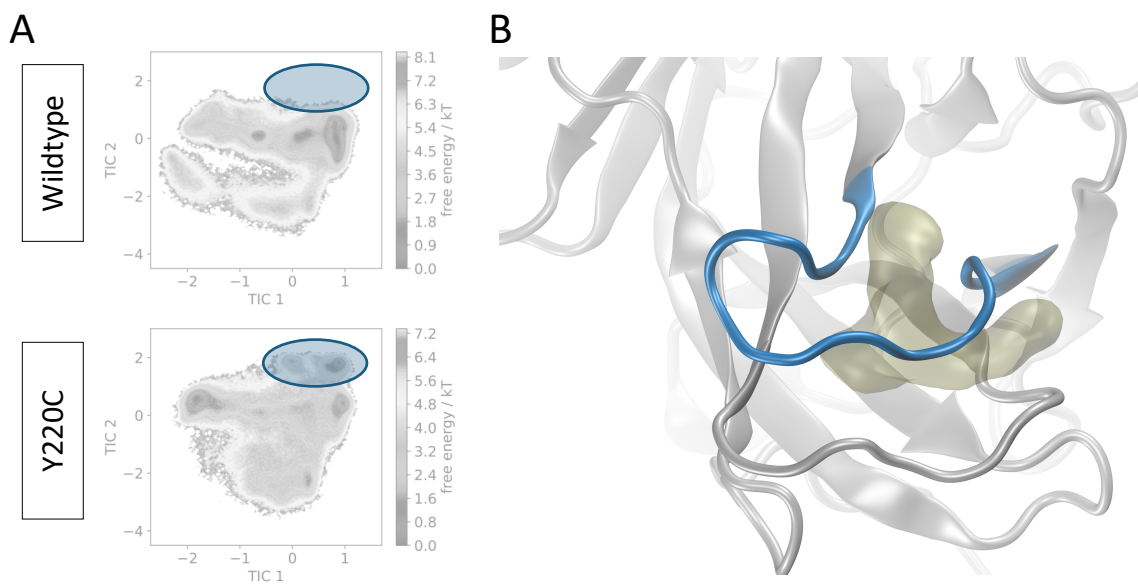


Figure 5.7. (a) Indication of mutant-exclusive states in the free energy landscape. (b) Representation of L6 “bent” conformation seen in the mutant exclusive state.

Remarkably, our models suggest a molecular explanation to the rescuing effect observed by the initial Y220C hit compounds: since in the mutant the recessed L6 conformation is slightly destabilized (33% of the Y220C population versus 50% for wildtype) with a preference for the extended conformations (Figure 5.6 green state and Figure 5.7), binding of a small molecule into the crevice underneath L6 should prevent the transition towards the extended conformations and could lead to a shift in the equilibrium towards a wildtype-like, recessed loop conformational ensemble. Additionally, since the investigation of the full p53 conformational flexibility suggest a high degree of correlation between L1 and L6 dynamics (Figures 5.2 and 5.3), this could further indicate a functional link between L6 conformation and p53 function.

5.5 Conclusions

Our combined tICA and MSM approach proved the existence of a novel dynamic loop, namely loop L6, that exhibits motions at longer timescales than other characterized structural motifs and presents potential for the rationalization of mutational effects on p53 function and for mutant-rescue therapeutic opportunities. The conformational landscape suggests some degree of allostery between L6 and the functionally-important loop L1, likely promoted by hydrogen bonds formed when both loops are in the recessed conformation and thus in close proximity to each other.

The Y220C mutation, which characterizes one of the most common cancer mutants, is located at the N terminus of L6, and we find that the mutation promotes the stabilization of novel protein conformations, which exhibit loop L6 in a novel extended state instead of the only other characterized and targeted recessed L6 conformation. The stabilization of the extended conformation induces the formation of a deep hydrophobic pocket within L6 due to the removal of the bulky tyrosine, as well as the population of two mutant-exclusive states that could be promising avenues for mutant-exclusive therapies.

In summary, the comparison of the dynamics of wildtype and mutant p53 DBD's using MD simulations and Markov state models evidenced for the first time the existence of significant motions involving loop L6 and presents applications for mutant-specific drug discovery efforts. We anticipate that this approach will be useful in the study of the conformational ensembles of other p53 cancer mutants or protein targets, as a way to provide atomic-level information on these proteins' motions combined with thermodynamic and kinetic details in tandem with experimental observations.

5.6 Acknowledgements

Chapter 5, in full, is currently being prepared for submission for publication of the material. Barros, E. P., Demir, O., Amaro, R.E. “Towards mutant-specific therapies: Uncovering the dynamical landscape and druggability of p53 DNA binding domain”. The dissertation author is the primary investigator and author of this paper.

5.7 Supplementary Information

Table 5.S1. Stepwise tICA-based selection of features for model building.

Iteration	Number of features	Number of tICs	Correlation cutoff	Constraints for next round
0	18,336			Remove pairs located $< 3\text{\AA}$ or $> 10\text{\AA}$ apart in all frames
1	7,183			Remove pairs with distance variance < 0.05
2	2,225			Remove pairs involving terminal residues
3	729	315	0.4	
4	499	122	0.5	
5	354	91	0.6	
6	194	57	0.6	
7	90	29	-	Remove features that involve residues close to termini
8	82	26	-	Remove similar pairs
9	35	16	0.75	Remove similar pairs
Final	24	13		

Table 5.S2. Pairs used for featurization of the simulations for model construction

Member 1 (Anchor residue)	Member 2
Ser116	Leu145
Ser116	Val147
Ser116	Thr150
Ser116	Tyr220
Ser116	Cys229
Ser116	Gly279
Ser116	Arg280
Pro223	Gly112
Pro223	Leu114
Pro223	Val143
Pro223	Leu145
Pro223	Thr230
Glu224	Pro153
Glu224	Gly154
Glu224	Cys229
Glu224	Ser260
Glu224	Ser261
Gly226	Thr155
Gly226	Arg156
Gly226	Pro219
Gly226	Tyr220
Gly226	Glu221
Gly226	Glu258
Gly226	Ser260

Table 5.S3. Persistence of L6-S3/S4 hydrogen bonds (in % of frames in the simulation)

Donor atom	Acceptor atom	Wildtype	Y220C
Thr149 - N	Gly225 - O	0	0.7
Thr149 - N	Asp227 – OD1	1.4	2.4
Thr149 - N	Asp227 – OD2	1.3	3.19
Cys219 – N	Thr154 - O	97.0	85.1
Ser226 – N	Thr149 – OG1	7.3	4.9
Asp227 - N	Thr149 – OG1	0.2	1.9
Thr149 – OG1*	Pro222 – O*	0.09	9.0
Thr149 – OG1	Val224 – O	0.02	1.0
Thr149 – OG1	Gly225 – O	0.04	1
Thr149 – OG1	Ser226 – OG	0.06	0.8
Thr149 – OG1	Ser226 – O	1.5	0.9
Thr149 – OG1	Asp227 – OD1	3.7	6.6
Thr149 – OG1	Asp227 – OD2	3.8	7.3
Thr149 – OG1	Asp227 – O	0.07	0.9
Thr154 – OG1	Cys219 - O	0.2	5.8
Ser226 - OG	Asp147 - O	0.01	1.1
Ser226 - OG	Thr54149 – OG1	0.3	1.0

* Interaction formed in the mutant-exclusive “sideways-bent” extended state

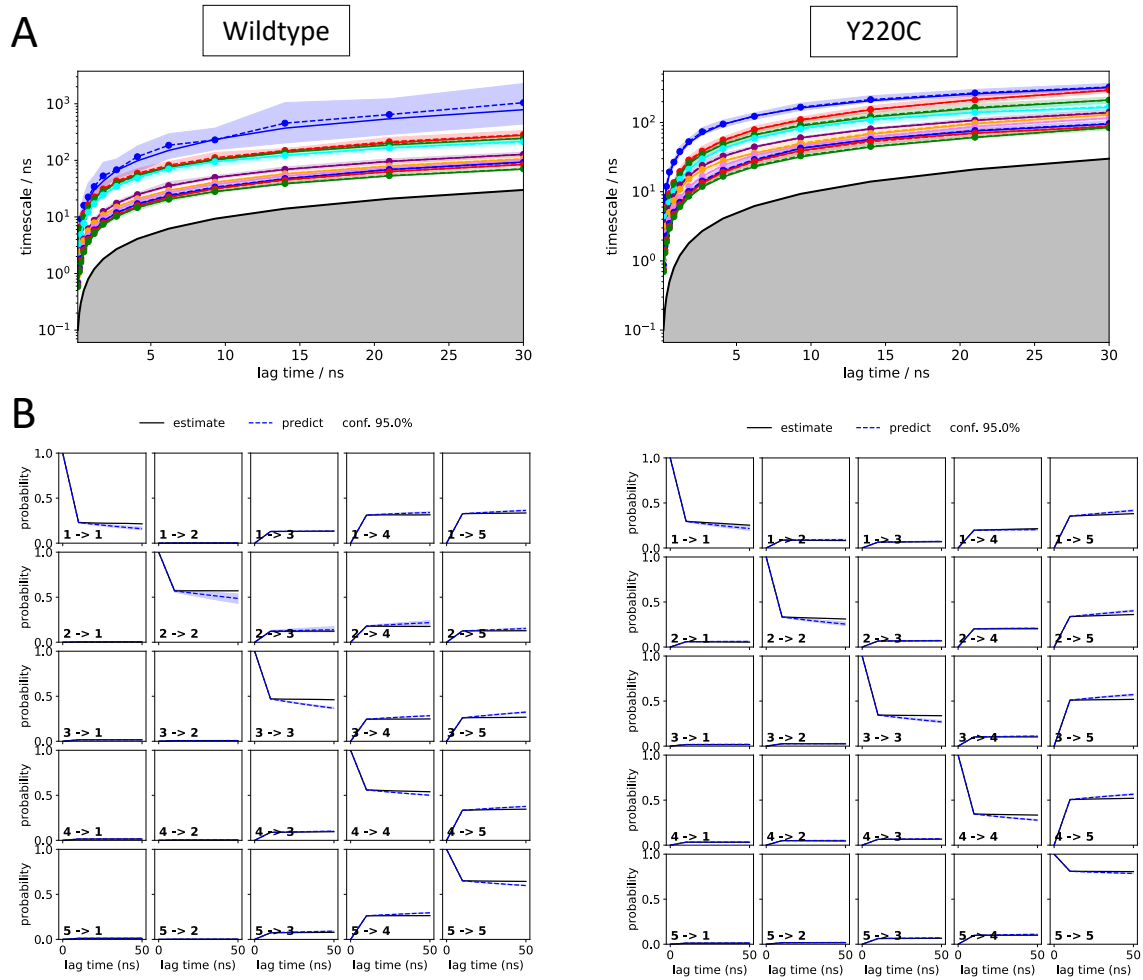


Figure 5.S1. L1 model validation analysis: (a) Implied timescale plots and (b) Chapman-Kolmogorov tests.

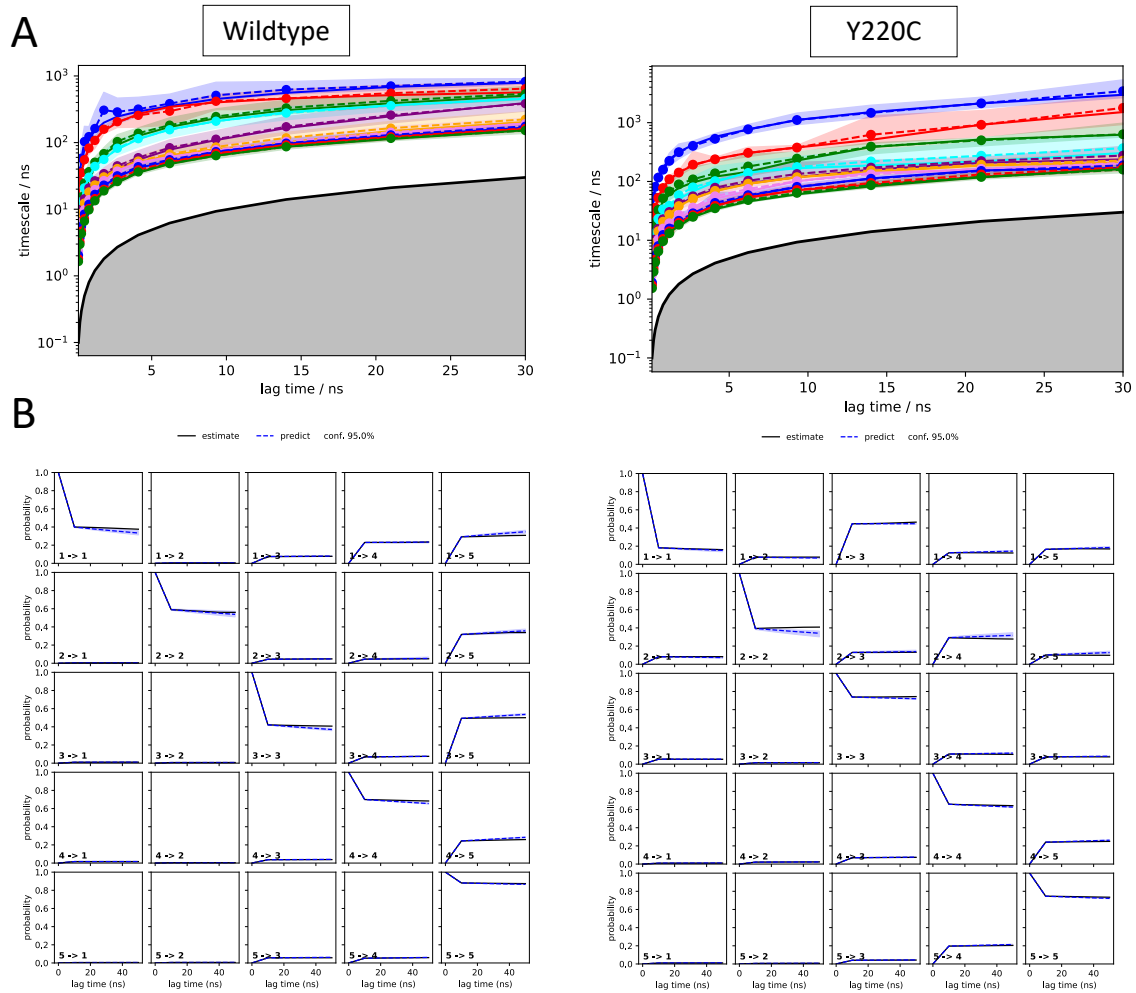


Figure 5.S2. L6 model validation analysis: (a) Implied timescale plots and (b) Chapman-Kolmogorov tests.

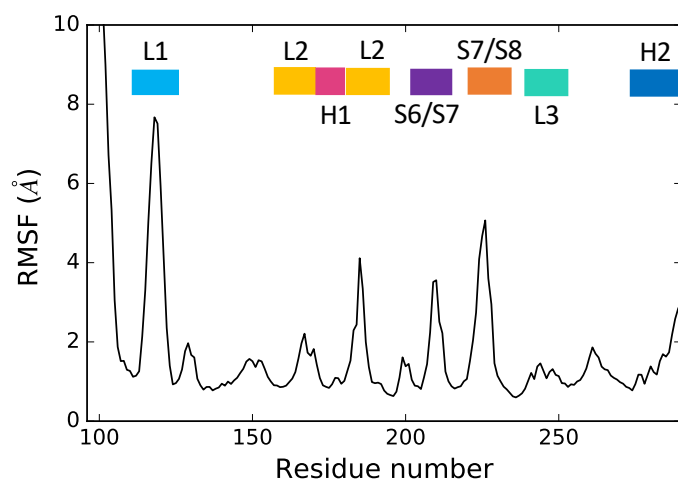
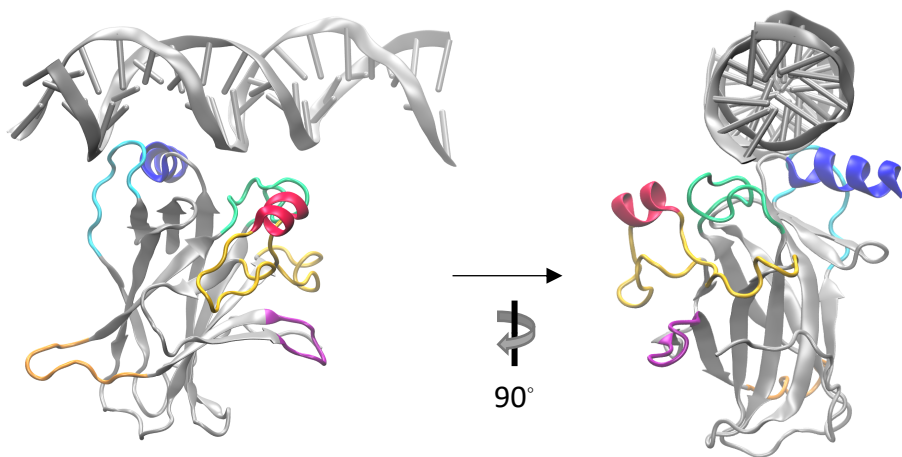


Figure 5.S3. Alpha carbon RMSF

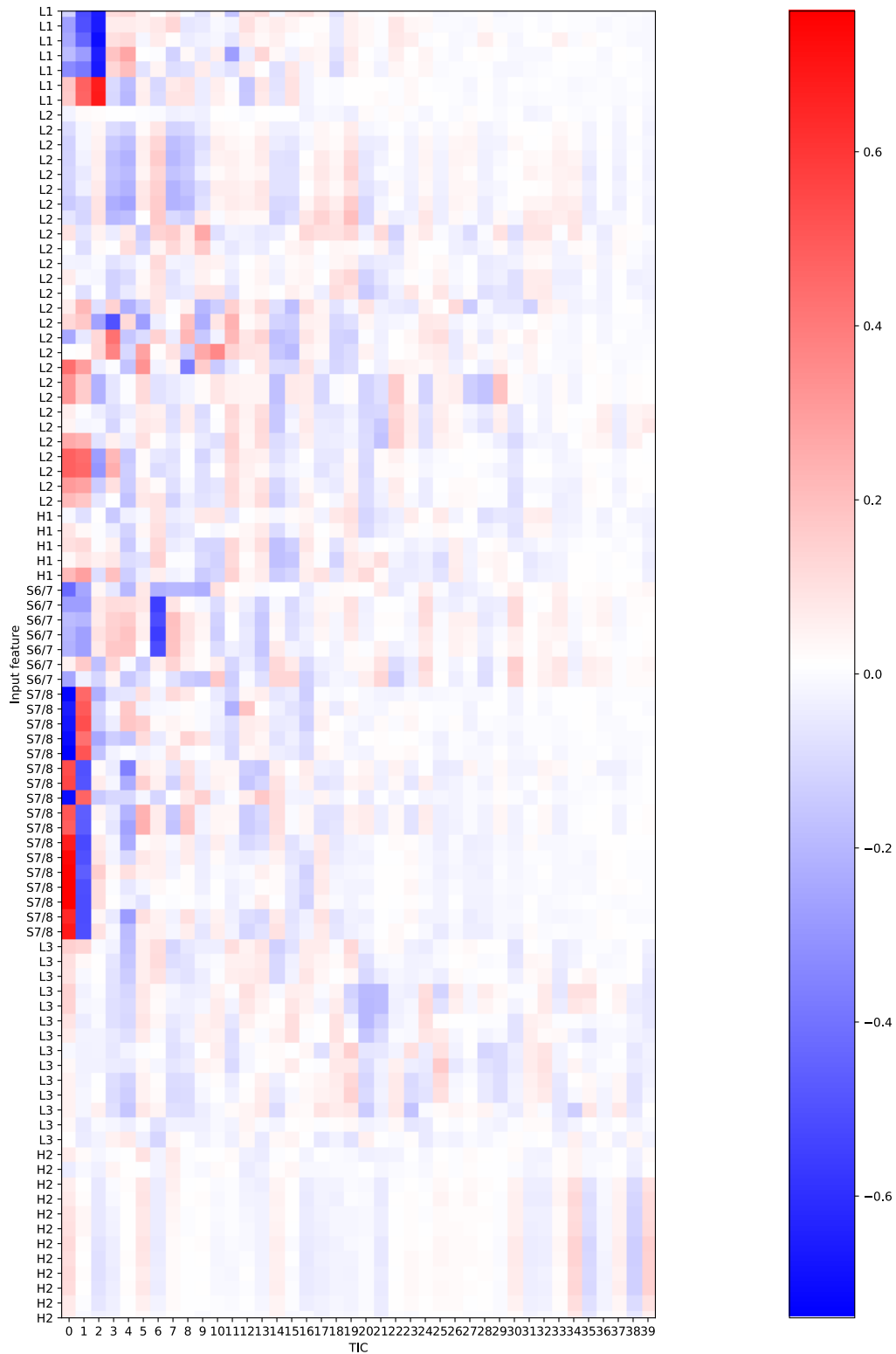


Figure 5.S4. tICA correlation for features incorporating functionally-important motifs in the protein (H1, H2, L2, L3, S6/7) in addition to L1 and L6.

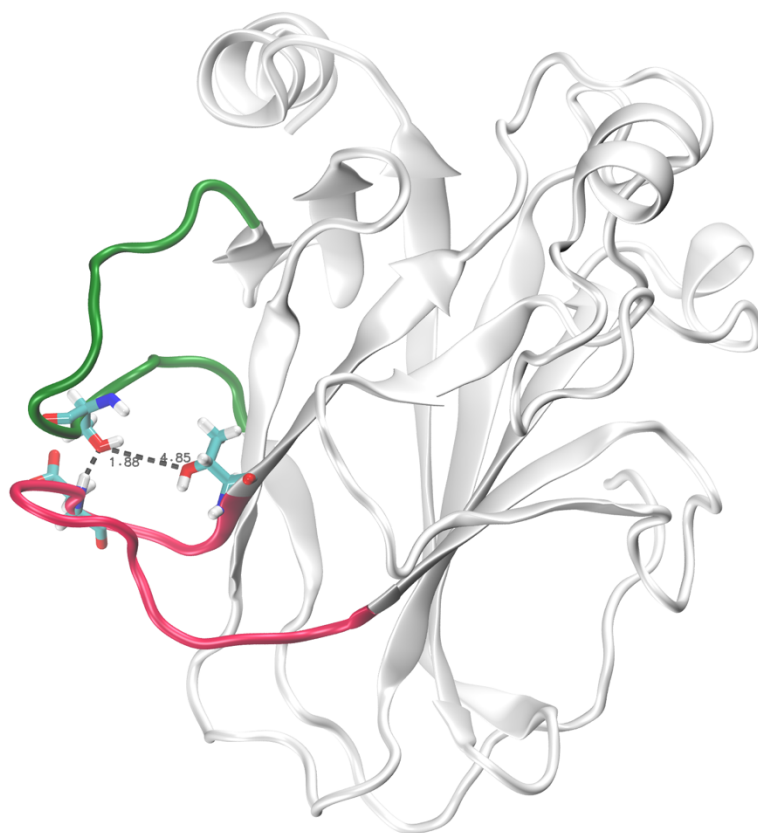


Figure 5.S5. Example of frame exhibiting most stable intra-loop hydrogen bonds, involving Ser116 in L1 and Asp228 or Thr231 in L6.

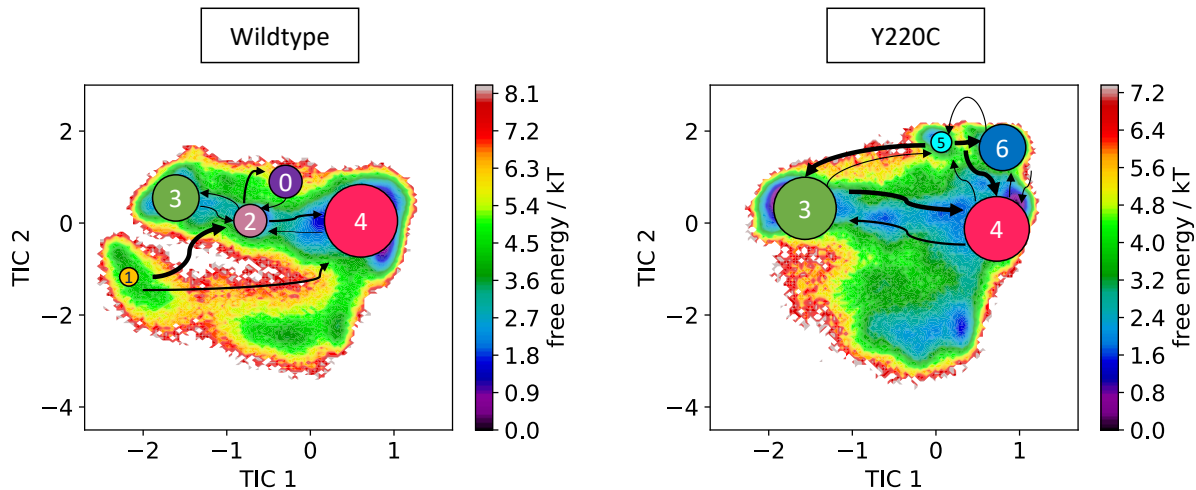


Figure 5.S6. Metastable states identified via Hidden Markov models overlaid over wildtype and Y220C L6-features free energy landscape.

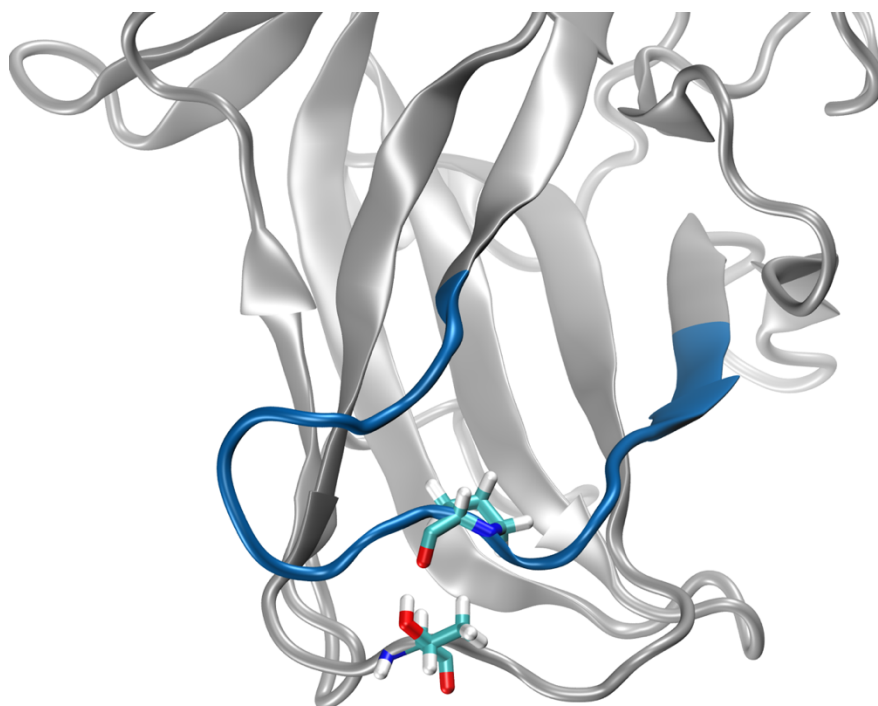


Figure 5.S7. Representation of the Thr149-Pro222 interaction thought to stabilize the bent L6 conformation observed in the mutant-exclusive states.

5.8 References

- (1) Biegging, K. T.; Attardi, L. D. Deconstructing P53 Transcriptional Networks in Tumor Suppression. *Trends Cell Biol.* **2012**, *22* (2), 97–106. <https://doi.org/10.1016/j.tcb.2011.10.006>.
- (2) Lujambio, A.; Akkari, L.; Simon, J.; Grace, D.; Tschaharganeh, D. F.; Bolden, J. E.; Zhao, Z.; Thapar, V.; Joyce, J. A.; Krizhanovsky, V.; et al. Non-Cell-Autonomous Tumor Suppression by P53. *Cell* **2013**, *153*, 449–460.
- (3) Vogelstein, B.; Lane, D.; Levine, A. J. Surfing the P53 Network. *Nature* **2000**, *408*, 307–310.
- (4) Soussi, T.; Dehouche, K.; Bérout, C. P53 Website and Analysis of P53 Gene Mutations in Human Cancer: Forging a Link Between Epidemiology and Carcinogenesis. *Hum. Mutat.* **2000**, *15*, 105–113.
- (5) Olivier, M.; Eeles, R.; Hollstein, M.; Khan, M. A.; Harris, C. C.; Hainaut, P. The IARC TP53 Database: New Online Mutation Analysis and Recommendations to Users. *Hum. Mutat.* **2002**, *19*, 607–614. <https://doi.org/10.1002/humu.10081>.
- (6) Soussi, T.; Bérout, C. Assessing TP53 Status in Human Tumours to Evaluate Clinical Outcome. *Nat. Rev. Cancer* **2001**, *1*, 233–240. <https://doi.org/10.1038/35106009>.
- (7) Ventura, A.; Kirsch, D. G.; Mclaughlin, M. E.; Tuveson, D. A.; Grimm, J.; Lintault, L.; Newman, J.; Reczek, E. E.; Weissleder, R.; Jacks, T. Restoration of P53 Function Leads to Tumour Regression in Vivo. *Nature* **2007**, *445*, 661–665. <https://doi.org/10.1038/nature05541>.
- (8) Parrales, A.; Iwakuma, T. Targeting Oncogenic Mutant P53 for Cancer Therapy. *Front. Oncol.* **2015**, *5*, 288. <https://doi.org/10.3389/fonc.2015.00288>.
- (9) Martins, C. P.; Brown-Swigart, L.; Evan, G. I. Modeling the Therapeutic Efficacy of P53 Restoration in Tumors. *Cell* **2006**, *127*, 1323–1334. <https://doi.org/10.1016/j.cell.2006.12.007>.
- (10) Selivanova, G.; Wiman, K. G. Reactivation of Mutant P53: Molecular Mechanisms and Therapeutic Potential. *Oncogene* **2007**, *26*, 2243–2254. <https://doi.org/10.1038/sj.onc.1210295>.
- (11) Xue, W.; Zender, L.; Miething, C.; Dickins, R. A.; Hernando, E.; Krizhanovsky, V.; Cordon-Cardo, C.; Lowe, S. W. Senescence and Tumour Clearance Is Triggered by P53 Restoration in Murine Liver Carcinomas. *Nature* **2007**, *445*, 656–660. <https://doi.org/10.1038/nature05529>.
- (12) Freed-Pastor, W. A.; Prives, C. Mutant P53 : One Name , Many Proteins. *Genes Dev.* **2012**, *26*, 1268–1286. <https://doi.org/10.1101/gad.190678.112.1268>.

- (13) Muller, P. A. J.; Vousden, K. H. Mutant P53 in Cancer : New Functions and Therapeutic Opportunities. *Cancer Cell* **2014**, *25* (3), 304–317. <https://doi.org/10.1016/j.ccr.2014.01.021>.
- (14) Sabapathy, K.; Lane, D. P. Therapeutic Targeting of P53: All Mutants Are Equal, but Some Mutants Are More Equal than Others. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 13–30. <https://doi.org/10.1038/nrclinonc.2017.151>.
- (15) Eldar, A.; Rozenberg, H.; Diskin-posner, Y.; Rohs, R.; Shakked, Z. Structural Studies of P53 Inactivation by DNA-Contact Mutations and Its Rescue by Suppressor Mutations via Alternative Protein – DNA Interactions. *Nucleic Acids Res.* **2013**, *41*, 8748–8759. <https://doi.org/10.1093/nar/gkt630>.
- (16) Joerger, A. C.; Fersht, A. R. Structure-Function-Rescue: The Diverse Nature of Common P53 Cancer Mutants. *Oncogene* **2007**, *26*, 2226–2242. <https://doi.org/10.1038/sj.onc.1210291>.
- (17) Bullock, A. N.; Henckel, J.; Fersht, A. R. Quantitative Analysis of Residual Folding and DNA Binding in Mutant P53 Core Domain: Definition of Mutant States for Rescue in Cancer Therapy. *Oncogene* **2000**, *19*, 1245–1256. <https://doi.org/10.1038/sj.onc.1203434>.
- (18) Wilcken, R.; Wang, G. Z.; Boeckler, F. M.; Fersht, A. R. Kinetic Mechanism of P53 Oncogenic Mutant Aggregation and Its Inhibition. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 13584–13589. <https://doi.org/10.1073/pnas.1211550109>.
- (19) Wang, G. Z.; Fersht, A. R. Multisite Aggregation of P53 and Implications for Drug Rescue. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E2634–E2643. <https://doi.org/10.1073/pnas.1700308114>.
- (20) Bykov, V. J. N.; Issaeva, N.; Zache, N.; Shilov, A.; Hulterantz, M.; Bergman, J.; Selivanova, G.; Wiman, K. G. Reactivation of Mutant P53 and Induction of Apoptosis in Human Tumor Cells by Maleimide Analogs. *J. Biol. Chem.* **2005**, *280*, 30384–30391. <https://doi.org/10.1074/jbc.M501664200>.
- (21) Beraza, N.; Trautwein, C. Restoration of P53 Function: A New Therapeutic Strategy to Induce Tumor Regression? *Hepatology* **2007**, *45*, 1578–1579. <https://doi.org/10.1002/hep.21789>.
- (22) Wassman, C. D.; Baronio, R.; Demir, Ö.; Wallentine, B. D.; Chen, C.-K.; Hall, L. V; Salehi, F.; Lin, D.; Chung, B. P.; Hatfield, G. W.; et al. Computational Identification of a Transiently Open L1/S3 Pocket for Reactivation of Mutant P53 ". *Nat. Commun.* **2013**, *4*, 1407. <https://doi.org/10.1038/ncomms2361>.
- (23) Russo, D.; Ottaggio, L.; Foggetti, G.; Masini, M.; Masiello, P.; Fronza, G.; Menichini, P. PRIMA-1 Induces Autophagy in Cancer Cells Carrying Mutant or Wild Type P53. *Biochim. Biophys. Acta* **2013**, *1833*, 1904–1913. <https://doi.org/10.1016/j.bbamcr.2013.03.020>.
- (24) Izetti, P.; Hautefeuille, A.; Abujamra, A. L.; De Farias, C. B.; Giacomazzi, J.; Alemar, B.;

- Lenz, G.; Roesler, R.; Schwartzmann, G.; Osvaldt, A. B.; et al. PRIMA-1, a Mutant P53 Reactivator, Induces Apoptosis and Enhances Chemotherapeutic Cytotoxicity in Pancreatic Cancer Cell Lines. *Invest. New Drugs* **2014**, *32*, 783–794. <https://doi.org/10.1007/s10637-014-0090-9>.
- (25) Bykov, V. J. N.; Wiman, K. G. Mutant P53 Reactivation by Small Molecules Makes Its Way to the Clinic. *FEBS Lett.* **2014**, *588* (16), 2622–2627. <https://doi.org/10.1016/j.febslet.2014.04.017>.
- (26) Perdrix, A.; Najem, A.; Saussez, S.; Awada, A.; Journe, F.; Ghanem, G.; Krayem, M. PRIMA-1 and PRIMA-1Met (APR-246): From Mutant/Wild Type P53 Reactivation to Unexpected Mechanisms Underlying Their Potent Anti-Tumor Effect in Combinatorial Therapies. *Cancers (Basel)*. **2017**, *9*, 172. <https://doi.org/10.3390/cancers9120172>.
- (27) Zache, N.; Lambert, J. M. R.; Wiman, K. G.; Bykov, V. J. N. PRIMA-1MET Inhibits Growth of Mouse Tumors Carrying Mutant P53. *Cell. Oncol.* **2008**, *30*, 411–418. <https://doi.org/10.3233/CLO-2008-0440>.
- (28) Zache, N.; Lambert, J. M. R.; Rökaeus, N.; Shen, J.; Hainaut, P.; Bergman, J.; Wiman, K. G.; Bykov, V. J. N. Mutant P53 Targeting by the Low Molecular Weight Compound STIMA-1. *Mol. Oncol.* **2008**, *2*, 70–80. <https://doi.org/10.1016/j.molonc.2008.02.004>.
- (29) Brown, C. J.; Lain, S.; Verma, C. S.; Fersht, A. R.; Lane, D. P. Awakening Guardian Angels: Drugging the P53 Pathway. *Nat. Rev. Cancer* **2009**, *9*, 862–873. <https://doi.org/10.1038/nrc2763>.
- (30) Lambert, J. M. R.; Gorzov, P.; Veprintsev, D. B.; Söderqvist, M.; Segerbäck, D.; Bergman, J.; Fersht, A. R.; Hainaut, P.; Wiman, K. G.; Bykov, V. J. N. PRIMA-1 Reactivates Mutant P53 by Covalent Binding to the Core Domain. *Cancer Cell* **2009**, *15*, 376–388. <https://doi.org/10.1016/j.ccr.2009.03.003>.
- (31) Zandi, R.; Selivanova, G.; Christensen, C. L.; Gerds, T. A.; Willumsen, B. M.; Poulsen, H. S. PRIMA-1Met/APR-246 Induces Apoptosis and Tumor Growth Delay in Small Cell Lung Cancer Expressing Mutant P53. *Clin. Cancer Res.* **2011**, *17*, 2830–2841. <https://doi.org/10.1158/1078-0432.CCR-10-3168>.
- (32) Yu, X.; Vazquez, A.; Levine, A. J.; Carpizo, D. R. Allele-Specific P53 Mutant Reactivation. *Cancer Cell* **2012**, *21*, 614–625. <https://doi.org/10.1016/j.ccr.2012.03.042>.
- (33) Lehmann, S.; Bykov, V. J. N.; Ali, D.; Andreñ, O.; Cherif, H.; Tidefelt, U.; Uggla, B.; Yachnin, J.; Juliusson, G.; Moshfegh, A.; et al. Targeting P53 in Vivo: A First-in-Human Study with P53-Targeting Compound APR-246 in Refractory Hematologic Malignancies and Prostate Cancer. *J. Clin. Oncol.* **2012**, *30*, 3633–3639. <https://doi.org/10.1200/JCO.2011.40.7783>.
- (34) Liu, X.; Wilcken, R.; Joerger, A. C.; Chuckowree, I. S.; Amin, J.; Spencer, J.; Fersht, A. R. Small Molecule Induced Reactivation of Mutant P53 in Cancer Cells. **2013**, *41* (12), 6034–6044. <https://doi.org/10.1093/nar/gkt305>.

- (35) Joerger, A. C.; Ang, H. C.; Fersht, A. R. Structural Basis for Understanding Oncogenic P53 Mutations and Designing Rescue Drugs. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 15056–15061. <https://doi.org/10.1073/pnas.0607286103>.
- (36) Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R. Targeted Rescue of a Destabilized Mutant of P53 by an in Silico Screened Drug. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 10360–10365.
- (37) Basse, N.; Kaar, J. L.; Settanni, G.; Joerger, A. C.; Rutherford, T. J.; Fersht, A. R. Toward the Rational Design of P53-Stabilizing Drugs: Probing the Surface of the Oncogenic Y220C Mutant. *Chem. & Biol.* **2010**, *17*, 46–56. <https://doi.org/10.1016/j.chembiol.2009.12.011>.
- (38) Wilcken, R.; Liu, X.; Zimmermann, M. O.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Halogen-Enriched Fragment Libraries as Leads for Drug Rescue of Mutant P53. *J. Am. Chem. Soc.* **2012**, *134*, 6810–6818. <https://doi.org/10.1021/ja301056a>.
- (39) Joerger, A. C.; Bauer, M. R.; Wilcken, R.; Boeckler, F. M.; Spencer, J.; Fersht, A. R. Exploiting Transient Protein States for the Design of Small-Molecule Stabilizers of Mutant P53. *Struct. Des.* **2015**, *23* (12), 2246–2255. <https://doi.org/10.1016/j.str.2015.10.016>.
- (40) Bauer, M. R.; Jones, R. N.; Tareque, R. K.; Springett, B.; Dingler, F. A.; Verduci, L.; Patel, K. J.; Fersht, A. R.; Joerger, A. C.; Spencer, J. A Structure-Guided Molecular Chaperone Approach for Restoring the Transcriptional Activity of the P53 Cancer Mutant Y220C. *Futur. Med. Chem.* **2019**, *11* (19), 2491–2504. <https://doi.org/10.4155/fmc-2019-0181>.
- (41) Demir, Ö.; Jeong, P. U.; Amaro, R. E. Full-Length P53 Tetramer Bound to DNA and Its Quaternary Dynamics. **2016**, No. July, 1–10. <https://doi.org/10.1038/onc.2016.321>.
- (42) Wagner, R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. **2016**. <https://doi.org/10.1021/acs.chemrev.5b00631>.
- (43) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48* (2), 414–422. <https://doi.org/10.1021/ar5002999>.
- (44) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99–105. <https://doi.org/10.1016/j.ymeth.2010.06.002>.
- (45) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105. <https://doi.org/10.1063/1.3565032>.
- (46) Bowman, G. R.; Pande, V. S. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*.
- (47) Chodera, J. D.; Noe, F. Markov State Models of Biomolecular Conformational Dynamics.

- Curr. Opin. Bstructural Biol.* **2014**, *25*, 135–144. <https://doi.org/10.1016/j.sbi.2014.04.002>.
- (48) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; T.E. Cheatham, I.; Darden, T. A.; Duke, R. E.; Gohlke, H.; et al. AMBER 14. University of California, San Francisco 2014.
- (49) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys* **1983**, *79*, 926–932.
- (50) Pang, Y. P. Novel Zinc Protein Molecular Dynamics Simulations: Steps toward Antiangiogenesis for Cancer Treatment. *J. Mol. Model.* **1999**, *5*, 196–202. <https://doi.org/10.1007/s008940050119>.
- (51) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713. <https://doi.org/10.1021/acs.jctc.5b00255>.
- (52) Malmstrom, R. D.; Kornev, A. P.; Taylor, S. S.; Amaro, R. E. Allosteric through the Computational Microscope: CAMP Activation of a Canonical Signalling Domain. *Nat. Commun.* **2015**, *6* (May), 7588. <https://doi.org/10.1038/ncomms8588>.
- (53) Pierce, L. C. T.; Salomon-Ferrer, R.; Augusto F. De Oliveira, C.; McCammon, J. A.; Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 2997–3002. <https://doi.org/10.1021/ct300284c>.
- (54) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuild2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419. <https://doi.org/10.1021/ct200463m>.
- (55) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noe, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542. <https://doi.org/10.1021/acs.jctc.5b00743>.
- (56) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys* **2013**, *139*, 015102. <https://doi.org/10.1063/1.4811489>.
- (57) Durrant, J. D.; Votapka, L.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory Comput.* **2014**, *10*, 5047–5056.
- (58) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand Binding Hot Spots of Proteins. *Nat. Protoc* **2015**, *10*, 733–755. <https://doi.org/10.1038/nprot.2015.043>.The.

- (59) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532.
- (60) Petty, T. J.; Emamzadah, S.; Costantino, L.; Petkova, I.; Stavridi, E. S.; Saven, J. G.; Vauthey, E.; Halazonetis, Thanos, D. An Induced Fit Mechanism Regulates P53 DNA Binding Kinetics to Confer Sequence Specificity. *EMBO J.* **2011**, *30* (11), 2167–2176. <https://doi.org/10.1038/emboj.2011.127>.
- (61) Emamzadah, S.; Tropia, L.; Halazonetis, T. D. Crystal Structure of a Multidomain Human P53 Tetramer Bound to the Natural CDKN1A (P21) P53-Response Element. *Mol. Cancer Res.* **2011**, *9*, 1493–1500. <https://doi.org/10.1158/1541-7786.MCR-11-0351>.
- (62) Lukman, S.; Lane, D. P.; Verma, C. S. Mapping the Structural and Dynamical Features of Multiple P53 DNA Binding Domains : Insights into Loop 1 Intrinsic Dynamics. *PLoS One* **2013**, *8* (11), e80221. <https://doi.org/10.1371/journal.pone.0080221>.
- (63) Lu, Q.; Tan, Y. H.; Luo, R. Molecular Dynamics Simulations of P53 DNA-Binding Domain. *J. Phys. Chem. B* **2007**, *111*, 11538–11545. <https://doi.org/10.1021/jp0742261>.
- (64) Pradhan, M. R.; Siau, J. W.; Kannan, S.; Nguyen, M. N.; Ouaray, Z.; Kwoh, C. K.; Lane, D. P.; Ghadessy, F.; Verma, C. S. Simulations of Mutant P53 DNA Binding Domains Reveal a Novel Druggable Pocket. *Nucleic Acids Res.* **2019**, *47* (4), 1637–1652. <https://doi.org/10.1093/nar/gky1314>.