Washington University in St. Louis

Washington University Open Scholarship

Arts & Sciences Electronic Theses and **Dissertations**

Arts & Sciences

Spring 5-15-2022

Combining computer simulations and deep learning to understand and predict protein structural dynamics

Michael D. Ward Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds



Part of the Computer Engineering Commons

Recommended Citation

Ward, Michael D., "Combining computer simulations and deep learning to understand and predict protein structural dynamics" (2022). Arts & Sciences Electronic Theses and Dissertations. 2729. https://openscholarship.wustl.edu/art_sci_etds/2729

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences Computational and Systems Biology

Dissertation Examination Committee:
Gregory R. Bowman, Chair
Gautam Dantas
Roman Garnett
Alex Holehouse
Janice Robertson

Combining Computer Simulations and Deep Learning to Understand and Predict Protein
Structural Dynamics
by
Michael D. Ward

A dissertation presented to The Graduate School of Washington University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

> August 2022 St. Louis, Missouri

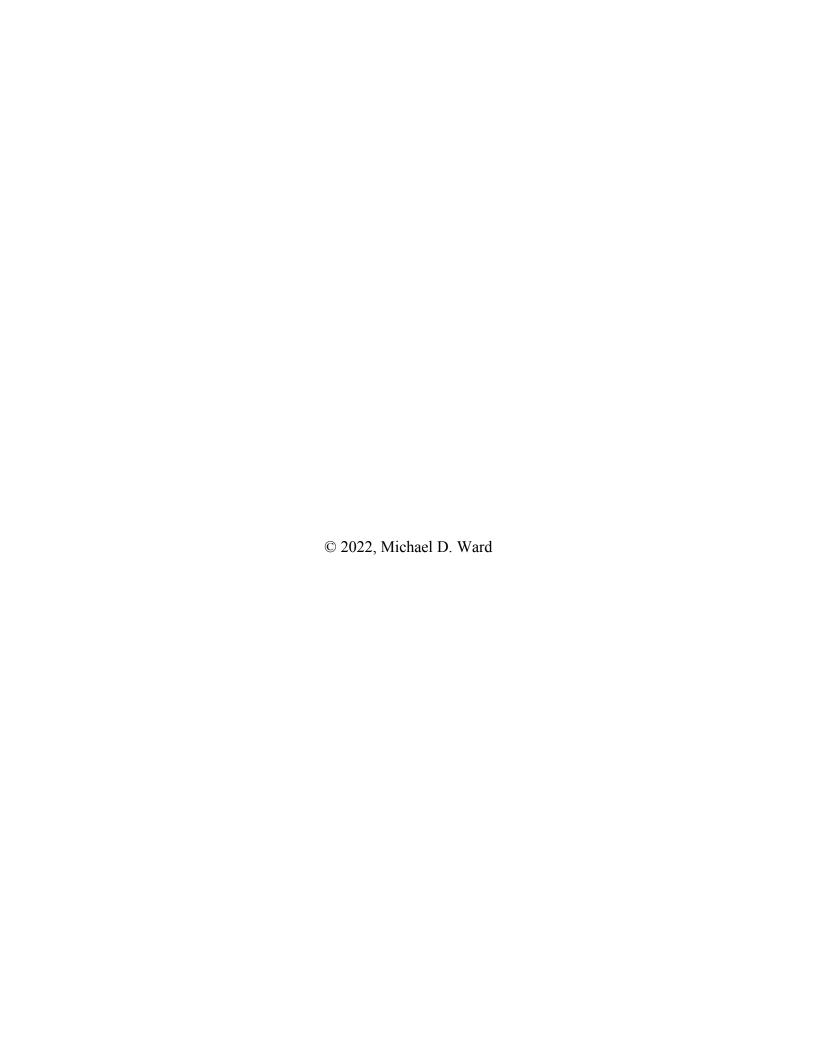


Table of Contents

List of Figures	vi
List of Tables	viii
Acknowledgments	ix
Abstract of the Dissertation	xvi
Chapter 1	1
Introduction	1
1.1 Proteins are the molecular machinery of the living	1
1.2 Molecular Dynamics simulations provide atomistic resolution of a prodynamics	
1.3 Deep learning has made a lasting impact on protein biophysics	6
1.4 Scope of Thesis	10
Bibliography	14
Chapter 2	18
Deep learning the structural determinants of protein biochemica comparing structural ensembles with DiffNets	18
2.1 Preamble	
2.2 Introduction	18
2.3 Results	
2.3.1 The DiffNet Architecture	
2.3.2 The Classification Task Reorganizes the Latent Space to Emphasize Important Structural Features	
2.3.3 Self-Supervised DiffNets Learn Structural Signatures Associated wi	
2.3.4 DiffNets works for other proteins and more divergent sequences	39
2.4 Conclusions	42
2.5 Methods	44
2.5.1 MD Simulations	44
2.5.2 DiffNet Model	
2.5.3 EM algorithm	
2.5.4 Featurization	
2.5.5 Classification Targets	

2.5.6 Neural Network Training	47
2.5.7 Reconstruction Experiment	50
2.5.8 Classification Labels	51
2.5.9 β-lactamase expectation maximization experiment	51
2.5.10 Myosin expectation maximization experiment	52
2.5.11 Code Availability	
2.5.12 Data Availability	53
Bibliography	54
Chapter 3	61
Naturally-ocurring genetic variants in the oxytocin receptor alter receptor	
signaling profiles	61
3.1 Preamble	61
3.2 Introduction	61
3.3 Methods	62
3.3.1 Cell culture	
3.3.2 cDNA constructs	
3.3.3 Ca ²⁺ assays	
3.3.4 Bioluminescence resonance energy transfer (BRET) assays	
3.3.5 Quantitative flow cytometry	
3.3.6 Data processing for Ca ²⁺ , BRET, desensitization, and internalization assays	
3.3.8 DiffNet Analysis	
3.3.9 Markov State Model construction and analysis	
·	
3.4 Results	
3.4.1 Genetic variation occurs in several locations within OXTR	
3.4.3 OXTR variants alter cell surface localization	
3.4.4 V45L, P108A, and E339K impair OXTR desensitization and internalization	
3.4.5 Variants that reduce desensitization and internalization alter OXTR structural	
	75
3.4.6 Conformational changes in V45L and P108A OXTRs disrupt putative β-arrestin b	_
site	_
3.4.7 Structural conformations in V281M OXTR	
3.5 Conclusions	78
Bibliography	91
Chapter 4	95
SARS-CoV-2 Simulations Go Exascale to Predict Dramatic Spike Opening and	1
Cryntic Packets Across the Proteame	95

4.1 Preamble	95
4.2 Introduction	95
4.3 Results	97
4.3.1 To the exascale and beyond!	97
4.3.2 Extreme spike opening reveals cryptic epitopes	98
4.3.3 Cryptic pockets and functional dynamics are present throughout the proteome	
4.4 Conclusions	106
4.5 Methods	107
4.5.1 System preparation	
4.5.2 Adaptive sampling simulations	
4.5.3 Folding@home simulations	
4.5.4 Markov state models	
4.5.5 Spike/ACE2 binding competency	
4.5.6 Cryptic pockets and solvent accessible surface area	
4.5.7 Sequence conservation	
4.5.8 Data availability	
4.5.9 Code availability	
Bibliography	119
Chapter 5	. 126
SARS-CoV2 Nsp16 activation mechanism and a cryptic pocket with pan-coronavirus antiviral potential	. 126
5.1 Preamble	126
5.2 Introduction	126
5.3 Methods	129
5.3.1 System Preparation	129
5.3.2 Adaptive sampling simulations	130
5.3.3 DiffNets	131
5.3.4 Markov State Models	132
5.3.5 Distance and SASA calculations	134
5.3.6 Cryptic pocket detection	134
5.3.7 Sequence Conservation	
5.4 Results	136
5.4.1 Nsp10 promotes opening of Nsp16's SAM- and RNA-binding pockets	136
5.4.2 A cryptic pocket in Nsp16 is a potential therapeutic target	139
5.4.3 Conservation of the cryptic pocket in Nsp16 makes it a promising target for broa spectrum inhibitors	
·	
5.5 Conclusions	
Bibliography	

Chapter 6	152
Predicting cryptic pocket opening from protein structures using graph neuro	
networks	152
6.1 Preamble	152
6.2 Introduction	152
6.3 Results 6.3.1 Predicting cryptic pockets with Geometric Vector Perceptrons 6.3.2 Graph neural networks accurately predict residue level pocket volume changes simulation data 6.3.3 Graph neural networks accurately identify cryptic pockets from experimental structures	153 from 156
6.4 Conclusions	159
Bibliography	160
Chapter 7	161
Conclusions	161
7.1 Main Findings	161
7.2 Future Directions	163
Bibliography	166
Appendices	167
A.1 Appendix to Chapter 2	167 169
A.2 Appendix to Chapter 3	174
A.3 Appendix to Chapter 4	181
A.4 Appendix to Chapter 5	190
Curriculum Vitae	. 198

List of Figures

Figure 2.1 Comparison of autoencoder and DiffNet architectures	22
Figure 2.2 Helix 9 compaction distinguishes structural ensembles	27
Figure 2.3 DiffNets accurately reconstruct protein structures	30
Figure 2.4 Classification task reorganizes DiffNets latent space	31
Figure 2.5 DiffNets learn helix 9 compaction is important for distinguishing variants	37
Figure 2.6 Automated feature detection reveals what DiffNets learn	
Figure 2.7 DiffNets capture known P-loop motions that distinguish myosin isoforms	42
Figure 3.1 Screen identifies <i>OXTR</i> variants that alter oxytocin response in Ca ²⁺ assays and β-	
arrestin recruitment assays	84
Figure 3.2 Genetic variants alter quantity of OXTR on the cell membrane	85
Figure 3.3 Method and data processing for desensitization and internalization assays	86
Figure 3.4 OXTR variants alter receptor activation, desensitization, and internalization	87
Figure 3.5 DiffNets identify distances associated with V45L, P108A, and V281M OXTR	88
Figure 3.6 Potential mechanism for altered β-arrestin function in V45L and P108A OXTR	89
Figure 3.7 Conformational changes in V281M OXTR may reduce G protein binding	90
Figure 4.1 Summary of Folding@home's computational power	.112
Figure 4.2 Structural characterization of Spike opening and conformational masking for thre	e
Spike homologues	.113
Figure 4.3 Effects of glycan shielding and conformational masking on the accessibility of	
different parts of the Spike to potential therapeutics	.115
Figure 4.4 Examples of cryptic pockets and functionally-relevant dynamics	.116
Figure 5.1 Structural view of NSP16	.148
Figure 5.2 DiffNets and MSMs reveal the mechanism of NSP16 activation	
Figure 5.3 Cryptic pocket opening in NSP16	.150
Figure 5.4 Cryptic pocket opening is conserved across coronavirus homologs	.151
Figure 6.1 Training scheme to predict cryptic pocket opening	.155
Figure 6.2 Evaluation of graph neural network's ability to predict cryptic pocket opening in	
simulation	.157
Figure 6.3 Graph neural network prediction on experimentally determined cryptic pockets	.158
Figure A.1.1 Self-supervised DiffNets are robust across a range of expectation maximization	
bounds.	
Figure A.1.2 Self-supervised DiffNets improve ability to predict property of a variant outside	the
training	
Figure A.1.3 Impact of expectation maximization on what features a DiffNet uses to distingu	ıish
variants	
Figure A.1.4 DiffNet analysis suggests conformational changes on switch-II are important for	
distinguishing high-and low-duty myosin isoforms	
Figure A.2.1 Atoms included in DiffNets analysis	
Figure A.2.2 Comparison of equilibrium properties calculated from simulations using differe	
clustering methods	179

Figure A.2.3 Oxytocin-induced ß-arrestin recruitment to wild type (WT) and variant OXTRs	.180
Figure A.2.4 Bias plots for wild type (WT) and variant OXTRs	.181
Figure A.3.1 Distribution of SARS-CoV-2 Spike RBD opening	.182
Figure A.3.2 Simulations of the SARS-CoV-2 Spike complex reveal the existence of an "open"	,
state	.182
Figure A.3.3 Gylcosylated SARS-CoV-2 spike protein transitions to an extremely open state	
through simultaneous rotation of the RBD.	. 183
Figure A.3.4 The discovery of cryptic pockets on NSP5 is robust to the choice of forcefield	. 185
Figure A.3.5 NSP3-PL2Pro domain transition from closed to open state	. 185
Figure A.3.6 NSP5 (dimer) transition from closed to open state	. 185
Figure A.3.7 NSP7 transition from closed to open state	.186
Figure A.3.8 NSP8 transition from closed to open state	.186
Figure A.3.9 NSP9 (dimer) transition from closed to open state	.186
Figure A.3.10 NSP10 transition from closed to open state	. 187
Figure A.3.11 NSP12 transition from closed to open state	. 187
Figure A.3.12 NSP13 transition from closed to open state	. 187
Figure A.3.13 NSP14 transition from closed to open state	. 188
Figure A.3.14 NSP15 transition from closed to open state	. 188
Figure A.3.15 NSP16 transition from closed to open state	. 188
Figure A.3.16 NSP10/NSP14 (complex) transition from closed to open state	.189
Figure A.3.17 NSP10/NSP16 (complex) transition from closed to open state	.189
Figure A.3.18 Nucleoprotein dimerization domain transition from closed to open state	. 189
Figure A.3.19 Human ACE2 transition from closed to open state	.190
Figure A.3.20 Human IL6 transition from closed to open state.	.190
Figure A.3.21 Human IL6-R transition from expanded to closed state	.190
Figure A.4.1 Implied timescales plot	.191
Figure A.4.2 Distance distribution replicates.	
Figure A.4.3 SASA calculation replicates	. 192
Figure A.4.4 Cryptic pocket opening distribution replicates	
Figure A.4.5 Change in root mean square fluctuation (rmsf) of Nsp16 upon Nsp10 associatio	n.
Figure A.4.6 DiffNets predict that $\beta4$ peels away from $\beta3$ in Nsp16 inactive structural states.	
Figure A.4.7 Displacement of Nsp10 binding residues by cryptic pocket opening	. 195
Figure A.4.8 Structural comparison of $\beta 3\text{-}\beta 4$ cryptic pocket in SARS-CoV2 Nsp16 and human	
CMTr1	
Figure A.4.9 Multiple sequence alignment of Nsp16 homologs from coronaviruses	.197

List of Tables

Table 3.1 OXTR variants for study83
Table 4.1 Summary of protein systems we have simulated on Folding@home, organized by viral
strain117
Table A.2.1 Oxytocin response in Ca ²⁺ assays for wild type (WT) and variant OXTRs174
Table A.2.2 Oxytocin-induced β -arrestin-1 recruitment for wild type (WT) and variant OXTRs 176
Table A.2.3 Oxytocin-induced β -arrestin-2 recruitment for wild type (WT) and variant OXTRs 177
Table A.2.4 Log(IC50)s for desensitization and internalization curves for wild type (WT) and
variant OXTR
Table A.4.1 Timescales for transitioning between the pocket closed and open states in Nsp16
homologs

Acknowledgments

"The weird thing is, now I'm exactly where I want to be and I'm still just thinking about my old pals. Only now, they're the ones I made here. I wish there was a way to know you're in the good ole days before you've left them."

The Office, Andy Bernard (aka "Nard-dog")

The path my life has taken over the course of my thesis work is not what I could have predicted, and it has come with euphoric highs, crushing lows, times of panic, and stretches of manic enthusiasm. Yet, I am eternally grateful to the universe because I feel incredibly lucky when I reflect on where I am now and all of the pivot points that could have led me to less prosperous outcomes, both personally and professionally. I dedicate the rest of this section to the many people who played an integral role in helping me achieve an unbelievably enjoyable and productive five years.

First, I'd like to thank my advisor, Greg Bowman. Greg was the reason I was drawn to WashU in the first place, making him the initial catalyst of this beautiful experience. I did a research rotation in Greg's lab right when I arrived at WashU and had an incredible time. In retrospect, it should have been obvious that this was the lab to do my thesis work in, but I chose to work in another lab. It wasn't until a year and a half later that I came crawling back to Greg and Greg graciously welcomed me back to the lab. Greg has been a mentor to me, a model of success, and a kind person who has always let me put my goals first. Greg was initially very hands on and taught me the value of trying as many things as possible. I used to overthink and hesitate before doing anything, but Greg always encouraged a bias toward action, which has paid great dividends. Greg has also progressed rapidly in his own career and has been transparent

throughout the process. Watching this evolution has given me the confidence that I too can continually progress as long as I am constantly working on the next step. Finally, Greg has generously helped me expand my network professionally, which among other opportunities directly led to an internship that has been highly rewarding for my career without benefitting his. Thanks for everything Greg.

Similarly, I want to thank my committee for their guidance. To Gautam, thank you for emboldening me to make the crucial decision that changed my career for the good. To Roman, thank you for giving me random career advice, as I've admired your line of work from afar. To Janice, thank you for being attentive. You made my thesis updates a productive endeavor by providing useful feedback. To Alex, thank you for exemplifying balance. Your dedication to science is out of this world, and you maintain a high level of integrity and friendliness.

Naturally, I want to extend my gratitude to the whole Bowman lab group. Throughout my life, I've had an insufferable tendency to think I am the best in whatever peer group I enter. I can say emphatically that this quickly ended when I became a member of the Bowman lab. I quickly realized the people around me were not only more experienced than I was, they were also clearly more intelligent than I am. It was humbling, but mostly exciting that I had so many people to guide the way I think in a way I never had before.

In my early lab days, Maxwell Zimmerman and Justin Porter had a particularly strong role shaping my graduate school experience. Max has improved every idea I've come up with, has given me an unreasonable amount of generosity with his time, and became one of my closest companions. From daily sessions of playing ping-pong, piano, singing, working out together, etc. to arguments over capitalism and hostile AI, Max has played an inimitable role as a friend and

mentor throughout my graduate school experience. Like when I met Max, it was clear to me that Justin was just on another level as far as intelligence and ability go. By example, he has helped me improve my software engineering skills. Moreover, he has helped me hone the skill of dropping one's ego to be able to relentlessly battle in the space of ideas without personal attacks. I can't think of a more important ability for discovering ways to bring about *good* change in the world.

My later lab days were shaped by Matthew Cruz, Artur Meller, and Neha Vithani. Matt is the rock that makes me feel like I have community here in grad school. He has been my most consistent friend here, my comrade across the many different groups of people we end up hanging out together with. Like Justin and Max have been my mentors in the world of computers, Matt has been my mentor in experiment land. He continually keeps my thinking grounded in terms of what people can actually do, or care about, in biochemistry. Neha has been the friend with the perfect balance of playful and productive. Neha and I co-wrote a paper together and it was the smoothest collaboration I've ever had. She also helped out when I was helping build a tiny house, gone to see Christmas lights with me, and has tried to optimize the food she cooks by using me and my continuous glucose monitor as a guinea pig! Finally, Artur quickly progressed from labmate to my ride-or-die buddy. I've also respected the extraordinary depth to which Artur takes questions or problems, it has definitely increased my ability and desire to do so. In the last year we've also been roommates and co-leads on a project and somehow we're not yet worst enemies. From cooking together with friends, to fancy nights out, to ping pong battles, to skiing and playing spikeball in the mountains of Salt Lake, I'm grateful to have made a lifelong friend. The Bowman lab is quite large, so I will spare the reader more highlights and conclude by saying I am grateful for all of the people I've gotten to know in the

lab. I've had special interactions and moments of admiration for each and every person who has been in this lab, and I'm lucky to have ended up here in the end.

I would be remiss if I did not thank the members of Joshua Swamidass' lab for their guidance and friendship during my thesis work. The Swamidass' lab is where I started to internalize what deep learning is and how it works, which set me up for future success. In particular, I'm grateful to have collaborated with Noah Flynn extensively, received excellent mentorship from the brilliant Tyler Hughes, and had Arghya Datta and Rohit Farmer to vent and spend time with.

I also want to thank two of my close collaborators, Meghana Kshirsagar and Manasi Malik. Meghana, you have been a mentor that has helped me hone my expertise with machine learning, and you've been a good friend. Manasi, your work inspires me, and you gave me something to be excited about working on when I wasn't always feeling the same way about my own work. To both, thank you for your kindness, it has been a joy to work with you.

Outside of lab, I was lucky to be immediately absorbed into a tightknit group of biophysics friends. I was a bit of a loner on the first day of orientation, while my cohort of ~80 graduate students watched the 2017 solar eclipse. All of the ~8 biophysics grad students had already made friends and formed a group, but I wasn't a part of it until Jasmine Cubuk noticed me standing alone and dragged me over to the group. Since then we've had a group chat that has had at least one message in it every day for each of the last 5 years (roughly 1700 days since it started!). From potlucks to qualifying exam practices to holiday parties to movie nights to dungeons and dragons etc. I have never been part of such a closely knit and supportive friend

group. Thank you Victoria Ismail, Huiming Xia, JJ Alston, Jasmine Cubuk, Kellan Weston, Matthew Cruz, Naomi Wilson, Paige Hall, and Rose Reis for being my grad school family.

Outside of science, I am thankful that I got to pursue my fiery passion for singing. I participated in *The Addam's Family Musical* and *Shrek The Musical*, as well as, *The Histones* a capella singing group. Both of these groups are student run and include medical students, occupational therapy students, physical therapy students, audiology students, and students in my own graduate program. I am indebted to those who organized the groups and made the shows happen. There are too many friends from these groups to name, but those who I made deep connections with, you know who you are, and I'm elated to have met you! I also want to give a special shout out to Nicole Huang, an occupational therapist who is forever my close friend and foodie buddy.

Outside of the WashU community, I want to thank my best friend Collin Walsh and his wife Sarah for being my family throughout this entire journey. Collin (nicknamed "K") and I wanted to live together at some point ever since we'd been kids. He packed up and moved from Florida to come live with me in Saint Louis while I did the grad school thing. Words don't do justice to the companionship he's provided me along this journey, and honestly, I don't think I really need to find the right words because he knows how much I love him and have valued his company here in The Lou. To Sarah, you're newer to the family, but it wouldn't be the same with you. I am going to miss the many nights we've all spent together, nothing beats that quality time.

To my other best friend, Nick Marino, you've played an instrumental role in who I've become even if your influence has been less direct in the past few years because of physical

distance. No one's ideas have inspired and influenced my life path more than Nick. Whether you know it or not, you're still pulling levers in my brain that shape my present and my future. Thank you for inspiring me to be a scientist and engineer.

Last, but not least, I am eternally grateful to Cristina Macklem for her companionship and her influence on my life. I am in awe of your work ethic, your conviction to always do the right thing, your desire to provide for those around you, and your empathy. Cristina and I built (mostly her) and lived in a tiny house together during the middle of my thesis work. We also adopted an incredibly cuddly pitbull who got us out of the house daily during the pandemic. Thank you for being a sounding board for my ideas, for encouraging me to feel proud of my work, for helping me be a better writer, and for serving as an example as the most disciplined, hardest working person I've ever met. Thank you for your companionship through all of it. You are responsible, in large part, for why I view these past 5 years as such a wonderful part of my life. I would also like to thank her family, Estrella, Chris, Dave, and Christine who all provided support and love throughout.

Oh, oops, I almost forgot my own family! I am fortunate enough to be extremely close with everyone in my immediate family, and I am sure that helps explain why I've felt such a strong sense of stability and security during my thesis work. First and foremost, to my mom, thank you for being where home is for my entire life. To my dad, you have instilled your extreme work ethic in me, which I needed to complete this work. To my brother Kevin, I have always felt one step behind you in terms of mental acuity, thank you for giving me the chip on my shoulder to strive for excellence, not only with my intellectual pursuits, but with my personal responsibilities as well. To my brother Jeff, thanks for showing me the kind of personality it

takes to make others feel deeply loved, admired, and cool. You give people a strong sense of worth in the way that you build them up, and I try to offer this to my peers every day. To my sister Sarah, thank you for helping me stay grounded. You, more than anyone in this family, are in tune with what is important in life, and that has helped me keep valuable perspective when I have felt discouraged during grad school. To my cousin Kaitlin, thank you for bringing me familial camaraderie by being my only family member in science. You've provided me with hands-on guidance with my writing, and you've provided wisdom no one else could. To my Nana, thank you for interrogating me about science. You have truly asked me more questions about biology and my research than anyone else in this world. It is stunning to me the way you are able to string me along to enthusiastically describe everything I know about biology and AI, and that has kept me inspired to keep going.

Finally, I want to express gratitude for the Alcoholics Anonymous community. It feels like taboo to discuss substance abuse issues in the scientific community, but I think it is important for me to acknowledge that I struggled with alcohol abuse throughout my life bleeding into graduate school. The AA community helped me get sober over two years ago, and this has probably been the single most important change I have made in my life. Colin Kluender was a gentle, caring voice who made me feel understood and like I wasn't alone. May he rest in peace.

ABSTRACT OF THE DISSERTATION

Combining Computer Simulations and Deep Learning to Understand and Predict Protein

Structural Dynamics

by

Michael D. Ward

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2022

Professor Gregory R. Bowman, Chair

Molecular dynamics simulations provide a means to characterize the ensemble of structures that a protein adopts in solution. These structural ensembles provide crucial information about how proteins function, and these ensembles also reveal potential drug binding sites that are not observable from static protein structures (i.e. cryptic pockets). However, analyzing these high-dimensional datasets to understand protein function remains challenging. Additionally, finding cryptic pockets using simulation data is slow and expensive, which makes the appeal of computationally screening for cryptic pockets limited to a narrow set of circumstances. In this thesis, I develop deep learning based methods to overcome these challenges. First, I develop a deep learning algorithm, called DiffNets, to deal with the high-dimensionality of structural ensembles. DiffNets takes structural ensembles from similar systems with different biochemical properties and learns to highlight structural features that distinguish the systems, ultimately connecting structural signatures to their associated biochemical properties. Using DiffNets, I

provide structural insights that explain how naturally occurring genetic variants of the oxytocin receptor alter signaling. Additionally, DiffNets help reveal how a SARS-CoV-2 protein involved in immune evasion becomes activated. Next, I use MD simulations to hunt for cryptic pockets across the SARS-CoV-2 proteome, which led to the discovery of more than 50 new potential druggable sites. Because this effort required an extraordinary amount of resources, I developed a deep learning approach to predict sites of cryptic pockets from single protein structures. This approach reduces the time to identify if a protein has a cryptic pocket by ~10,000-fold compared to the next best method.

Chapter 1

Introduction

1.1 Proteins are the molecular machinery of the living

When organizing the physical levels of biology, proteins exist toward the bottom of the scale. At the lower end there are atoms and small molecules (e.g. metabolites). At the higher end there are large molecular complexes, organelles, cells, tissues, organs, humans, and so on. Recent estimates suggest there are roughly 4 trillion cells in a human¹ and roughly 42 million protein molecules per cell², which implies there are $\sim 1.6 * 10^{\circ}19$ protein molecules in an average human body. For reference, there are an estimated $\sim 3 * 10^{\circ}23$ stars in the entire universe.³ Importantly, there are an estimated 20,000 protein coding genes in the human body⁴ and there are likely many more distinct proteins (e.g. splice variants) that behave differently and exist in different environments in the body.

Proteins are chains of amino acids that can be composed in nearly infinite ways to accomplish a dizzying breadth of functions that are crucial to life. While an exhaustive list of different protein functions is not practical, a few examples are worth highlighting. Proteins catalyze chemical reactions,⁵ proteins regulate how many (and which types) of other proteins should be in the cell,⁶ proteins recognize and deactivate other proteins from external pathogens,⁷ proteins transport cargo across and between cells,⁸ etc.

The geometric structure that a protein adopts is an important determinant of how it functions. Upon production, many proteins fold into a unique three-dimensional structure, which

is determined by the protein's primary sequence, or the composition of amino acid residues that make up the protein. A protein's structure is one crucial feature that determines what it can interact with in its environment, and therefore is crucial to its function. For example, when a protein catalyzes a chemical reaction, it must "find" the chemical (i.e. small molecule) that it has evolved to interact with. This is achieved in part by having a pocket that is topologically complementary to the small molecule.

Given the importance of protein structure, many decades of research have gone into determining protein structures with atomistic resolution. In 1958, myoglobin was the first protein to have its structure solved. Today, there are over 180,000 protein structures deposited in the protein data bank (PDB), which were solved using a variety of techniques including Nuclear Magentic Resonance (NMR), X-ray crystallography, and cryogenic electron microscopy (cryo-EM). More recently, in 2020, a computational method called AlphaFold was shown to accurately predict protein structures without the need for experimental data. With AlphaFold predicted structures there are now millions of structures available that provide detail on the exact positioning of all the atoms in a protein within ~1-2 Å error.

Beyond a protein's native, folded structure, the way the protein structure fluctuates in solution is another important determinant of its function. Proteins are composed of atoms that are constantly in motion, which means that proteins adopt an ensemble of structural configurations when in solution. The structural diversity in these ensembles may range from subtle amino acid sidechain motions, which are important for coordinating ligand binding during enzyme catalysis, ¹⁴ to large scale reorganizations like the 12 nm swinging of the myosin lever arm during muscle contraction. ¹⁵ Many methods have been developed to characterize protein conformational

ensembles including NMR, fluorescence resonance energy transfer (FRET),¹⁶ Hydrogen Deuterium Exchange (HDX),¹⁷ and Molecular Dynamics simulations (MD).¹⁸

1.2 Molecular Dynamics simulations provide atomistic resolution of a protein's structural dynamics

Molecular Dynamics (MD) simulations can serve as a computational microscope, providing a time-course trajectory of how a protein structure fluctuates in solution. Typically, MD simulations start with a protein molecule situated in a "box" of water where all atoms and bonds are modeled as balls and springs, respectively. MD simulations iteratively apply forces to each atom in the box over discrete timesteps to update the positions of all atoms. These forces come from bonded and non-bonded interactions between atoms in the system, where the interactions are typically parameterized by a combination of empirical observations and theoretical calculations, which makes up a "force field". Provided a force field with perfect physical fidelity and a simulation of infinite timescale, one can accurately calculate the thermodynamic and kinetic behavior of the protein of interest. Of course, neither of these are possible in practice. Still, improvements to force fields and compute power over time have come with substantial progress in improving our ability to determine a protein's conformational ensemble with MD simulations. ¹⁹

Simulating protein dynamics contributes to our understanding of basic science. While experimental methods like NMR and FRET can measure protein dynamics on longer timescales than MD simulations, they cannot determine how the exact positions of a protein's atoms evolve over time. The unique ability for MD simulations to capture this level of atomistic detail over the course of up to ~milliseconds of time^{20,21} means that MD simulations can provide mechanistic

insights impossible otherwise. As such, MD simulations have been instrumental for shaping our understanding of how proteins fold. In more recent work, MD simulations have helped us understand how protein conformational changes are an important part of protein function. For example, MD simulations of the SARS-CoV-2 spike protein have helped uncover how the spike protein opens to engage ACE2 receptors on human cells to infect them.^{22,23} Knowledge about important contacts in the spike trimer from these simulations helps inform how new variants of the spike may increase or decrease infectivity.

MD simulations have been particularly useful in situations where a system behaves differently with and without a perturbation (e.g. apo vs ligand bound), but experimentally determined structures do not explain why the systems behave differently. For example, different isoforms of the protein myosin have varying roles in different environments, but have almost identical native, folded structures.²⁴ It has been shown that they manage these different functions, in part, by having altered preferences for certain structural states in their conformational ensemble. In a similar vein, Sultan et. al. used simulations to understand how structural fluctuations across seven different Src kinase family members tune their functions.²⁵ Importantly, this study identified structural states unique to each kinase, which could help in the development of drugs that can target each kinase specifically, rather than hitting the active site that is common to all seven kinases.

While MD simulations provide a wealth of protein structural data to explain protein function, it is challenging to determine which structural fluctuations are critical for function. Let us assume, for a typical MD simulation study, structural snapshots are saved every 20 picoseconds, ~5 microseconds of data are accumulated, and the protein simulated contains ~250

amino acids. The end product is 250,000 structural snapshots, which contain thousands of atom positions in each snapshot. Analyzing every piece of this raw data is time consuming and, more importantly, it is challenging to wrangle this data into human interpretable conclusions. Therefore, there has been lots of research towards developing methods to analyze simulation data. Because of the high dimensionality of the problem, dimensionality reduction algorithms are often used. Common approaches include transforming the data into a Markov state model (MSM),²⁶ applying principal component analysis (PCA)²⁷ or similar algorithms like time-based independent component analysis (tICA),²⁸ training neural networks,²⁹ and simply choosing to focus exclusively on "collective variables" such as the distance between two residues based on some *a priori* knowledge about the system of choice.

Beyond advancing our understanding of basic science, MD simulations play an important role in the development of new medicines. One important goal when developing a new medicine is to identify a ligand that will bind to a protein with high affinity. In the MD simulation community, accurately computing the free energy of a ligand binding to a protein is a heavy area of focus. The SAMPL challenge is one way researchers evaluate how well MD simulation methods are progressing at this challenge. As one of its challenges, SAMPL holds out a test set of know protein-ligand binding affinities and methods are submitted to try and accurately predict this property. MD simulation methods have been progressing at this challenge, which has come with tangible benefits for drug development. For example, the COVID moonshot project has used simulation-based free energy calculations to figure out which small molecules should be prioritized as potential inhibitors for SARS-CoV-2 proteins, which could help treat the deadly COVID19 disease. 31

Identification and targeting of a protein's "cryptic pockets" is another promising direction in medicine where MD simulations are making strong contributions. A structure of a protein's native, folded state can reveal potential drug binding pockets, but leaves us blind to other potential pockets that form as the protein structure fluctuates in solution. There are over 100 confirmed examples of these "other" binding pockets where a small molecule binds in a pocket on a protein which was not observable from any previously determined structure of that protein (i.e. a "cryptic pocket"). Since alternative structural states are known to be functionally important (e.g. for allostery, acatalysis, ligand binding, fee.) and drugs designed to target cryptic pockets can modulate time spent in alternate states, cryptic pockets make for compelling drug targets. In fact, McCammon et. al. used MD simulations to discover a novel binding trench in HIV integrase, which led to the development of an antiretroviral drug raltegravir by Merck. Importantly, this binding trench was not observed in any previous determined structures of the HIV integrase, highlighting the utility of identifying cryptic pockets in simulation.

1.3 Deep learning has made a lasting impact on protein biophysics

In most scenarios where there is an abundance of data, machine learning algorithms are a suitable choice for effectively using the data. Machine learning algorithms are algorithms that learn through experience or by the use of data and can generally be categorized as *supervised* or *unsupervised*. Unsupervised machine learning algorithms try to highlight patterns in data without any explicit instruction on what might be important. Unsupervised algorithms such as clustering, PCA and neural network-based autoencoders³⁹ are popular choices to apply to MD simulation data as they can highlight structural/dynamic properties in a protein that might be difficult to

notice otherwise. Supervised machine learning algorithms are trained to map inputs to outputs, which often requires that they find features in the data that separate two or more classes. Supervised machine learning algorithms have been used to predict sites of cryptic pockets in proteins, ³² predict protein-protein interactions, ⁴⁰ predict protein/RNA structures, ⁴¹ among many other tasks.

Most machine learning approaches developed over the past decade are "deep learning" methods based on artificial neural networks. ⁴² Artificial neural networks emerged as a simple, more powerful extension of logistic regression. In logistic regression, input features (e.g. # of amino acid residues, net charge of protein, radius of protein) get mapped to an output (e.g. enzyme or not?) by multiplying each input feature by a weight, summing the values, and then applying a function that compress the value between 0 (e.g. not an enzyme) to 1 (e.g. is an enzyme). The weights are learned via example. Specifically, weights are changed such that they would make a more accurate prediction next time seeing the example, and this is done through a stochastic gradient descent algorithm. ⁴³ Artificial neural networks extend this by adding intermediate ("hidden") layers that transform the input features through a successive series of multiplications with matrices of learned weights, with nonlinear "activation" functions between each matrix multiplication. The weights are updated by example through an algorithm termed backpropagation. ⁴⁴ The added layers allow neural networks to be more expressive than a logistic regression model giving neural networks the ability to approximate any function.

Deep learning algorithms have become the dominant machine learning method employed across most domains including protein biophysics. In 2012, Krizhevsky et. al, developed AlexNet, a deep learning model that performed substantially better than any previous approach

for classifying images across ~20,000 classes. ⁴⁵ Since this accomplishment, deep learning has become the dominant approach used to tackle algorithmic challenges across many fields including computer vision, ⁴⁶ natural language processing, ⁴⁷ competitive gaming, ⁴⁸ and most recently protein biophysics. Within protein biophysics, the protein structure prediction problem has been deemed "solved" based on a deep learning algorithm called AlphaFold. ¹³ This has enabled the characterization of millions of protein structures that were previously undetermined. Other notable deep learning breakthroughs include state-of-the-art work predicting RNA structure, ⁴¹ predicting protein-protein interactions, ⁴⁰ predicting protein-ligand interactions, ⁴⁹ and predicting what protein sequence is capable of folding into a predetermined structure. ⁵⁰

Accurately predicting drug binding sites is one area where deep learning is helping advance drug development. Before the rise of deep learning, the most common way to evaluate if a small molecule was a good candidate to bind a protein was via "docking". ⁵¹ In docking, one docks a small molecule drug to a pocket in the protein structure and scores the docked pose, using physics-inspired scoring functions, to determine if they are likely to bind to the protein target. Now, several companies are using deep-learning based algorithms to score docked poses after training the models to discriminate between small molecules that bind to a protein target from those that do not using known examples derived from the PDB. ^{49,52} In a similar spirit, several deep learning algorithms have been developed to determine if a region on a protein structure is a "hot spot" for ligand binding, which helps researchers decide whether a protein is worth targeting. ^{53,54} The deep learning based methods are not trained to identify cryptic pockets (i.e. pockets that are not present in the native structure), but other types of machine learning algorithms have shown that this goal can be achieved with reasonable accuracy. ³² It is possible that a deep learning based approach would improve the performance on this task.

The impact of deep learning has mainly been felt in the world of static structures, but encouraging results are emerging that show the potential of applying deep learning to gain information about ensembles of protein structures. Improving our ability to sample structural ensembles is one major area of progress that can be attributed to deep learning examples. For example, Noe et. al. designed "Boltzmann Generators" which generate physically realistic alternate protein structural configurations in a manner that is Boltzmann distributed meaning thermodynamic information about the ensemble can be calculated from the generated structures.⁵⁵ This accomplishes many of the goals of MD simulations, but has the potential to use substantially fewer computational resources. While promising, this method remains a proof of concept and has not been vetted at a large scale. In a similar spirit, several groups have made progress replacing traditional force fields, which are used to calculate Newton's law of motion, by training deep learning models to accurately compute these forces. 56,57 Through this type of approach one could, in principal, remove the bulk of atoms in simulation (i.e. water) and still accurately calculate a protein's structural fluctuations. While this promises to reduce the resources needed to run simulations, the approach is not yet used regularly in practice. Finally, there has been an effort to use deep learning to speed up MD simulations by guiding the systems to sample functionally relevant regions of conformational space. 58-60 This reduces the amount of compute resources spent sampling irrelevant conformations. In practice, none of the discussed approaches have been shown to generalize beyond a single protein per networked trained, which has greatly limited their utility.

Deep learning is also starting to play a role in the analysis of conformational ensembles. Many non deep learning algorithms have been developed and employed successfully to analyze conformational ensembles, but the utility of each is limited by assumptions that are not

universally appropriate. For example, PCA finds linear combinations of features that retain as much of the geometric variance in the original data as possible, ²⁷ effectively assuming that large structural changes are more important than subtle ones. Unfortunately, there are many cases where this assumption is invalid, as in enzymes where arbitrary motions of a large floppy loop may dwarf subtle but functionally-relevant sidechain motions in the active site. Autoencoders, ³⁹ deep learning based models, are a more powerful alternative since they consider nonlinear combinations of features. These neural networks learn a low-dimensional projection of data called the latent space—that is optimized to produce a high-fidelity geometric reconstruction of a protein configuration. However, like PCA, autoencoders still focus on capturing large geometric variations. Time-lagged independent component analysis (tICA) is another common approach.²⁸ It is similar to PCA but focuses on slowly varying degrees of freedom rather than emphasizing large geometric changes. However, there are many situations where the conformational changes of interest are fast relative to others (e.g. allostery within the native ensemble that is faster than folding and unfolding of the protein). Another recent approach, VAMPnets, ²⁹ combines ideas from autoencoders and tICA to achieve a dimensionality reduction that maps protein structures to metastable states. This allows VAMPnets to capture non-linearities that tICA cannot, but the assumption that slowly varying degrees of freedom are more important than faster ones is still limiting in many cases.

1.4 Scope of Thesis

While the application of deep learning models to protein biophysics has come with swift progress, there are areas of scientific and medicinal interest that could benefit from the development of new deep learning approaches. As discussed above, deep learning-based approaches for analyzing conformational ensembles are still in the nascent stages of

development. Specifically, comparing and contrasting conformational ensembles across similar systems remains a challenge. The intersection of deep learning, conformational ensembles, and drug development is another poorly studied area. Therefore, studies that employ deep learning methods to learn from conformational ensembles to inform drug development are needed. These studies would shed new light on the importance (or not) of conformational ensembles in drug development and inform on whether or not deep learning-based approaches are suitable for this problem.

Chapter 2 details the development of DiffNets, a deep learning approach for comparing and contrasting conformational ensembles. Before the development of DiffNets, PCA, tICA, and autoencoders were commonly used to compare and contrast conformational ensembles derived from MD simulations. The main limitation of these approaches is that they have no explicit mechanism built in to them to highlight differences between datasets. DiffNets are an extension of autoencoders; they are augmented with a classification task that constrains them to learn a low dimensional representation of data that highlights features that distinguish two distinct datasets (e.g. two protein variants with different biochemical properties). Chapter 2 demonstrates the utility of DiffNets on a couple of examples. First, we showed that DiffNets can identify a structural signature of stability that distinguishes single-point variants of the bacterial protein TEM β-lactamase. Next, we showed that DiffNets can identify a structural signature that distinguishes low sequence identity isoforms of the motor protein, myosin. This work demonstrated that DiffNets is applicable to larger proteins that have perturbations beyond single point mutations.

While chapter 2 details the application of DiffNets to systems with well documented differences, chapter 3 uses DiffNets to highlight differences between conformational ensembles across systems that had not been previously described. Namely, we applied DiffNets to 4 commonly occurring genetic variants of the human oxytocin receptor. Two of these variants have impaired signaling related to their interactions with β -arrestin, one variant has impaired signaling related to its interaction with Gq, and the other variant is the wild-type. After simulating all four variants, we trained separate DiffNets to look for signatures of β -arrestin and Gq impairment. Identifying the signatures that distinguish the variants helped identify several structural motions that appear to be critical for normal interaction with Gq and β -arrestin.

Like many graduate students in my cohort, my research was massively interrupted by the onset of the COVID19 pandemic.⁶¹ This interruption shifted my research focus from understanding structural mechanisms (e.g. with DiffNets) to providing insight that may be relevant for the development of therapeutics. Toward this effort, I contributed to work that uncovered more than 50 new sites with druggability potential across the SARS-CoV-2 proteome (chapter 4). Specifically, we simulated most of the SARS-CoV-2 proteome to uncover cryptic pockets. The structural and thermodynamic characterization of the pockets provides a template for designing new antiviral inhibitors.

In chapter 5, I explored the mechanism of activation and the druggability potential of a SARS-CoV-2 protein, nonstructural protein (NSP) 16. NSP16 is a methyltransferase that plays a key role disguising the SARS-CoV-2 genome from human immune proteins that evolved to recognize pathogens. Importantly, NSP16 is only active when in the presence of its binding partner, NSP10. Before the work in chapter 5, the mechanism of activation was unknown.

Comparing structural ensembles of the NSP16 monomer and the NSP10/16 dimer using DiffNets we found that NSP10 stimulates NSP16 by increasing its propensity to adopt structural configurations with an open active site. We also discovered a cryptic pocket in NSP16 that, when open, coincides with closing of the active site. Therefore, wedging this cryptic pocket open with a small molecule would serve to inactivate NSP16. This cryptic pocket opens in the SARS-CoV and MERS-CoV homologs, but not in the human homolog. Therefore, this discovery could help in the development of a pan-coronavirus inhibitor.

Chapter 6 details a deep learning approach I developed to accelerate the discovery of cryptic pockets. Cryptic pockets are commonly discovered via molecular dynamics simulations or drug screening campaigns. Both of these methods are relatively slow and expensive. An algorithm that can quickly identify whether a protein has a cryptic pocket, and ideally locate where, would help researchers prioritize which proteins are worth simulating or screening. The state-of-the-art approach for this task is a supervised machine learning algorithm that was trained to identify cryptic ligand binding sites from ~90 examples in the PDB. The algorithm is slow (~1 day to run on a single protein) and has only mediocre accuracy, which is likely due to the dearth of available data. Instead of training a model on known cryptic binding sites, I developed a model to predict protein breathing motions, which was trained on MD simulation data. This model predicts sites of known cryptic ligand binding pockets with similar accuracy to the previous state-of-the-art and it does so ~10,000x faster allowing it to be used in a high throughput manner.

Chapter 7 concludes the thesis summarizing the achievements and discoveries made across studies and highlight future directions that can emerge from this work.

Bibliography

- 1. Bianconi E, Piovesan A, Facchin F, et al. An estimation of the number of cells in the human body. *Ann Hum Biol.* 2013. doi:10.3109/03014460.2013.807878
- 2. Ho B, Baryshnikova A, Brown GW. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.* 2018. doi:10.1016/j.cels.2017.12.004
- 3. Kahane G. Our cosmic insignificance. *Nous*. 2014. doi:10.1111/nous.12030
- 4. Salzberg SL. Open questions: How many genes do we have? *BMC Biol*. 2018. doi:10.1186/s12915-018-0564-x
- 5. Robinson PK. Enzymes: principles and biotechnological applications The nature and classification of enzymes. *Essays Biochem*. 2015.
- 6. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell.* 2018. doi:10.1016/j.cell.2018.01.029
- 7. Basler CF, Mikulasova A, Martinez-Sobrido L, et al. The Ebola Virus VP35 Protein Inhibits Activation of Interferon Regulatory Factor 3. *J Virol*. 2003. doi:10.1128/jvi.77.14.7945-7956.2003
- 8. Ross JL, Ali MY, Warshaw DM. Cargo transport: molecular motors navigate a complex cytoskeleton. *Curr Opin Cell Biol.* 2008. doi:10.1016/j.ceb.2007.11.006
- 9. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*. 1958. doi:10.1038/181662a0
- 10. Mlynárik V. Introduction to nuclear magnetic resonance. *Anal Biochem*. 2017. doi:10.1016/j.ab.2016.05.006
- 11. Johnson LN. Protein crystallography. *New Compr Biochem*. 1985. doi:10.1016/S0167-7306(08)60564-5
- 12. Hebert H. CryoEM: a crystals to single particles round-trip. *Curr Opin Struct Biol.* 2019. doi:10.1016/j.sbi.2019.05.008
- 13. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020. doi:10.1038/s41586-019-1923-7
- 14. Ojeda-May P, Mushtaq AUI, Rogne P, et al. Dynamic Connection between Enzymatic Catalysis and Collective Protein Motions. *Biochemistry*. 2021. doi:10.1021/acs.biochem.1c00221
- 15. Holmes KC. The swinging lever-arm hypothesis of muscle contraction. *Curr Biol.* 1997. doi:10.1016/s0960-9822(06)00051-0
- 16. Ha T. Single-molecule fluorescence resonance energy transfer. *Methods*. 2001. doi:10.1006/meth.2001.1217
- 17. Narang D, Lento C, Wilson DJ. HDX-MS: An analytical tool to capture protein motion in action. *Biomedicines*. 2020. doi:10.3390/BIOMEDICINES8070224
- 18. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol*. 2002. doi:10.1038/nsb0902-646
- 19. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. *J Comput Chem*. 2005. doi:10.1002/jcc.20291
- 20. Shaw DE, Dror RO, Salmon JK, et al. Millisecond-scale molecular dynamics simulations

- on Anton. In: *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09.*; 2009. doi:10.1145/1654059.1654099
- 21. Shirts M, Pande VS. Screen savers of the world unite. *Science* (80-). 2000. doi:10.1126/science.290.5498.1903
- 22. Zimmerman MI, Porter JR, Ward MD, et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat Chem.* 2021. doi:10.1038/s41557-021-00707-0
- 23. Sztain T, Ahn SH, Bogetti AT, et al. A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat Chem.* 2021. doi:10.1038/s41557-021-00758-3
- 24. Porter JR, Meller A, Zimmerman MI, Greenberg MJ, Bowman GR. Conformational distributions of isolated myosin motor domains encode their mechanochemical properties. *Elife*. 2020. doi:10.7554/eLife.55132
- 25. Sultan MM, Kiss G, Pande VS. Towards simple kinetic models of functional dynamics for a kinase subfamily. *Nat Chem.* 2018. doi:10.1038/s41557-018-0077-9
- 26. Bowman GR, Ensign DL, Pande VS. Enhanced modeling via network theory: Adaptive sampling of markov state models. *J Chem Theory Comput*. 2010. doi:10.1021/ct900620b
- 27. David CC, Jacobs DJ. Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol Biol.* 2014. doi:10.1007/978-1-62703-658-0 11
- 28. Naritomi Y, Fuchigami S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J Chem Phys.* 2011. doi:10.1063/1.3554380
- 29. Mardt A, Pasquali L, Wu H, Noé F. VAMPnets for deep learning of molecular kinetics. *Nat Commun*. 2018. doi:10.1038/s41467-017-02388-1
- 30. Amezcua M, El Khoury L, Mobley DL. SAMPL7 Host–Guest Challenge Overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *J Comput Aided Mol Des.* 2021. doi:10.1007/s10822-020-00363-5
- 31. Achdout H, Aimon A, Bar-David E, et al. COVID moonshot: Open science discovery of SARS-CoV-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *bioRxiv*. 2020.
- 32. Cimermancic P, Weinkam P, Rettenmaier TJ, et al. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J Mol Biol*. 2016. doi:10.1016/j.jmb.2016.01.029
- 33. Guo J, Zhou HX. Protein Allostery and Conformational Dynamics. *Chem Rev.* 2016. doi:10.1021/acs.chemrev.5b00590
- 34. Petrovic D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM. Conformational dynamics and enzyme evolution. *J R Soc Interface*. 2018. doi:10.1098/rsif.2018.0330
- 35. Seo MH, Park J, Kim E, Hohng S, Kim HS. Protein conformational dynamics dictate the binding affinity for a ligand. *Nat Commun*. 2014. doi:10.1038/ncomms4724
- 36. Knoverek CR, Mallimadugula UL, Singh S, et al. Opening of a cryptic pocket in β-lactamase increases penicillinase activity. *Proc Natl Acad Sci U S A*. 2021. doi:10.1073/pnas.2106473118
- 37. Schames JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, McCammon JA. Discovery of a Novel Binding Trench in HIV Integrase. *J Med Chem.* 2004. doi:10.1021/jm0341913
- 38. Hazuda DJ, Anthony NJ, Gomez RP, et al. A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1

- integrase. Proc Natl Acad Sci U S A. 2004. doi:10.1073/pnas.0402357101
- 39. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* (80-). 2006. doi:10.1126/science.1127647
- 40. Townshend RJL, Bedi R, Suriana PA, Dror RO. End-to-end learning on 3D protein structure for interface prediction. In: *Advances in Neural Information Processing Systems*.; 2019.
- 41. Townshend RJL, Eismann S, Watkins AM, et al. Geometric deep learning of RNA structure. *Science* (80-). 2021. doi:10.1126/science.abe5650
- 42. Yegnanarayana B. Artificial neural networks for pattern recognition. *Sadhana*. 1994. doi:10.1007/BF02811896
- 43. Tolstikhin I, Bousquet O, Schölkopf B, et al. An overview of gradient descent optimization algorithms. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2018.
- 44. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986. doi:10.1038/323533a0
- 45. Krizhevsky A, Sutskever I, Hinton GE, et al. ImageNet Classification with Deep Convolutional Neural Networks Alex. *Proc 31st Int Conf Mach Learn*. 2012.
- 46. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.; 2016. doi:10.1109/CVPR.2016.90
- 47. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*.; 2017.
- 48. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*. 2017. doi:10.1038/nature24270
- 49. Feinberg EN, Sur D, Wu Z, et al. PotentialNet for Molecular Property Prediction. *ACS Cent Sci.* 2018. doi:10.1021/acscentsci.8b00507
- 50. Ingraham J, Garg VK, Barzilay R, Jaakkola T. Generative models for graph-based protein design. In: *Advances in Neural Information Processing Systems*.; 2019.
- 51. Chen Y, Shoichet BK. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat Chem Biol.* 2009. doi:10.1038/nchembio.155
- 52. Wallach I, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. 2015:1-11. http://arxiv.org/abs/1510.02855.
- 53. Mylonas SK, Axenopoulos A, Daras P. DeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics*. 2021. doi:10.1093/bioinformatics/btab009
- 54. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*. 2017. doi:10.1093/bioinformatics/btx350
- 55. Noé F, Olsson S, Köhler J, Wu H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* (80-). 2019. doi:10.1126/science.aaw1147
- 56. Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci.* 2017. doi:10.1039/C6SC05720A
- 57. Husic BE, Charron NE, Lemm D, et al. Coarse graining molecular dynamics with graph neural networks. *J Chem Phys.* 2020. doi:10.1063/5.0026133

- 58. Hernández CX, Wayment-Steele HK, Sultan MM, Husic BE, Pande VS. Variational encoding of complex dynamics. *Phys Rev E*. 2018. doi:10.1103/PhysRevE.97.062412
- 59. Wang Y, Ribeiro JML, Tiwary P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat Commun*. 2019. doi:10.1038/s41467-019-11405-4
- 60. Wehmeyer C, Noé F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J Chem Phys.* 2018. doi:10.1063/1.5011399
- 61. Varadarajan J, Brown AM, Chalkley R. Biomedical graduate student experiences during the COVID-19 university closure. *PLoS One*. 2021. doi:10.1371/journal.pone.0256687
- 62. Amaro RE. Will the Real Cryptic Pocket Please Stand Out? *Biophys J.* 2019. doi:10.1016/j.bpj.2019.01.018

Chapter 2

Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets

2.1 Preamble

This chapter is adapted from the following article: Ward, M.D, Zimmerman, M.I., Meller, A., Chung M., Swamidass, S.J., and Bowman, G.R. (2021). "Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets", *Nature Communications*, 12, 3023.

2.2 Introduction

A mechanistic understanding of how a protein's sequence determines its structural preferences and, ultimately, its biochemical properties is crucial for advancing our understanding of fundamental biology and for applications in precision medicine and protein engineering. Sequence variations can modulate a protein's biochemical properties in a deleterious manner leading to morbidity and mortality, 1,2 or in a manner that can improve a species fitness, e.g. conferring the ability to metabolize new substrates. Moreover, entire protein families with a wide range of functions and biochemical properties emerge after long timescale evolution of protein sequences. In either case, identifying the structural and dynamical differences between protein variants is a powerful means to understand the mechanism that connects a protein's sequence and biochemical properties. 4-9 Streamlining this process would make it easier to infer the behavior of new protein variants, which would accelerate protein engineering and the

interpretation of newly discovered variants. Understanding the structural basis for protein function and dysfunction can also accelerate the development of drugs and other therapeutics.

Identifying the structural features that determine the biochemical differences between protein variants is often a difficult challenge, requiring one to consider the entire ensemble of structures that a protein adopts. Techniques like crystallography and cryoEM sometimes reveal dramatic structural differences between protein variants that readily explain their biochemical differences. However, there are also many cases where structural snapshots do not provide a clear explanation for the differences between variants, ¹⁰ suggesting that one must consider the entire ensemble of thermally accessible configurations these proteins adopt to understand the biochemical differences between them^{11–13}. Molecular dynamics simulations can provide access to these ensembles. 14 However, there are many factors that make comparing these ensembles difficult. First of all, proteins have thousands of degrees of freedom that enable them to adopt an enormous number of different configurations^{15,16}. Moreover, two ensembles may be highly overlapping, requiring one to identify differences in the probabilities of structural features that are present in both ensembles, rather than simply identifying features that are only present in one ensemble. For example, mutations in the enzyme TEM β-lactamase were found to determine its specificity by modulating the relative probabilities of different structures, 12 but all the variants considered had a reasonable probability of adopting any of these structures.

Dimensionality reduction algorithms play a crucial role in dealing with the enormity of conformational ensembles. Many powerful algorithms have been developed and employed successfully, but the utility of each is limited by assumptions that are not universally appropriate. For example, principal component analysis (PCA)^{17,18} finds linear combinations of features that retain as much of the geometric variance in the original data as possible, effectively assuming

that large structural changes are more important than subtle ones. Unfortunately, there are many cases where this assumption is invalid, as in enzymes where arbitrary motions of a large floppy loop may dwarf subtle but functionally-relevant sidechain motions in the active site. Autoencoders¹⁹ are a more powerful alternative since they consider nonlinear combinations of features. These neural networks learn a low-dimensional projection of data—called the latent space—that is optimized to produce a high-fidelity geometric reconstruction of a protein configuration (Fig. 2.1). However, like PCA, autoencoders still focus on capturing large geometric variations. Time-lagged independent component analysis (tICA)^{20,21} is another common approach. It is similar to PCA but focuses on slowly varying degrees of freedom rather than emphasizing large geometric changes. However, there are many situations where the conformational changes of interest are fast relative to others (e.g. allostery within the native ensemble that is faster than folding and unfolding of the protein). Another recent approach, VAMPnets,²² combines ideas from autoencoders and tICA to achieve a dimensionality reduction that maps protein structures to metastable states. This allows VAMPnets to capture nonlinearities that tICA cannot, but the assumption that slowly varying degrees of freedom are more important than faster ones is still limiting in many cases. Recent work suggests supervised machine learning algorithms aid in identifying features that distinguish structural states²³. Here, we explore the idea of integrating supervised machine learning and dimensionality reduction algorithms.

We hypothesized that requiring a dimensionality reduction algorithm to predict the biochemical differences between protein variants would be a powerful means to ensure that it identifies the relevant structural differences without being misled by a priori assumptions.

Instead of assuming what type of variation is important (e.g. that large structural changes are

more important than smaller ones), such an algorithm would simply assume there are differences between two or more classes of data and then search for features that separate these classes.

To test this hypothesis, we introduce DiffNets, a dimensionality reduction algorithm that uses a self-supervised autoencoder to learn features of a protein's structural ensemble that are predictive of the biochemical differences between protein variants (Fig. 2.1). While we focus on protein variants, the algorithm should be equally applicable to other perturbations, such as understanding the impact of post-translational modifications and interactions with binding partners. DiffNets takes two inputs: 1) a set of molecular dynamics simulations for each protein variant and 2) the biochemical property of interest (e.g. stability or activity) for each variant. The algorithm then learns a low-dimensional projection (latent space) of the protein structures that is explicitly organized to separate structural configurations based on how closely they are associated with the biochemical property of interest. DiffNets achieve this by combining supervised autoencoders²⁴ with self-supervision. Supervised autoencoders are multi-task networks. Like standard (unsupervised) autoencoders, they must learn a low-dimensional projection of the data that retains sufficient geometric information to reconstruct the original high-dimensional input (Fig. 2.1, left). However, supervised autoencoders add the additional requirement that the low-dimensional projection of the data be sufficient to predict a label, in this case one related to the biochemical property of interest. This second requirement forces the dimensionality reduction to dedicate representational power to identifying degrees of freedom that are important for the label instead of focusing exclusively on large structural changes. The classification task can be based on the entire latent space to minimize assumptions, or on a subset of the inputs (e.g. the region around a mutation, as in Fig. 2.1) to focus attention on critical areas. Self-supervision provides an automated way to deal with the fact that we know the biochemical

properties of variants (i.e. their entire structural ensemble), but the association between any specific structure and that biochemical property is unknown. This problem is non-trivial because there is likely to be overlap between ensembles (i.e. structures that are visited by all variants). Therefore, classifying all structures from variants without the property of interest as different than all structures from variants with the property is likely a misleading oversimplification. To overcome this limitation, we present an expectation maximization scheme that iteratively updates training labels to identify a subset of structures that are more probable for variants with the biochemical property of interest while allowing for overlap between the conformational ensembles of different variants.

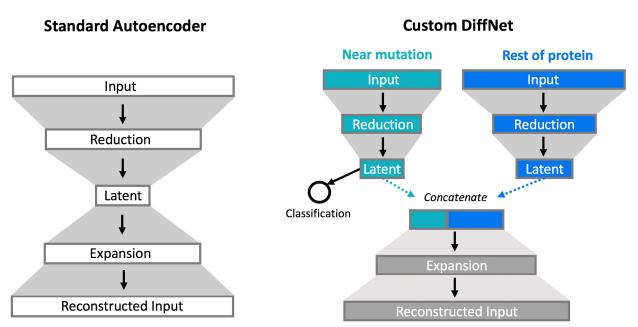


Figure 2.1 Comparison of autoencoder and DiffNet architectures.

Standard autoencoder architecture (left) and an example DiffNet architecture (right). Autoencoders have an encoder that compress the input data to a bottleneck, or latent, layer and a decoder that expands the latent representation to reconstruct the original input. The DiffNet adds a classification task to the latent space. In the example shown, the input is split into two encoders. One is a supervised encoder that operates on atoms near the mutation (cyan) and must predict the biochemical property associated with a structure. The second encoder is unsupervised and operates on the rest of the protein (blue). The latent layers from these two encoders are concatenated and trained to reconstruct the original input.

To test the performance of DiffNets, we apply them to a set of four TEM β -lactamase variants, which differ by single point mutations, and to a set of eight myosin isoforms. First, we demonstrate how the DiffNet classification task alters dimensionality reduction of protein structures compared to standard autoencoders. Then, we use DiffNets to recapitulate known differences in β -lactamase variants' folded ensembles that are predictive of changes in stability between variants. The relevant changes are geometrically subtle (< 1 Å distance change) compared to other motions, and thus, it originally took our group several months to identify them. Therefore, attempting to recapitulate this result is a challenging test case for new methods, such as DiffNets. Finally, we use DiffNets to understand the structural determinants of duty ratio (i.e. the amount of time a myosin protein spends attached to actin) among eight myosin isoforms. This is a difficult test case since small loop motions are critical for determining duty ratio, which is difficult to pick out in large (e.g. ~800 residues) myosin motor proteins. Further, the underlying amino acid sequences of isoforms are highly divergent, so success on this task would demonstrate that DiffNets are applicable to variants with more complex perturbations compared to single-point mutations.

2.3 Results

2.3.1 The DiffNet Architecture

The DiffNet architecture is based on an autoencoder, which is a deep learning framework commonly used for dimensionality reduction^{4,25–36}(Fig. 2.1). Like standard autoencoders, DiffNets connect an encoder and decoder network to compress and reconstruct input data, respectively. In our case, the input is protein XYZ coordinates (C,CA,N,CB) from a simulation frame, which are whitened for normalization (see methods). First, the encoder network transforms the input to progressively reduce the dimensionality of the input to a bottleneck layer,

called the latent space. Then, the latent space vector is used as input to the decoder network that attempts to reconstruct the original input. Mechanically, both the encoder and decoder operate via successive matrix multiplications and non-linear activation functions. DiffNets (and autoencoders) are initialized with random matrix multiplications, and the network improves by iteratively tuning the matrix values (weights) by training across many examples. Concretely, the weights are trained to minimize a loss function that measures the difference between the input and output of the model, called the input reconstruction error. Ultimately, if a DiffNet (or autoencoder) can compress and then reconstruct the original input with high accuracy, this implies that the low-dimensional latent space vector retains the salient features that describe the input.

Inspired by supervised autoencoders²⁴, DiffNets augment autoencoders with a loss function that measures how accurately the latent space vector performs a user-defined classification task (e.g. did the protein structure come from a wild-type or variant simulation?). Therefore, DiffNets must learn weights that simultaneously minimize protein reconstruction error and classification error. The constraint to minimize protein reconstruction error enforces that the low-dimensional representation of data retains a structural basis, and the classification constraint is designed to reconfigure the latent space such that data points are separated to highlight differences between datasets (e.g. biochemical differences between protein variants). While supervised autoencoders have been previously used as a way to obtain better performance on a classification task²⁴, we use the classification task to learn a more interpretable low-dimensional projection of data. Additionally, we propose an expectation maximization scheme such that classification labels are updated between DiffNet training epochs. This self-supervision provides an automated way to deal with the fact that we know the biochemical properties of

variants (i.e. their entire structural ensemble), but the association between any specific structure and that biochemical property is unknown.

The DiffNet architecture can be split to focus the classification task on a region of interest within a protein. If there is a region of interest known a priori (e.g. region around a mutation, or an enzyme active site) the input may be split into two encoder networks. In this case, only the encoder with inputs from the region of interest performs a classification task, then the latent spaces from each encoder are concatenated for input to the decoder (see Fig. 2.1). This split architecture guides a DiffNet to search in the region of interest to find differences between variants. This is a reasonable default to use when studying single point mutations as the region of a mutation is root of differences between variants. Moreover, classifying based on a region of interest does not preclude the identification of relevant distal structural differences between variants. If a mutation causes biochemically relevant differences at distal regions then these regions are inherently linked to the state of the region of the mutation and, thus, are implicitly linked to the classification task

2.3.2 The Classification Task Reorganizes the Latent Space to Emphasize Biochemically-Important Structural Features

Dimensionality reduction algorithms are only helpful for identifying differences between two classes of data if the two classes of data are separated in the latent space. Unsupervised autoencoders learn a latent representation of data that focuses on large geometric variations, so structures with large geometric differences are separated, while structures with subtle differences are close together. As a result, if biochemical differences between protein variants are related to subtle geometric changes, then the variants will be highly overlapping in the latent space and

thus, the autoencoder will fail to provide a useful way to distinguish variants. We hypothesized that augmenting a standard autoencoder with a classification task, as with DiffNets, would reorganize the latent space to highlight relevant differences between datasets, even if they are subtle structurally.

In order to test this hypothesis, we applied DiffNets and autoencoders to a set of variants of the enzyme TEM β -lactamase. β -lactamase is an enzyme that confers bacteria with antibiotic resistance by metabolizing β -lactam drugs like penicillin³⁷. Bacteria are quick to evolve new variants of TEM that have activity against new drugs, but these mutations are often destabilizing, so compensatory mutations evolve to restore stability^{38–40}. M182T is one stabilizing mutation that frequently appears in clinical isolates^{41,42}. While crystal structures of the wild-type and M182T proteins had been solved, comparing them did not provide a conclusive mechanism for stabilization capable of predicting the impact of other variants. Recently, our group combined simulations, NMR experiments, and x-ray crystallography to demonstrate that compaction of helix 9 is a structural signature that distinguishes more stable variants (like M182T) from less stable ones (Fig. 2.2). This compaction is associated with stronger h-bonds along helix 9 that stabilize this secondary structure element. Helix 9 is part of a crucial interdomain interface, so stabilizing it ultimately stabilizes the native state relative to an intermediate where one domain is at least partially unfolded. Importantly, this helix compaction includes distance changes of less than 1 Ångstrom between hydrogen bonding partners. Given that this is geometrically subtle compared to nearby loop motions, we expect that compact and extended helix states will not be well separated in the latent space of a standard autoencoder. However, we do expect that a DiffNet trained to classify compact and extended helix states will learn a latent space that separates these states.

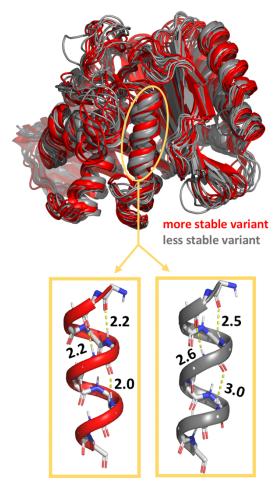


Figure 2.2 Helix 9 compaction distinguishes structural ensembles.Structural configurations sampled from molecular dynamics simulations of wild-type TEM β-lactamase (grey) and an M182T variant (red) that is far more stable. Helix 9 is circled in yellow and shown below in a compact configuration (left, red) and a more extended configuration (right, grey). Hydrogen bond distances are shown in Ångstroms.

To evaluate if the DiffNet classification layer alters the latent space in a way that helps identify differences between two classes of data, we compared the latent space of DiffNets to the latent space of unsupervised autoencoders after training on a dataset that includes two classes of data distinguishable by a subtle difference in helix 9 compaction. From the original set of 650,210 structures (from wild-type and M182T simulations) we curated a dataset of 178,402 simulation frames from wild-type and M182T simulations where half of the frames have a

compact helix 9 (helix compaction criteria described in Methods) and half have a more extended helix. Then, we trained DiffNets and unsupervised autoencoders using a split architecture described in the methods and visualized in Figure 2.1b. The DiffNets and autoencoders we trained were identical, except the DiffNet has an additional output layer such that it has to classify helix 9 as compact or extended in addition to reconstructing protein structures. The classification labels are not updated with expectation maximization in this case. This dataset was selected specifically to evaluate how the classification task of the DiffNet alters the dimensionality reduction compared to a standard autoencoder. In a normal setting we would not have a priori knowledge about the importance of helix 9 compaction. However, this is an important test to determine if adding a classification task can reorganize the latent space to highlight differences between datasets, which is a property that DiffNets will ultimately need to identify differences between variants.

Requiring DiffNets to perform a classification task in tandem with dimensionality reduction successfully reconfigures the latent space to disentangle compact helix configurations from more extended helix configurations. First, we note that DiffNets and unsupervised autoencoders have similar ability to reconstruct protein structures (~1 Ångstrom error - see Figure 2.3) using as few as three latent variables and as many as fifty, which is in line with another study reporting autoencoder reconstruction error²⁷. To compare latent spaces, we analyze a split architecture that has twenty-five latent variables including three in encoder A (which receives input including helix 9 and performs the classification task in the DiffNet) and twenty-two in encoder B (takes input from the rest of the protein). This architecture provides a low reconstruction error (< 1 Ångstrom) and few enough latent variables so that all dimensions in encoder A's latent space can be visualized. In the unsupervised autoencoder, simulation frames

of compact and extended helices are overlapping in encoder A's latent space (Fig. 2.4a). This demonstrates that training an unsupervised autoencoder on two classes of data does not necessarily yield a latent representation that provides any insight into how the two classes of data are different. To explore this point further, we held the autoencoder's latent space constant and then trained it to classify whether a structure has a compact or extended helix 9 (i.e. performed logistic regression). The resulting receiver operating characteristic (ROC) curve, which measures classification performance, shows a classification performance similar to random guessing (area under the curve [AUC]=0.54) providing further evidence that the latent representation does not help distinguish the two classes of data. In contrast, the DiffNet encoder A latent space clearly separates the two classes of data (Fig. 2.4a) and has excellent performance classifying compact and extended helix states (AUC=0.91, Fig. 2.4b). This result demonstrates that adding a classification component to the learning task provides a powerful means to learn a low dimensional representation that highlights crucial differences between datasets. It follows that DiffNets trained with a classification task that must predict a biochemical property should learn a low dimensional representation of data that highlights structural features that are predictive of biochemical differences between protein variants.

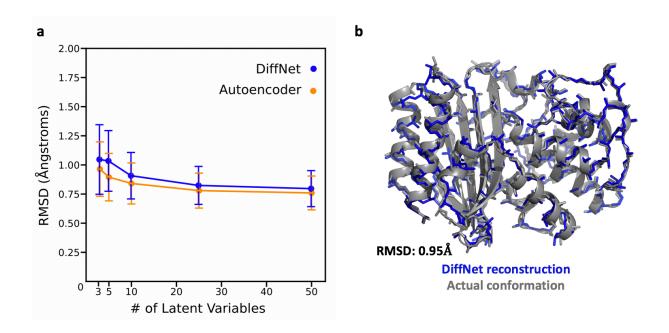


Figure 2.3 DiffNets accurately reconstruct protein structures.

Autoencoders and DiffNets can both compress protein structures and then reconstruct them. (a) Reconstruction error plots showing the root-mean-square deviation (RMSD) between a protein structure from simulation and the corresponding protein structure generated by unsupervised autoencoders (yellow) or DiffNets (blue). One of every ten structures from wild-type and M182T simulation data was used (n=65,210) and the standard deviation is shown with error bars. (b) Structure representing the difference between a structure generated by the DiffNet (blue) vs. the actual conformation from simulation (grey) when training with 3 latent variables.

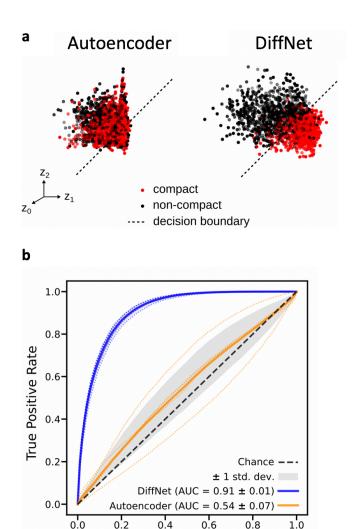


Figure 2.4 Classification task reorganizes DiffNets latent space.

Adding a classification component to the learning task (as in DiffNets) results in a latent representation that separates different datasets more clearly than an unsupervised autoencoder. (a) Simulation frames that have a compact helix (red) and an extended helix (black) are projected onto the three-dimensional latent space learned by an unsupervised autoencoder (left) and a DiffNet (right). The decision boundary (black dotted line) indicates the plane that each neural network uses to separate compact helix states from extended helix states. (b)Receiver operating characteristic (ROC) curve showing the average classification performance of the unsupervised autoencoder (dark yellow) and the DiffNet (dark blue) as well as the performance for each of the 5 folds of cross validation (faded dotted lines). Mean area under the ROC curve (AUC) is shown in the bottom right corner with the standard deviation across the 5-folds of training.

False Positive Rate

2.3.3 Self-Supervised DiffNets Learn Structural Signatures Associated with Protein Stability

While the classification task can help DiffNets learn a useful dimensionality reduction, realizing this potential is non-trivial because we know the biochemical properties of variants (e.g. their

entire structural ensembles) but not individual structures. The simplest approach to providing these classification labels would be to assign ones to structures from simulations of variants with the biochemical property of interest and zeros to structures from simulations of variants without the property. However, it is likely that variants fall on a continuum rather than having a biochemical property or not, that their conformational ensembles overlap, and that only a subset of conformations are relevant for determining the property of interest.

This problem is similar to multiple instance learning. During multiple-instance learning, learners are given bags of training examples where each bag is labelled negative, indicating that the bag contains all negative examples, or positive, indicating that there are at least some positive examples in the bag. The learner then must figure out how to label all of the individual instances as positive or negative by identifying features that are consistent in positive bags, but absent in negative bags. This is similar to our situation where we know the biochemical property of each protein variant (i.e. negative bag or positive bag), but we do not know if a given structural configuration is associated with a biochemical property, or inconsistent with a biochemical property.

We propose a self-supervised approach for learning the relationship between individual structures and the biochemical property of interest using an iterative expectation maximization algorithm based on work from Zaretski et. al.⁴³ Expectation maximization is a statistical method that allows the parameters of a model to be fit, even when the outputs of the model cannot be observed directly in the training data⁴⁴ (i.e. when they are hidden). In our case, the hidden variables are labels for each structure that specify the probability that a structure is associated with the biochemical property of interest. These labels are initially set to ones for all structures from variants with a given biochemical property (e.g. more stable β-lactamase variants) and

zeros for variants without that property (e.g. variants with lower stabilities). Then the expectation maximization algorithm iteratively alternates between a maximization step and an expectation step to identify a self-consistent set of labels. During the maximization step, a DiffNet is trained to predict the current labels for each structure. Then, the expectation step refines the training labels by computing the expected values of the labels, y, using the output from the DiffNet, \hat{y} , conditioned on constraints about what fraction of structures from each variant we expect to be associated with the property of interest. This constraint provides a way to enforce that more high probability values are assigned to structures from variants with the biochemical property. The expectation is the probability-weighted average of all binary realizations of binomial distributions parameterized by \hat{y} , excluding binary realizations that do not meet the constraint. Formally, we update training labels as,

$$y_i = E[\hat{y}_i \mid S_L \le \hat{y}_r \le S_U] \tag{1}$$

$$= P(\hat{y}_i \text{ is } 1) * \left(\frac{P(S_L - 1 \le \hat{y}_r - \hat{y}_i \le S_U - 1)}{P(S_L \le \hat{y}_r \le S_U)} \right)$$
 (2)

where y_i is the updated label for each individual frame, \hat{y}_i is the DiffNet output, S_L and S_U are the lower and upper bounds on how many conformations in a batch are associated with the biochemical property, \hat{y}_r is the sum of the binary outcomes of a batch which contains conformation i, $P(S_L - 1 \le \hat{y}_r - \hat{y}_i \le S_U - 1)$ is the probability that the number of conformations in a batch is within the limits if conformation i is ignored, and $P(S_L \le \hat{y}_r \le S_U)$ is the probability that the number of conformations in a batch is within the limits, including

conformation *i*. Ultimately, the desired outcome is that the expectation maximization algorithm redistributes training labels from all 0s and 1s for simulation frames of variants without and with a biochemical property, respectively, to values that indicate the probability that a given structural configuration is associated with the biochemical property of interest. This mechanism is self-supervised since the training labels are learned by the algorithm, rather than explicitly curated.

To test this approach, we trained a self-supervised DiffNet to identify structural preferences that distinguish two highly stable β-lactamase variants (M182T and M182S) from two less stable variants (wild-type [WT] and M182V). In this case, the DiffNet receives no a priori information about features, like helix 9 compaction, that are associated with increased stability in M182T and M182S. If self-supervision of DiffNets works as expected, then training should produce a latent space where it is easy to identify the structural features that are associated with the stability of M182T and M182S, relative to WT and M182V. For example, we expect to see structural configurations with a compact helix 9 in one region of the latent space and structures with a more extended helix elsewhere. Beyond helix compaction, DiffNets may even capture additional structural features that were missed in our previous manual analysis. To evaluate if the DiffNet learns these biochemically relevant structural differences between variants, we trained a DiffNet on 6.5µs of simulation data for each variant: M182T, M182S, WT, and M182V. All frames from M182T and M182S (highly stable variants) were initially assigned classification labels of 1, and simulation frames from M182V and WT were initially assigned 0s. During the expectation maximization procedure, we calculate the expected values (updated labels) conditioned on the constraint that 0-30% of less stable variants frame are likely to be stabilizing, and 60-90% of frames for highly stable variants. In general, it should be sufficient to base bounds on qualitative a priori knowledge rather than precise, quantitative information. In

this case, we chose these bounds as a way to allow overlap between ensembles, but still provide a clear signal to distinguish more and less stable variants. Empirically, we find that DiffNets are robust across a wide range of bounds (Fig. A.1.1).

Expectation maximization aids the DiffNet in learning a low-dimensional representation that accurately identifies that helix 9 compaction is associated with highly stable variants. First, we trained two supervised autoencoders (one with and one without expectation maximization) and compared the distribution of output classification labels. Without expectation maximization, almost all structures from more stable variants have output labels close to 1, and structures from less stable variants have output labels close to 0 (Fig. 2.5a). This is at odds with the fact that there is structural overlap between the ensembles. It indicates that the supervised autoencoder essentially memorizes which ensemble each structure comes from instead of learning a useful association between individual structures and stability. In contrast, when expectation maximization is applied the output labels span the full spectrum from 0-1 for each variant (Fig. 2.5b), which is consistent across a wide range of expectation maximization bounds (Fig A.1.1). Moreover, as the labels increase from 0 to 1, helix 9 compaction smoothly decreases, which indicates that DiffNets learn a latent space with a continuum of structures that are less/more closely associated with stability (Fig. 2.5c). Without expectation maximization, the extreme labels (i.e. 0,1) track well with helix stability, but structures labelled between 0.1 and 0.9 do not show a clear trend of helix compaction.

Using DiffNets to predict on a variant outside of training provides further support that expectation maximization aids in learning structural features associated with stability. We compared each model's ability to predict the stability of a less stable variant not seen during training (M182N), and we find that this prediction is improved when expectation maximization

is applied (Fig. A.1.2). This suggests expectation maximization helps DiffNets hone in on biochemically relevant structural features, and that DiffNets could be used as a predictive tool. However, we caution that autoencoders will fail anytime they are applied to data that is highly dissimilar from the training set, so a DiffNet will not perform well on new variants that visit conformations not visited in the training set. Future studies would be necessary to optimize DiffNets for prediction and should be evaluated against related methods such as by Riesselman et al⁴⁵.

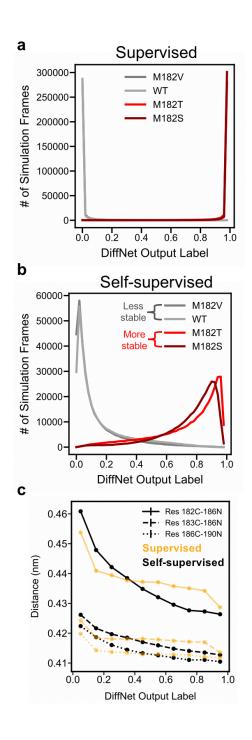


Figure 2.5 DiffNets learn helix 9 compaction is important for distinguishing variants.

Self-supervision improves the DiffNet's ability to organize structural configurations based on their biochemical property. Histogram showing DiffNet output labels across all simulation frames from M182T and M182S (red – highly stable variants in training set) versus WT and M182V (grey – less stable variants in training set) for a supervised autoencoder (a) and a self-supervised autoencoder (b). (c) Three key hydrogen bond lengths in helix 9 as a function of the DiffNet output label (n=1,300,420) (yellow – supervised, black – self-supervised), which ranges from zero for structures associated with low stability to one for structures associated with high stability. The distances are between the carbonyl carbon of the i'th residue and the nitrogen of the (i+4)'th residue. Standard error bars are not visible since the standard error is smaller than scatter points.

While many deep learning approaches are criticized for their lack of interpretability, the DiffNet architecture provides opportunities to understand what the network learned, which provides biophysical insight. To automate DiffNet interpretation, we measured all inter-atom distances within 1nm of the mutation using 2000 cluster centers calculated from all simulations and then measured the linear correlation between each distance and the DiffNet output label. We plot the top 1% of distances correlated with the DiffNet output label to visualize the conformational changes that the DiffNet views as important for distinguishing stable variants from less stable variants. Encouragingly, the distance correlations strongly point to helix 9 compaction as an important feature of more stable variants (Fig. 2.6). While the helix compaction is striking, DiffNets also captured other trends that our previous computational analyses did not detect. For example, our NMR data suggested that the packing between helix 9 and adjacent β-sheet differs in more stable vs less stable variants,⁵ but our computational analysis did not detect a clear trend. On the same simulation dataset, the DiffNet clearly learns that this interface becomes more tightly packed for more stable variants (Fig. 2.6). Specifically, the DiffNet analysis suggests more stable variants have tighter packing at the helix 9 and β-sheet interface (Fig. 2.6b). Often times the important features that distinguish protein variants can be complicated and, therefore, easily missed even with months of analysis. DiffNets can learn complicated features and help automate the process of identifying biochemically relevant structural features that distinguish protein variants.

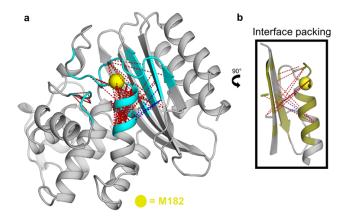


Figure 2.6 Automated feature detection reveals what DiffNets learn.

Visualization of the features that DiffNets find important for increased stability of M182T and M182S variants. (a) Crystal structure of TEM β -lactamase (PDB ID: <u>1JWP</u>) overlaid **with** dotted lines that indicate distances between two atoms that change in a way that is strongly correlated with an increased DiffNet output label. Red indicates the atoms move closer together as the output label increases, blue indicates atoms moving away from each other. The mutated residue is highlighted with a yellow sphere. Protein atoms are colored cyan if they are near the mutation, which indicates that they were included in the classification task and considered for the distance correlation calculation. (b) Rotated inset of (a) showing DiffNet predicted packing at the interface of helix 9 and the adjacent β -sheet. Residues with chemical-shift perturbations in M182S relative to wild-type are shown in deep olive.

2.3.4 DiffNets works for other proteins and more divergent sequences

In order to explore the broad applicability of DiffNets, we also trained a self-supervised DiffNet to identify structural features that distinguish high duty myosin motor domains from low duty myosins. Myosins are a ubiquitous class of motor proteins that perform an extraordinary diversity of functions despite sharing a common mechanochemical cycle. ⁴⁶ In order to perform roles as diverse as muscle contraction and intracellular transport, myosins have precisely tuned their duty ratios, or the fraction of time a myosin spends attached to actin during one full pass through its mechanochemical cycle. Recent work from Porter et al. ⁴⁷ suggests that the conformational ensemble of the active site P-loop encodes duty ratio through the balance of nucleotide favorable and unfavorable states. Specifically, low duty motors have an increased propensity to adopt a P-loop "up" state, where the S180 carbonyl group sterically occludes

nucleotide binding, whereas high duty motors favor a "down" state, where the P-loop is nucleotide compatible (see Fig. 2.7b).

We trained a DiffNet using molecular dynamics simulation data from the active sites of four low duty motors and four high duty motors to see if we could recapitulate the trend between P-loop dynamics and duty ratio (Fig. 2.7a). Importantly, this test case is especially challenging because it includes eight different proteins with a low degree of sequence conservation in the area of interest (i.e. 34% of residues were perfectly conserved within the training area). Low duty motors were given an initial label of zero and high duty motors were initially given a label of one.

A DiffNet trained to distinguish high and low duty myosin motors substantiates previous work that identified P-loop dynamics to be important for distinguishing these myosins. To determine if a DiffNet captures the importance of P-loop "up" and "down" states, we examined structures with low and high DiffNet output labels (i.e. predicted low and high duty respectively) from a single isoform. We saw a consistent trend in the orientation of the S180 carbonyl group, where structures with high DiffNets labels are in the "down" orientation and structures with low labels are in the "up" orientation (see Figure 2.7b). This indicates that the DiffNet correctly learned that high-duty motors are more likely to be in the "down" state and vice-versa. To more precisely quantify this trend, we examined the correlation between DiffNet output labels and nucleotide compatibility (as defined previously⁴⁷) for all frames. We find that as the DiffNet output labels increase (i.e. shift from low duty to high duty), there is a concurrent increase in the ratio of nucleotide favorable:unfavorable states (Fig. 2.7c).

Automated interpretation of a DiffNet captures the importance of P-loop dynamics and suggests other order parameters that may distinguish high and low duty myosins. Similar to

Figure 2.6, we calculate the correlation between interatomic distances and DiffNet output labels for all 139,129 distances around the active site (Figure 2.7a) and then project the top 100 correlated distances onto the structure (Fig. 2.7d). This analysis finds 78 distances between the P-loop and the loop connecting the third beta sheet with the SH2 helix (referred to as the β3-SH2 loop), again highlighting that the DiffNet learns that P-loop dynamics are important for discriminating high and low duty motors. We compared this result to a model trained without expectation maximization and find that expectation maximization improves the quality of this analysis. Specifically, changes in Ser180 are strongly detected when expectation maximization is applied, but without expectation maximization these changes are not detected at all (Fig. A.1.3). The DiffNet also infers that high duty motors are more likely to occupy states where the P-loop is near to the β3-SH2 loop, and indeed this finding is confirmed using previously published Markov State Models of the motor domains (see Figure 2.7e). Since the β3-SH2 loop is below the P-loop, this provides further evidence that the DiffNet is correctly learning that high duty motors prefer the "down" state. While this order parameter is the predominant feature of this analysis, the DiffNet suggests that other distances may be important for distinguishing high and low duty motors. In particular, there are two residues on switch-II with distances that are strongly correlated with the DiffNet label indicating that conformational changes in switch-II may be important for determining the duty ratio, which is consistent with previous findings⁴⁸ (see Fig. A.1.4).

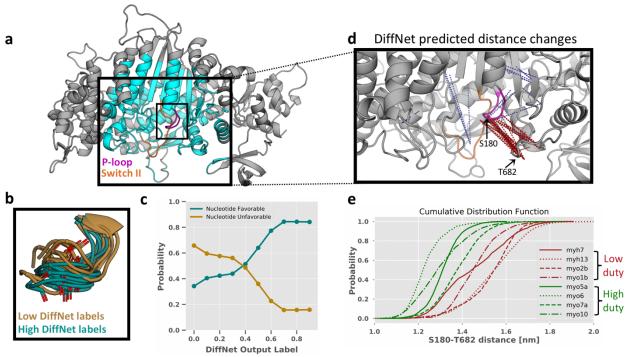


Figure 2.7 DiffNets capture known P-loop motions that distinguish myosin isoforms.

DiffNets capture the importance of P-loop motions in distinguishing high and low duty myosin motor proteins. (a) Structure of a myosin motor protein (PDB ID: 4PA0) showing the DiffNet classification region (cyan), the P-loop (magenta), and Switch-II (Orange). (b) Twenty states predicted by the DiffNet as high duty (teal) and low duty (dark gold). Predicted high duty states are mostly in a nucleotide compatible, P-loop "down" conformation and vice-versa for predicted low duty states. (c) Percentage of nucleotide favorable (teal) and unfavorable (dark gold) states as a function of the DiffNet output label, measured with 10 equally spaced bins with labels spanning 0-1. Most structures with low Diffnets labels are nucleotide unfavorable, and vice-versa. (d) Inset of the myosin active site. Dotted lines indicate distances between two atoms that change in a way that is strongly correlated with an increased DiffNet output label. Red indicates the atoms move closer together as the output label increases, blue indicates atoms moving away from each other. (e) Cumulative distribution function showing the distance between S180 (P-loop) and T682 (β 3-SH2 loop). Probabilities come from a previously published MSM⁴⁷. This distance clearly separates high and low duty motors (green and light brown, respectively) as predicted by the DiffNet in (d).

2.4 Conclusions

We have introduced DiffNets, a deep learning framework for identifying the structural signatures that are predictive of biochemical differences between protein variants from molecular dynamics simulations. Such simulations contain valuable information about the structural mechanisms that determine proteins' biochemical properties. However, extracting this insight is often difficult because of factors like the high dimensionality of the spaces involved and overlap between the

structural ensembles for different variants. Our results suggest that self-supervised DiffNets learn a low-dimensional latent representation of protein structures that separates them based on their association with biochemical properties, such as higher or lower stability. This success relies on two key innovations. First, performing dimensionality reduction simultaneously with a classification task helps yield a latent representation that organizes protein structural configurations based on their association with biochemical properties. Second, challenges with labelling each structure with a biochemical property can be overcome using an expectation maximization scheme inspired by multiple instance learning.

As a proof of principle, we demonstrated that DiffNets automatically identify structural changes that explain biochemical differences between variants in several systems including β -lactamase and myosin proteins. Success identifying helix 9 compaction (< 1 Å) as an important distinguishing factor between β -lactamase variants demonstrated that DiffNets finds biochemically relevant structural features even if they are geometrically subtle relative to other structural fluctuations in the protein. Success identifying the importance of P-loop dynamics for determining the duty ratio across myosin isoforms demonstrated that DiffNets is generalizable to large proteins (~800 residues) with low sequence conservation. Looking ahead, we expect the same architecture to be applicable to other perturbations, such as post-translational modifications or the presence/absence of a binding partner.

While these results are promising, future work can be done to expand the utility of DiffNets further. For example, the DiffNet architecture is not translationally, nor rotationally, invariant, which means the results depend on the quality of the initial alignment of simulations. Future work exploring equivariant architectures may improve DiffNets. Additionally, the current study included an abundance of data, so there were no optimizations for working with small

datasets. It is yet to be seen how well DiffNets performs on smaller (sub-microsecond) datasets. Lastly, after training a DiffNet it is possible to use the model to predict the biochemical property of a variant for which the property has not been determined experimentally. Toward this end, we showed that DiffNets accurately classified the stability of a β-lactamase variant (M182N) that was not seen during the training. However, accurate predictions will require that the variants of interest have high conformational overlap, and future studies are required to optimize a model for this task.

2.5 Methods

2.5.1 MD Simulations

All molecular dynamics simulation data were generated in previous manuscripts by Zimmerman et. al⁵ and Porter et. al⁴⁷. Briefly, all simulations were run with Gromacs 5.1.1 at a temperature of 300K using the AMBER03 force field with explicit TIP3P solvent^{49,50}. β-lactamase simulations were initialized from the TEM-1 β-lactamase crystallographic structure (PDB ID: 1JWP [https://www.rcsb.org/structure/1JWP])³⁹ and ran at 300K using the AMBER03 force field with explicit TIP3P solvent^{49,50}. Each variant, wild-type, M182V, M182T, M182S, and M182N was simulated for 6.5 μs including 4 μs of FAST-RMSD adaptive sampling⁵¹ and 2.5 μs of conventional sampling. Conformations were stored every 20 ps. Myosin simulations were performed mostly on Folding@Home⁵² to obtain ~2 milliseconds of total sampling across four low duty (MYH13, MYH7, MYH10, and MYO1B) and four high duty motors (MYO5A, MYO6, MYO7A, and MYO10), where the initial structures were built from homology models in SWISS-MODEL⁵³ using the 4PA0 [https://www.rcsb.org/structure/4PA0]⁵⁴ as a guide template structure.

2.5.2 DiffNet Model

DiffNets are neural networks with a supervised autoencoder architecture (as shown in Figure 2.1). These models take as input a vector of features that describe a protein structural configuration and output a score which indicates how closely a structure is associated with a certain biochemical property, as well as, a vector that matches the input vector (i.e. reconstructs a protein structure).

2.5.3 EM algorithm

The goal of the algorithm is to find a vector K, that maps each structure to a value between 0 and 1 that maps to the biophysical property of the structure (e.g. stability). We initialize K with all 1s for structures from variants with the biophysical property of interest, and all 0s for structures from variants without the biophysical property of interest. Then, we alternate between M- and E-steps to update the vector K. First, the M-step fits a neural network using K as classification targets. Next, the neural network outputs a vector of scores for structures, Y. Then, we apply an E-step to update the values in K. Specifically, we compute the expected value of each structure where we treat a set of structures as binomial random variables parameterized by Y, conditioned on user-defined bounds on the number of successes (i.e. structures with the biochemical property) for each variant. The expected values are computed as the probability-weighted average of all binary realizations of binomial distributions parameterized by Y that are within the user-defined bounds. These expected values provide an updated K, allowing us to repeatedly iterate between M- and E- steps. We refer the reader to Appendix and our previous work for a more thorough discussion of the algorithm.

2.5.4 Featurization

Simulation data was preprocessed before becoming input to the DiffNets. Simulation trajectories and the original crystallographic structure (PDB ID: 1JWP

[https://www.rcsb.org/structure/1JWP]) are stripped down to the XYZ coordinates of the protein backbone without carbonyl oxygens (C, CA, CB, and N). Then, the trajectories are centered at the origin and aligned to the crystallographic structure. Next, we follow a procedure similar to Wehmeyer and Noe³¹ to mean-shift the XYZ coordinates to zero, followed by whitening. First, we mean shift,

$$x^{mean-free} = \sum_{i=1}^{N_t} x_i - \bar{x} \tag{3}$$

where $x^{mean-free}$ is the mean-shifted trajectory of XYZ coordinates, x_i is a single frame with XYZ coordinates, \bar{x} is the mean of the XYZ coordinates across all trajectories, and N_t is the number of frames in all trajectories.

Next, we whiten the data,

$$\tilde{\chi} = C_{00}^{-\frac{1}{2}} \chi^{mean-free} \tag{4}$$

where \tilde{x} is the whitened trajectory of XYZ coordinates and C_{00} is the covariance matrix for the XYZ coordinates. Whitening decorrelates the inputs and adjusts their variance to be unity. After

whitening, we use one out of every ten simulation frames for each epoch of DiffNet training. In practice, whitening and unwhitening of the data is performed on the input XYZ coordinates directly in the DiffNet with frozen (untrainable) weights. For myosin, we subsampled the data to use only one of every ten simulation frames.

2.5.5 Classification Targets

To train the model we need a target for each protein structural configuration. We assign initial, binary targets based on the observed biochemical property (e.g. 1s for more stable variants, 0s for less stable variants). Our assumption that individual configurations can be mapped to biochemical properties is consistent with studies that attribute specific structural states to a biochemical property (e.g. an enzyme primed for catalysis) and designate other individual structural states as being incompatible with a biochemical property (e.g. an enzyme in an inactive state). Next, we iteratively update the initial labels with an expectation maximization algorithm (described above). This relaxes the labels such that structural configurations are on a continuum. This effectively turns the problem into a regression problem instead of a classification problem, which is consistent with the observation that most biophysical observables are on a continuum.

2.5.6 Neural Network Training

We trained DiffNets with three loss functions to minimize protein reconstruction error (ℓ_{Recon}), minimize feature classification error (ℓ_{Class}), and minimize the correlation of latent space variables (ℓ_{Corr}).

$$\mathcal{L}_{DiffNet} = \ell_{Recon} + \ell_{Class} + \ell_{Corr}$$
 (5)

47

The reconstruction loss term attempts to tune the network weights to properly reconstruct the original XYZ coordinates of the protein. This loss combines an absolute error (L1), which funnels reconstructions to the proper XYZ coordinates, and a mean-squared error (L2) to strongly discourage outliers. Explicitly,

$$\ell_{Recon} = \frac{1}{N_b} \sum_{i=1}^{N_b} \frac{1}{N_n} \sum_{j=1}^{N_n} \left[\left| x_{ij} - \hat{x}_{ij} \right| + \left(x_{ij} - \hat{x}_{ij} \right)^2 \right]$$
 (6)

where N_n is the number of output nodes (all XYZ coordinates), N_b is the number of examples in a training batch, x_{ij} is a target value (actual XYZ coordinate), and \hat{x}_{ij} is the output value from the DiffNet.

The classification error is a binary cross entropy error that penalizes misclassifications by the latent space. This classification loss attempts to constrain the latent space to learn a dimensionality reduction that can also classify a biophysical feature. Explicitly,

$$\ell_{Class} = \frac{1}{N_b} \sum_{i=1}^{N_b} y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)$$
 (7)

where N_b is the number of examples in a training batch, y_i is the target value, a binary value indicating if a simulation frame has a specific feature or not, and \hat{y}_i is the output of the classification layer by the DiffNet.

Finally, we include a loss function to minimize the covariance between latent space variables. This loss takes the form of

$$\ell_{Corr} = \sum_{i \neq j} Cov(z_i, z_j)^2$$
 (8)

where $Cov(z_i, z_j)$ is the covariance matrix of the latent vector, Z, across all N_b samples in a training batch. We reason that preventing redundancy in latent variables should maximize the amount of information one can gain in a small number of variables. Ideally, this sets us up to use just a few latent variables and still have a rich amount of information. With fewer latent variables, models are generally more interpretable.

Our training procedure uses several training iterations to progressively build in hidden layers of the DiffNet. First, we train a minimal version of a DiffNet. Explicitly, the encoders have an input layer and a reduction layer with a four-fold reduction in variables. There is no further reduction to a bottleneck layer. Instead, the decoder takes the reduction layer as input and passes it to an output layer. Training this simplified autoencoder is an easier task than training a full DiffNet because the dimensionality reduction it performs is modest. It has ~an order of magnitude more dimensions to explain the original data compared with a true bottleneck layer. We reason that this can generate useful priors for what the reduction layer should capture. In our second pretraining procedure, we freeze those priors and add the bottleneck layer in to train the full DiffNet. Therefore, this second pretraining step concentrates its representational power on

tuning how to properly reduce from the reduction layer to the bottleneck layer. Finally, we unfreeze the priors and train the full DiffNet to polish all weights. Each of these three procedures undergoes 20 training epochs. In the self-supervised setting, classification labels are updated using expectation maximization after each training epoch.

All training was performed in PyTorch 1.1⁵⁵. Training on ~120,000 simulation frames of β-lactamase takes under one hour on a single AMD Vega 20 GPU. Training with expectation maximization approximately doubled the training time for DiffNets trained on TEM. We used the Adam optimizer with a learning rate of 0.0001 and a batch size of 32.

We performed limited hyperparameter tuning to arrive at our final models. We found that the DiffNet performance was robust across a wide range of latent variables (Figure 2.3) and expectation maximization bounds (Fig. A.1.1, Fig. 2.5b). To choose a final number of latent variables, we chose the minimum number where reconstruction error no longer showed qualitative improvement. Additionally, we saved a trained model after every epoch of training and ultimately used the model that showed the best reconstruction performance on a validation set that contained 10% of the data.

2.5.7 Reconstruction Experiment

To analyze DiffNet reconstruction error (Figure 2.3), we trained on five architectures where we varied the numbers of latent variables. All architectures split the input (as in Figure 2.1b) such that any atom (C, CA, N, CB) within 1nm of residue 182 (source of single point mutation – colored cyan in Fig 2.6) was included in encoder A, while the rest of the protein was included in encoder B. Encoder A reduced down to 1, 2, 3, 5, and 10 latent variables, while encoder B reduced down to 2, 3, 7, 20, and 40 latent variables. After training, we use the neural networks to

reconstruct the protein structure from 1 of every 100 simulation frames and compute its rootmean squared deviation from the actual structure obtained via simulation.

2.5.8 Classification Labels

To provide classification labels for Figure 2.4, we designated simulation frames as "compact helix" or "extended helix" based on a previous manuscript that identified three key hydrogen bond distances in Helix 9 that distinguish stabilizing variants from nonstabilizing variants (Res 182-186, Res 183-187, and Res 186-190)⁵. Specifically, we label helix 9 compact if the distance between the backbone nitrogen and the carbonyl oxygen is less than 4.2 Ångstroms for all residue pairs listed, and we label it extended otherwise.

2.5.9 β-lactamase expectation maximization experiment

When training on β-lactamase with expectation maximization (Fig 2.5, 2.6) we trained a split architecture DiffNet consisting of 2 encoders and 2 latent spaces (as visualized in Fig. 2.1). The input to encoder "A" is all XYZ coordinates within 1nm of residue 182 (1nm region around the mutation). The input to encoder "B" is the XYZ coordinates from the rest of the protein. These encoders reduce the input to 4 and 26 latent variables, respectively (30 total latent variables split proportionally into latent A and latent B based on the number of atoms input into each encoder). After training, we applied the trained DiffNet to all simulation data to obtain DiffNet output labels. These output labels can be thought of as a proxy for latent A (region around the mutation) as the output label is simply a linear combination of the values in latent A (then scaled between 0 and 1 using the PyTorch sigmoid activation function). We bin all structures into 10 equally

spaced bins from 0-1 based on their DiffNet output label. Then, we measure the average distance for Res 182-186, Res 183-187, and Res 186-190 in each bin (Figure 2.5a). We calculated an AUC evaluating how well the DiffNet output labels classify compact helix 9 states from extended stated where the labelling criteria is explained in the above section. To find distance changes that are correlated with changes in the DiffNet output label (as shown in Figure 2.6), we first cluster the simulation data into 2000 clusters using a hybrid k-centers and k-medoids approach with our open-source python package, Enspara⁵⁶. Then, we enumerate all possible distance pairs between atoms in encoder A (i.e. within 1nm of the mutation). For each distance pair, we perform a linear regression between the distance and the DiffNet output label across all 2000 cluster centers. We then select the distance pairs with the highest correlation coefficients (top 1%) and visualize them in PyMol (Figure 2.6).

2.5.10 Myosin expectation maximization experiment

When training on myosin (Fig. 2.7) we used an architecture with a single encoder (i.e. not split) that received C, CA, N, and CB atoms as input within a 2.25 nm radius around the P-loop (specifically residue S180, *Myh7* numbering). We used 50 latent variables. All frames from low duty motors were initially assigned classification labels of 0, and simulation frames from high duty motors were initially assigned 1s. During the EM procedure, we set bounds of 10-40% for low duty motor frames and 60-90% for high duty motor frames. To find distance changes that are correlated with changes in the DiffNet output label, we copied the scheme described in the previous section. To identify P-loop orientations with high/low DiffNet labels, we selected the 10 structures with DiffNet labels closest to 0.03 and 0.7 from a single isoform (Myh7). To calculate the cumulative distribution function in Fig. 2.7e, we used a previously published MSM.

Specifically, for each cluster center in the MSM, we measured its distance and weighted the distance by its equilibrium population. Lastly, for Fig. 2.7c, we grouped structures as nucleotide favorable/unfavorable as defined in a previous manuscript⁴⁷.

2.5.11 Code Availability

Data normalization and DiffNets training with, or without, expectation maximization is freely available on GitHub at https://github.com/bowman-lab/diffnets.

2.5.12 Data Availability

The datasets are not publicly deposited because they are several terabytes in size. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. We expect that it should take several business days to share the data upon a particular request. Once shared, we will not enforce any limitations for how the data may be used.

Bibliography

- 1. Erickson RP. Somatic gene mutation and human disease other than cancer: An update.

 Mutat Res Rev Mutat Res. 2010. doi:10.1016/j.mrrev.2010.04.002
- 2. Krawczak M, Ball E V., Fenton I, et al. Human Gene Mutation Database A biomedical information and research resource. *Hum Mutat.* 2000. doi:10.1002/(SICI)1098-1004(200001)15:1<45::AID-HUMU10>3.0.CO;2-T
- Davies J. Origins and evolution of antibiotic resistance. *Microbiologia*. 1996.
 doi:10.1128/mmbr.00016-10
- Sultan MM, Wayment-Steele HK, Pande VS. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *J Chem Theory Comput*. 2018. doi:10.1021/acs.jctc.8b00025
- Zimmerman MI, Hart KM, Sibbald CA, et al. Prediction of New Stabilizing Mutations
 Based on Mechanistic Insights from Markov State Models. ACS Cent Sci. 2017.
 doi:10.1021/acscentsci.7b00465
- 6. Perryman AL, Lin J-H, McCammon JA. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. *Protein Sci.* 2004. doi:10.1110/ps.03468904
- 7. Schwantes CR, Shukla D, Pande VS. Markov state models and tICA reveal a nonnative folding nucleus in simulations of NuG2. *Biophys J.* 2016. doi:10.1016/j.bpj.2016.03.026
- 8. Sang D, Pinglay S, Wiewiora RP, et al. Ancestral reconstruction reveals mechanisms of erk regulatory evolution. *Elife*. 2019. doi:10.7554/eLife.38805
- 9. Razavi AM, Voelz VA. Kinetic Network Models of Tryptophan Mutations in β-Hairpins Reveal the Importance of Non-Native Interaction. *J Chem Theory Comput.* 2015.

- doi:10.1021/acs.jctc.5b00088
- 10. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol*. 2009. doi:10.1038/nrm2805
- 11. James LC, Tawfik DS. Conformational diversity and protein evolution A 60-year-old hypothesis revisited. *Trends Biochem Sci.* 2003. doi:10.1016/S0968-0004(03)00135-X
- Hart KM, Ho CMW, Dutta S, Gross ML, Bowman GR. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nat Commun*. 2016. doi:10.1038/ncomms12965
- 13. Knoverek CR, Amarasinghe GK, Bowman GR. Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities. *Trends Biochem Sci.* 2019. doi:10.1016/j.tibs.2018.11.007
- 14. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol*. 2002. doi:10.1038/nsb0902-646
- 15. Bowman GR, Pande VS, Noé F. An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. *Springer*. 2014. doi:10.1007/978-94-007-7606-7
- Husic BE, McKiernan KA, Wayment-Steele HK, Sultan MM, Pande VS. A Minimum Variance Clustering Approach Produces Robust and Interpretable Coarse-Grained Models. *J Chem Theory Comput.* 2018. doi:10.1021/acs.jctc.7b01004
- David CC, Jacobs DJ. Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol Biol*. 2014. doi:10.1007/978-1-62703-658-0
 11
- 18. Teodoro ML, Phillips GN, Kavraki LE. Understanding Protein Flexibility through

- Dimensionality Reduction. In: *Journal of Computational Biology*.; 2003. doi:10.1089/10665270360688228
- 19. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science (80-)*. 2006. doi:10.1126/science.1127647
- 20. Naritomi Y, Fuchigami S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J Chem Phys.* 2011. doi:10.1063/1.3554380
- 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. doi:10.1063/1.4811489
- 22. Mardt A, Pasquali L, Wu H, Noé F. VAMPnets for deep learning of molecular kinetics.

 Nat Commun. 2018. doi:10.1038/s41467-017-02388-1
- Fleetwood O, Kasimova MA, Westerlund AM, Delemotte L. Molecular Insights from Conformational Ensembles via Machine Learning. *Biophys J.* 2020. doi:10.1016/j.bpj.2019.12.016
- 24. Le L, Patterson A, White M. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In: *Advances in Neural Information Processing Systems*.; 2018.
- 25. Lemke T, Peter C. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J Chem Theory Comput*. 2019. doi:10.1021/acs.jctc.8b00975
- 26. Greener JG, Moffat L, Jones DT. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci Rep.* 2018. doi:10.1038/s41598-018-34533-1
- 27. Degiacomi MT. Coupling Molecular Dynamics and Deep Learning to Mine Protein

- Conformational Space. Structure. 2019. doi:10.1016/j.str.2019.03.018
- 28. Noé F, De Fabritiis G, Clementi C. Machine learning for protein folding and dynamics.

 *Curr Opin Struct Biol. 2020. doi:10.1016/j.sbi.2019.12.005
- 29. Hernández CX, Wayment-Steele HK, Sultan MM, Husic BE, Pande VS. Variational encoding of complex dynamics. *Phys Rev E*. 2018. doi:10.1103/PhysRevE.97.062412
- 30. Tsuchiya Y, Taneishi K, Yonezawa Y. Autoencoder-Based Detection of Dynamic Allostery Triggered by Ligand Binding Based on Molecular Dynamics. *J Chem Inf Model*. 2019. doi:10.1021/acs.jcim.9b00426
- 31. Wehmeyer C, Noé F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J Chem Phys.* 2018. doi:10.1063/1.5011399
- 32. Chen W, Ferguson AL. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J Comput Chem*. 2018. doi:10.1002/jcc.25520
- 33. Teletin M, Czibula G, Bocicor MI, Albert S, Pandini A. Deep autoencoders for additional insight into protein dynamics. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.; 2018. doi:10.1007/978-3-030-01421-6 8
- 34. Wang Y, Ribeiro JML, Tiwary P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat Commun.* 2019. doi:10.1038/s41467-019-11405-4
- 35. Wang Y, Lamim Ribeiro JM, Tiwary P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Curr Opin Struct Biol.* 2020. doi:10.1016/j.sbi.2019.12.016

- 36. Lusch B, Kutz JN, Brunton SL. Deep learning for universal linear embeddings of nonlinear dynamics. *Nat Commun.* 2018. doi:10.1038/s41467-018-07210-0
- 37. Jacquier H, Birgy A, Le Nagard H, et al. Capturing the mutational landscape of the betalactamase TEM-1. *Proc Natl Acad Sci U S A*. 2013. doi:10.1073/pnas.1215206110
- 38. Orencia MC, Yoon JS, Ness JE, Stemmer WPC, Stevens RC. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nat Struct Biol*. 2001. doi:10.1038/84981
- 39. Wang X, Minasov G, Shoichet BK. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol*. 2002. doi:10.1016/S0022-2836(02)00400-X
- 40. Thomas VL, McReynolds AC, Shoichet BK. Structural Bases for Stability-Function Tradeoffs in Antibiotic Resistance. *J Mol Biol.* 2010. doi:10.1016/j.jmb.2009.11.005
- 41. Woodford N, Ellington MJ. The emergence of antibiotic resistance by mutation. *Clin Microbiol Infect*. 2007. doi:10.1111/j.1469-0691.2006.01492.x
- 42. Salverda MLM, de Visser JAGM, Barlow M. Natural evolution of TEM-1 β-lactamase: Experimental reconstruction and clinical relevance. *FEMS Microbiol Rev.* 2010. doi:10.1111/j.1574-6976.2010.00222.x
- 43. Zaretzki JM, Browning MR, Hughes TB, Swamidass SJ. Extending P450 site-of-metabolism models with region-resolution data. *Bioinformatics*. 2015. doi:10.1093/bioinformatics/btv100
- 44. Moon TK. The expectation-maximization algorithm. *IEEE Signal Process Mag.* 1996. doi:10.1109/79.543975
- 45. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation

- capture the effects of mutations. Nat Methods. 2018. doi:10.1038/s41592-018-0138-4
- 46. De La Cruz EM, Ostap EM. Relating biochemistry and function in the myosin superfamily. *Curr Opin Cell Biol*. 2004. doi:10.1016/j.ceb.2003.11.011
- 47. Porter JR, Meller A, Zimmerman MI, Greenberg MJ, Bowman GR. Conformational distributions of isolated myosin motor domains encode their mechanochemical properties. *Elife*. 2020. doi:10.7554/eLife.55132
- 48. Llinas P, Isabet T, Song L, et al. How Actin Initiates the Motor Activity of Myosin. *Dev Cell*. 2015. doi:10.1016/j.devcel.2015.03.025
- 49. Duan Y, Wu C, Chowdhury S, et al. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J Comput Chem.* 2003. doi:10.1002/jcc.10349
- Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys.* 1983. doi:10.1063/1.445869
- Zimmerman MI, Bowman GR. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J Chem Theory Comput.* 2015. doi:10.1021/acs.jctc.5b00737
- 52. Shirts M, Pande VS. Screen savers of the world unite. *Science* (80-). 2000. doi:10.1126/science.290.5498.1903
- 53. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* 2018. doi:10.1093/nar/gky427
- 54. Winkelmann DA, Forgacs E, Miller MT, Stock AM. Structural basis for drug-induced allosteric changes to human β-cardiac myosin motor activity. *Nat Commun.* 2015.

doi:10.1038/ncomms8974

- 55. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. In: *Advances in Neural Information Processing Systems 32.*; 2019.
- 56. Porter JR, Zimmerman MI, Bowman GR. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *J Chem Phys.* 2019.

doi:10.1063/1.5063794

Chapter 3

Naturally-ocurring genetic variants in the oxytocin receptor alter receptor signaling profiles

3.1 Preamble

This chapter is adapted from the following article: Malik, M., Ward, M.D, Fang Y., Porter, J.R., Zimmerman, Koelblen, T., Roh, M., Frolova, A.I., Burris, T.P., Bowman, G.R., Imoukhuede, P.I., and England S.K. (2021). "Naturally-ocurring genetic variants in the oxytocin receptor alter receptor signaling profiles", *ACS Pharmacol. Transl. Sci.*, 4, 5, 1543-1555.

3.2 Introduction

A synthetic form of the hormone oxytocin is administered to a large portion of pregnant patients in the United States to induce or augment labor¹ and to nearly all patients who deliver to prevent post-partum hemorrhage.² Oxytocin response varies widely between individuals.³ For labor induction and augmentation, maximal oxytocin infusion rates range from 2 milliunits/minute (the starting rate specified in low-dose protocols) to 40 milliunits/minute (the maximal infusion rate recommended by many providers).³ The duration of oxytocin infusion required before delivery also varies by 50 hours or more, contributing to wide variations in the total oxytocin dose received by patients.⁴ Patients who receive high oxytocin doses are at increased risk for uterine hyperstimulation and rupture⁵ and postpartum hemorrhage secondary to uterine atony.⁶⁻⁸ In contrast, patients who receive insufficient oxytocin doses may require

Cesarean delivery, which puts them at risk for surgical complications.⁹ To avoid these adverse

events, clinicians have sought to identify individual factors that predict oxytocin dose requirement and thus enable personalized dosing of oxytocin.

The oxytocin receptor (OXTR) is a member of the G-protein coupled receptor (GPCR) family. To bind to oxytocin, OXTR must first traffic to the myometrial smooth muscle cell surface. Upon oxytocin binding, OXTR activates Gq, leading to Ca²⁺ release from intracellular stores, which promotes myometrial smooth muscle contraction. OXTR signaling through Gq is counteracted by coupling to β-arrestin, which mediates desensitization and internalization of OXTR from the cell surface. OXTR desensitization after oxytocin exposure may impair myometrial contractions, leading to adverse events including uterine atony and post-partum hemorrhage. Oxforced by the oxforced by coupling to adverse events including uterine atony and post-partum

Several investigators have tested the hypothesis that variants in the *OXTR* gene affect the response to exogenous oxytocin. For example, Reinl *et al.* and Grotegut *et al.* identified single nucleotide *OXTR* variants in patients who required high or low doses of oxytocin to induce labor, but these studies were not powered to detect significant associations. ^{15, 16} In an *ex vivo* study, one coding and one noncoding *OXTR* variant altered the oxytocin-induced contractions of uterine tissue strips isolated from pregnant individuals. ¹⁷ Although exome sequencing studies have shown that missense variants in the *OXTR* gene are prevalent in the global human population, ¹⁸ the functional effects of most of these variants have not been determined. However, prevalent missense variants in other GPCRs genes lead to aberrant drug responses. ¹⁹ Here, we assessed the effects of genetic variants of unknown significance in *OXTR* on oxytocin response in cells.

3.3 Methods

3.3.1 Cell culture

HEK293T cells were maintained in Dulbecco's Modified Eagle Medium/Ham's F12 medium without phenol red and supplemented with 10% fetal bovine serum and 25 μg/mL gentamicin. Cells were kept in a humidified cell culture incubator at 37 °C with 5% CO₂.

3.3.2 cDNA constructs

The wild-type (WT) OXTR and P108A OXTR constructs in pcDNA3.1(+) vector were a kind gift from Dr. Jeffrey Murray (University of Iowa). Other missense single nucleotide variants were introduced by site-directed mutagenesis (Genewiz, South Plainfield, NJ). The WT OXTR sequence was identical to the coding region of the National Center for Biotechnology Information reference sequence NM_000916.3.

The β-arrestin-1-Rluc8 fusion construct in the vector pcDNA3.1(+) encoded β-arrestin-1 with a C-terminal linker SGGSTSA followed by Rluc8. The β-arrestin-2-Rluc8 fusion construct in the vector pcDNA3.1(+) encoded β-arrestin-2 with a C-terminal linker GGGSEF followed by Rluc8. The template cDNA clones for β-arrestin-1 (ARRB100002) and β-arrestin-2 (ARRB200001) were obtained from the cDNA Resource Center (Bloomsberg, PA, www.cdna.org). A plasmid containing the Rluc8 cDNA was a kind gift from Dr. Brian Finck (Washington University in St. Louis).

The OXTR-GFP10 fusion construct in the vector pcDNA 3.1(+) encoded OXTR with a C-terminal linker SGGKL followed by GFP10. A plasmid containing the GFP10 cDNA was a kind gift Dr. Céline Gales (INSERM, France).

The plasmid encoding OXTR-GFP was a gift from Christian Gruber (Addgene plasmid #67848; http://n2t.net/addgene:67848; RRID:Addgene_67848).²⁰ Note that this plasmid includes the missense single nucleotide variant A218T, which was corrected before introducing the

variants of interest. An N-terminal HA tag was added (linker GPT) to generate the HA-OXTR-GFP construct.

All plasmids were confirmed by bidirectional Sanger sequencing.

Oxytocin (Tocris Bioscience, Minneapolis, MN) stock solutions diluted to 500 μ M in water were stored at -80 °C until just before use.

$3.3.3 \text{ Ca}^{2+}$ assays

HEK293T cells (2×10^4) were plated in each well of 96-well black-walled, clear-bottom polystyrene microplates coated with poly-D-lysine. The following day, cells were transfected with a construct encoding WT or variant OXTR. Each variant was tested alongside WT controls on the same plate. For transfections, 50 ng of DNA and 0.5 μ L of TransIT-293 reagent (Mirus Bio, Madison, WI) diluted in Opti-MEM reduced-serum media (Thermo Fisher Scientific, Waltham, MA) were added to each well. After 24 hours, media was removed and replaced with 100 μ L Brilliant Calcium indicator solution (Ion Biosciences, San Marcos, TX), which was prepared by diluting Brilliant Calcium indicator, DrySolv, and TRS reagent in assay buffer. After incubation for one hour, a Synergy2 plate reader (BioTek, Winooski, VT) was used to add 100 μ L of oxytocin of the appropriate concentration and record fluorescence intensity (excitation filter = 485/20 nm, emission filter = 528/20 nm) every 0.14 s for 20 s/well. Fluorescence increase (increase in intracellular Ca²⁺) was calculated as the average of fluorescence intensity readings from 10 s to 20 s after oxytocin addition minus the minimum fluorescence intensity averaged over 5 points from 0 s to 10 s.

For desensitization assays, transfected cells were pre-treated with the indicated oxytocin concentrations for 30 minutes. Then, without washing out the pre-treatment oxytocin, a Synergy 2 plate reader was used to add a challenge dose of 1 μ M oxytocin and record response as above.

3.3.4 Bioluminescence resonance energy transfer (BRET) assays

HEK293T cells (4 x 10⁴) were plated in each well of 96-well white-walled, clear-bottom polystyrene microplates coated with poly-D-lysine. The following day, cells were transfected with WT or variant OXTR-GFP10 and β-arrestin-1-Rluc8 or β-arrestin-2-Rluc8 at a ratio of 15:1 (w/w). For transfections, 50 ng of DNA and 0.5 μL of Lipofectamine 2000 reagent (Thermo Fisher Scientific), both diluted in Opti-MEM reduced-serum media, were added to each well. After 24 hours, media was removed and replaced with 100 μL of Hank's Buffered Salt Solution (HBSS) supplemented with 20 mM HEPES. A Synergy2 plate reader was used to add 100 μL of assay buffer containing 10 μM coelenterazine 400a (Biotium, Fremont, CA) and the indicated concentrations of oxytocin to 10 wells at a time. Luminescence at 520 nm and 400 nm was read every 26 s for a total of 182 s. BRET ratio was calculated as the average ratios of emission at 520 nm/400 nm at the 5 time points from 78 to 182 s. WT controls were tested on each plate in parallel with variants.

3.3.5 Quantitative flow cytometry

HEK293T cells (1 x 10⁶) were plated in T25 flasks and transfected the next day with HA-OXTR-GFP, OXTR-GFP, or HA-OXTR. Cells were transfected with 300 ng of plasmid DNA and 4 uL of TransIT-LT1 reagent (Mirus Bio). Cells were detached 24 hours later with CellStripper (Corning) and collected by centrifugation. To measure receptor internalization, cells were incubated with the indicated concentration of oxytocin for 30 minutes before and during

65

detachment. Cells were incubated with an empirically-determined saturating concentration (8-16 μg/mL) of phycoerythrin (PE)-conjugated anti-HA antibody (901518, Biolegend, San Diego, CA) in staining buffer (0.5% BSA and 0.1% sodium azide in Ca²⁺/Mg²⁺-free PBS) on ice for 40 minutes, then washed twice in staining buffer before flow cytometry to quantify cell surface OXTR. For quantification of total OXTR, the PE-labelled living cells were fixed with 2% paraformaldehyde and permeabilized with 0.5% Tween20 in PBS. Cells were washed with 0.1% Tween 20 in PBS, incubated with 16 μg/mL PE anti-HA antibody for 40 minutes at room temperature, and washed twice before flow cytometry.

Flow cytometry was performed on a CytoFLEX flow cytometer (Beckman Coulter, Indianapolis, IN). Three technical replicates were performed for each experimental condition, and data from 5000 transfected cells were collected from each replicate. Three independent trials were performed. SYTOX Blue (Thermo Fisher Scientific) was used to exclude dead cells where appropriate. PE Quantibrite beads (BD Biosciences) were used for calibration. Flow cytometry gating was performed as follows: 1) forward and side scatter were used to exclude debris, 2) forward scatter-width vs. -height was used to exclude doublets, 3) SYTOX blue staining was used to identify dead cells, 4) GFP fluorescence was used to gate transfected cells (GFP+ population). The GFP+ threshold was determined relative to the GFP signal in GFP-negative control (cells transfected with HA-OXTR).

The number of receptors on transfected cells was calculated from the geometric mean of PE fluorescence intensity calibrated to PE standards as previously described.²¹ Values from nonspecific binding of PE-HA antibody to HA-negative cells (cells transfected with OXTR-GFP) were subtracted from all samples.

3.3.6 Data processing for Ca²⁺, BRET, desensitization, and internalization assays

For Ca²⁺ and BRET assays, responses were normalized by subtracting the average basal response from all samples, then dividing by the average WT response at the highest oxytocin concentration for each trial. For desensitization and internalization experiments, responses were normalized by dividing values from all samples by the average response from the corresponding non-pretreated sample(s). Normalization was performed separately for each replicate experiment.

Non-linear regression with least-squares fitting was used to generate dose-response curves and calculate E_{max} , EC50, and IC50 values (GraphPad Prism 8). The three-parameter regression method, which was used to fit the BRET data and internalization data, used the model: $Y = Bottom + (Top-Bottom)/(1+10^(LogEC50 \text{ or IC50-X}))$. The four-parameter regression method, which was used to fit the Ca^{2+} activation and desensitization data, used the equation $Y=Bottom + (Top-Bottom)/(1+10^((LogEC50 \text{ or IC50-X})*HillSlope))$. In these models, Y=response, X=log(oxytocin concentration), and no constraints were placed on any values. Buffer controls were assigned a nominal concentration value of 10^{-9} M for BRET assays or 10^{-12} M for all other assays.

All experiments were performed in triplicate, with WT controls tested alongside each variant on the same plate to control for day-to-day variation in assay response. Average values from three biological replicates were used to construct dose-response curves for each variant and the matched WT controls, which were compared by performing nested extra sum-of-squares F tests. F statistics were calculated and *P*-values were determined as previously described.^{22, 23} *P*-values shown reflect comparisons of logEC50 values or Top values (see equations above), as indicated.

3.3.7 Molecular Dynamics Simulations

The initial homology model of WT OXTR was provided by the I-TASSER GPCR homology model database. This model was then prepared for simulation by the CHARMM-GUI membrane protein input generator. Mutations (e.g., V281M) and palmitate lipid tails on C346 and C347 were introduced by the CHARMM-GUI PDB manipulator. All proteins were simulated in 0.15 M KCl (111 K⁺ ions and 92 Cl⁻ ions) in a rectangular box of size 99.5 x 99.5 x 171.2 Å with a membrane consisting of 121 (upper leaflet) or 120 (lower leaflet) POPC molecules and 12 cholesterol molecules (upper and lower leaflet). All systems contained ~100,000 TIP3P³⁰ water molecules. Systems were minimized in the default manner supplied by CHARMM-GUI. Briefly, using the CHARMM36m force field, ach system's energy was minimized by using gradient descent, then simulated NVT with progressively weaker and fewer restraints on positions of atoms and membrane components.

Production runs were performed in GROMACS.³² Hydrogen bonds were constrained with the LINCS algorithm.³³ Cutoffs of 1.2 nm were used for the neighbor list, Coulomb interactions, and van der Waals interactions. The force-switch modifier was used to smoothly switch forces from van der Waals interactions to zero between 1.0 and 1.2 nm. The Verlet cutoff scheme was used for the neighbor list. The Nose-Hoover thermostat was used to hold the temperature at 300 K.³⁴ The semi-isotropic Parrinello-Rahman barostat was used to maintain constant pressure of 1 bar as is standard in protein-membrane simulations.³⁵ Conformations were stored every 20 ps.

The FAST algorithm^{36, 37} was used to enhance conformational sampling for each OXTR sequence (WT, P108A, V281M, and V45L). Five FAST simulation rounds were conducted with 10 simulations per round. Each simulation was 50 ns in length (2.5 µs aggregate simulation). To

explore away from the starting structure, the FAST ranking function favored restarting simulations from states that had the fewest number of preserved native contacts. Additionally, a similarity penalty was added to the ranking to promote conformational diversity in starting structures, as described previously.³⁸

3.3.8 DiffNet Analysis

DiffNets can perform dimensionality reduction in a way that highlights biochemically relevant differences between datasets.³⁹ Two DiffNets were independently trained to learn about impairment of β-arrestin and Gq signaling. All DiffNet training and analysis was conducted under the assumption that the regions of Gq and β -arrestin binding were most likely to contain differences that explained impaired Gq or β-arrestin signaling. Therefore, the DiffNet analysis only considered atoms in the binding region (as shown in **Figure A.2.1**). All simulation data (2.5 μs per variant) was converted to DiffNet input as described previously.³⁹ Briefly, XYZ atom coordinates from simulations were mean-shifted to zero and then multiplied by the inverse of the square root of a covariance matrix, which was calculated from simulations. To learn about βarrestin impairment, a DiffNet was trained to classify all structures from V45L and P108A as βarrestin impaired (i.e., initial labels of one) and WT and V281M simulations as normal (i.e., initial labels of zero). To learn about Gq impairment, a DiffNet was trained to classify structures from V281M simulations as potentially Gq impaired and WT, V45L, and P108A simulations as normal. In both cases, the labels were iteratively updated in a self-supervised manner described previously³⁹ in which expectation maximization bounds of [0.1-0.4] were chosen for normal variants and [0.6-0.9] for impaired variants. Both training sessions used 10 latent variables, 10 training epochs in which the data were subsampled by a factor of 10 in each epoch, a batch size of 32, and a learning rate of 0.0001.

3.3.9 Markov State Model construction and analysis

A Markov State Model (MSM) is a statistical framework for analyzing molecular dynamics simulations and provides a network representation of a free energy landscape. 40-42 To quantify differences between variants, several measurements were made that relied on MSMs, each built with 2.5 µs of simulation data for each variant. All MSMs were constructed with Enspara, 43 a python library for clustering and building MSMs from molecular simulation data. In this work, Enspara was used to cluster OXTR structures, count transitions between clusters, and derive equilibrium probabilities of structural states explored during simulation. A separate MSM was built for each variant, using the same methodology for each variant. Namely, simulation frames were converted from XYZ atom coordinates to a vector containing a value indicating the amount of solvent-accessible surface area (SASA) of each residue sidechain (i.e., the data was SASA featurized). SASA calculations were computed by using the Shrake-Rupley algorithm⁴⁴ (with a solvent probe radius of 0.28 nm) as implemented in the python package MDTraj⁴⁵. SASA featurization was used for subsequent clustering because, unlike other clustering schemes (e.g., RMSD-based), SASA emphasizes the conformational changes of surface residues over internal residues, which should be most useful for understanding signaling of a transmembrane receptor that has a surface for binding ligands. Next, the SASA-featurized data were clustered with a hybrid clustering algorithm. First, a k-centers algorithm⁴⁶ was used to cluster the data into 1000 clusters. Next, three sweeps of k-medoids update steps were applied to refine the cluster centers to be in the densest regions of conformational space. Then, transition probability matrices were produced by counting transitions between states (i.e., clusters) using a 2 ns lag time, adding a prior count of $\frac{1}{N_{states}}$ and row-normalizing, as described previously.⁴⁷ Equilibrium populations were calculated as the eigenvector of the transition probability matrix with an eigenvalue of one.

For the distance histograms in **Figures 3.6** and **3.7**, the distance for each cluster center (i.e., representative structure of the cluster) was calculated and the distance was weighted by the corresponding equilibrium population calculated with the MSM. Similar calculations performed with an MSM built on an RMSD-based clustering scheme produced similar results (**Figure A.2.2**).

3.4 Results

3.4.1 Genetic variation occurs in several locations within OXTR

We searched the worldwide gnomAD v2.1 dataset, ¹⁸ which includes 141,456 exomes, to identify the most prevalent single nucleotide missense variants in *OXTR*. We identified 11 *OXTR* variants (**Table 3.1**) with allele counts greater than 50, indicating that they were detected in more than 50 heterozygous individuals. ¹⁸ These variants affected residues in multiple domains, including six residues in transmembrane domains (TMs), one in the first extracellular loop (ECL1), two in the third intracellular loop (ICL3), and two in the C-terminal tail (**Table 3.1**, **Figure 3.1A**). The gnomAD cohort includes homozygotes for the four most common variants: A218T, A238T, V172A, and L206V. The most prevalent variant, A218T, was found in 27% of gnomAD participants; the 11th most prevalent variant, P108A, was found in 0.05% of participants.

3.4.2 OXTR missense variants alter Ca^{2+} signaling and β -arrestin recruitment

We reasoned that the missense variants most likely to affect clinical oxytocin response would alter oxytocin-induced Ca^{2+} signaling, which is required for myometrial smooth muscle contraction, or recruitment of β -arrestin, which is thought to mediate OXTR desensitization. Therefore, to prioritize variants for further study, we transiently transfected plasmids encoding

wild-type (WT) OXTR or the 11 variants into HEK293T cells and then performed highthroughput assays to measure effects on these pathways. First, to measure increases in intracellular Ca²⁺ in response to oxytocin, we used a fluorescent Ca²⁺ indicator dye. Second, to measure β-arrestin recruitment in response to oxytocin, we performed bioluminescence resonance energy transfer assays in HEK293T cells transfected with green fluorescent protein (GFP)-tagged OXTR and luciferase-tagged β-arrestin-1 or β-arrestin-2. V45L, P108A, L206V, V281M, and E339K had the largest statistically significant effects on EC50 or E_{max} in two or more assays and were therefore selected for further study (Figure 3.1, Tables A.2.1, A.2.2, and **A.2.3**). V45L decreased the E_{max} for β -arrestin-1 recruitment and increased the EC50 for β arrestin-2 recruitment (Figure A.2.3). P108A increased the EC50 for β-arrestin-1 recruitment and increased both the EC50 and the E_{max} for β -arrestin-2 recruitment. L206V increased the E_{max} for β-arrestin-1 and β-arrestin-2 recruitment. V281M increased the EC50 for Ca²⁺ signaling and decreased the E_{max} for Ca²⁺ signaling and β-arrestin-2 recruitment. Finally, E339K increased the EC50 for Ca^{2+} signaling and decreased the E_{max} for Ca^{2+} signaling, β -arrestin-1 recruitment, and β-arrestin-2 recruitment (**Figure 3.1**).

3.4.3 OXTR variants alter cell surface localization

To quantify the effect of these five genetic variants on OXTR quantity and localization to the plasma membrane, we performed quantitative flow cytometry. A specific OXTR antibody is not commercially available, so we created a plasmid encoding the OXTR fusion protein HA-OXTR-GFP. We used GFP fluorescence to differentiate transfected from untransfected cells, and a phycoerythrin (PE) -conjugated anti-HA antibody to quantify the HA epitope on the extracellular N-terminus of OXTR. To quantify surface OXTRs, living cells were labelled by

PE; to quantify total OXTRs throughout the cell, an additional PE-labelling step was performed after fixing and permeabilizing the PE-labelled living cells.

No variants had a statistically significant effect on the total number of OXTRs per cell after adjusting for multiple comparisons (P>0.01 in one-sample t-tests, **Figure 3.2A**). However, two variants (P108A and L206V) increased the number of cell surface OXTRs by 23 \pm 3% and 41 \pm 4%, respectively (P=0.0003 and P=0.0002, one sample t-tests). Conversely, two variants (V281M and E339K) decreased the number of cell surface OXTRs by 49 \pm 0.7% and 36 \pm 2%, respectively (P<0.0001, one-sample t-tests, **Figure 3.2B**).

When we graphed cell surface OXTRs as a percentage of total OXTRs (**Figure 3.2C**), we found that $21 \pm 2\%$ of total WT OXTRs were localized to the plasma membrane. P108A and L206V increased OXTR surface localization to $25 \pm 1\%$ and $27 \pm 1\%$, respectively (adjusted P=0.03 for both). Conversely, V281M and E339K decreased OXTR surface localization to $12 \pm 1\%$ and $17 \pm 1\%$, respectively (adjusted P=0.01 for both).

3.4.4 V45L, P108A, and E339K impair OXTR desensitization and internalization

OXTR internalization and desensitization, mediated in part by β-arrestin recruitment, are thought to be responsible for some adverse effects associated with oxytocin exposure, including uterine atony and post-partum hemorrhage. ¹³ Thus, to assess the potential clinical implication of variants, we aimed to define their effects on OXTR desensitization and internalization. As expected, for all five variants, relative differences in the number of cell surface receptors (**Figure 3.2**) corresponded to the differences seen in maximal β-arrestin recruitment assays (**Figure 3.1E, 1G**). For example, P108A and L206V had elevated E_{max} values for β-arrestin-2 recruitment and elevated membrane localization, whereas V281M and E339K had decreased E_{max} values for β-

arrestin recruitment decreased membrane localization. In contrast, differences in the EC50 of β -arrestin recruitment did not correspond to changes in cell surface receptor number. For example, V45L increased the EC50 of β -arrestin-2 recruitment but had no effect on membrane localization, and P108A increased the EC50 of both β -arrestin-1 and β -arrestin-2 recruitment and increased membrane localization. We hypothesized that increased EC50 values would reflect functional deficits in OXTR desensitization and internalization.

To measure desensitization, we pretreated cells expressing WT OXTR or the five variants with varying concentrations of oxytocin for 30 minutes, then used Ca^{2+} indicator assays to measure the cellular response to a saturating concentration (1 μ M) of oxytocin (**Figure 3.3**). To measure internalization, we incubated cells with varying concentrations of oxytocin for 30 minutes, then performed quantitative flow cytometry to measure surface OXTRs (**Figure 3.3**). We found that V281M and L206V had no effect on either receptor desensitization or internalization (P>0.05, extra sum-of-squares F test). In contrast, V45L, P108A, and E339K caused a rightward shift in the dose-response curve and increased the IC50 for desensitization (P=0.0001, P<0.0001, and P<0.0001, sum-of-squares F test, **Figure 3.4B**, **Table A.2.4**). V45L and P108A caused a similar rightward shift in internalization assays (P=0.0098 and P=0.0003, extra sum-of-squares F test, **Figure 3.4C**, **Table A.2.4**). Although E339K did not cause a statistically significant increase in EC50 for internalization (P>0.05), it prevented maximal internalization, with 44% of E339K OXTRs versus 24% of WT OXTRs remaining on the cell surface (P=0.0001, **Figure 3.4C**).

Three of the five variants investigated had differential effects on OXTR activation (oxytocin-induced Ca²⁺ signaling in **Figure 3.4A**), desensitization (**Figure 3.4B**), and internalization (**Figure 3.4C**). These variants altered the balance between OXTR desensitization

and activation at any given dose of oxytocin (**Figure 3.4D**, **Figure A.2.4**). Of the three variants that impaired OXTR internalization and desensitization, only one, E339K, also altered potency and efficacy for OXTR activation, potentially due to decreased cell surface localization (**Figure 3.2B**). V281M had similar effects as E339K on OXTR cell surface localization and OXTR activation but had no effect on OXTR internalization or desensitization (**Figure 3.4**). In contrast, V45L and P108A impaired OXTR internalization and desensitization without altering OXTR activation (**Figure 3.4**).

3.4.5 Variants that reduce desensitization and internalization alter OXTR structural conformations

In our *in vitro* assays, two variants (V45L and P108A) reduced β-arrestin recruitment, OXTR internalization, and OXTR desensitization compared to WT OXTR. Thus, three lines of evidence suggest that V45L and P108A decrease OXTR's ability to activate β-arrestin. To define the structural basis of β-arrestin impairment, we used molecular dynamics simulations to computationally model the motions of all atoms in WT and variant OXTRs in solution over time (**Figure 3.5A, 3.5B**). We paired these simulations with the FAST algorithm (see **Methods**^{36, 37}) to enhance sampling of the conformational ensemble (i.e., the set of structural poses the receptor adopts) of each variant.

To identify the conformational changes most associated with β-arrestin impairment, we used DiffNets, deep-learning algorithms that are trained to identify biochemically relevant differences between multiple conformational ensembles (see **Methods**).³⁹ We first trained a DiffNet to identify differences between conformational ensembles of the two β-arrestin-impaired OXTRs (V45L and P108A) and two OXTRs (WT and V281M) with normal desensitization and internalization. From this training, the DiffNet learned a label for each simulation frame

(structural configuration) from zero to one that indicated the probability that it was associated with this classification. To interpret these labels, we calculated the correlation between interatom distances in the OXTR cytosolic region (71,289 possible distances, **Figure A.2.1**) and changes in the DiffNet label. We then plotted the 100 distances that were most correlated with the DiffNet label (**Figure 3.5C**). This analysis showed clear enrichment in distances that cluster at the interface between transmembrane domain 1 (TM1) and the first intracellular loop (ICL1), indicating that changes in this region were associated with β -arrestin impairment.

3.4.6 Conformational changes in V45L and P108A OXTRs disrupt putative β -arrestin binding site

DiffNets identified locations associated with reduced β-arrestin function without any prior information about functional sites in OXTR. To determine whether the DifffNet predictions corresponded to functional locations, we used the simulation data to build Markov State Models. Markov State Models provide a discrete map of structural configurations and an equilibrium population value that corresponds to the proportion of time a protein spends in a given configuration. The DiffNet prediction implicated the TM1-ICL1 region in β-arrestin impairment, so we used Markov State Models to more closely examine this region. In this analysis, V45L and P108A introduced an additional helical turn at the C-terminus of TM1 that was not present in WT and V281M OXTR. Specifically, we found that the hydrogen bond between Val⁶⁰ and Leu⁶⁴ was shorter in V45L and P108A OXTR than in WT and V281M OXTR (0.2 nm vs. 0.6 nm) (**Figure 3.6A**). Thus, β-arrestin-impaired OXTRs were predicted to have a shorter ICL1 than OXTRs with normal β-arrestin function.

This conformational change has important implications for β -arrestin binding. First, shortening ICL1 may prevent the interactions between ICL1 and the bottom loop of β -arrestin

(**Figure 3.6B**) previously described by Yin *et al.*⁴⁸ Second, shortening ICL1 reduces the distance between ICL1 and helix 8 (H8), causing a collapsed state (**Figure 3.6C**). When we superimposed bound structures of β -arrestin and G protein (from other GPCRs^{48, 49}) onto the OXTR homology model, the model predicted that this shortened distance created a steric clash between ICL1 and the β -arrestin finger loop, but not between ICL1 and the G protein (**Figure 3.6D**). Taken together, our data suggest that the mechanism underlying reduced β -arrestin function was similar in V45L and P108A OXTR.

3.4.7 Structural conformations in V281M OXTR

Our results in **Figure 3.4D** indicated that the balance between OXTR activation and desensitization in V281M OXTR deviated significantly from WT, with greater relative desensitization for any given unit of activation. We observed the opposite deviation in V45L and P108A OXTR, both of which had less relative desensitization for any given unit of activation. To investigate the structural basis of this difference, we used a similar approach as above and trained a second DiffNet to identify differences between conformational ensembles of V281M OXTR and V45L, P108A, and WT OXTR. We plotted the 100 distances that were most correlated with the DiffNet label in **Figure 3.5D**. This analysis showed enrichment for distances between transmembrane domains 3 and 5 (TM3 and TM5), indicating that structural rearrangements in this region were associated with V281M.

We then used Markov State Models to plot the probability that OXTR adopts a conformation with a given distance between TM3 and TM5. V281M OXTR was more likely to adopt conformations with a shorter distance between TM3 and TM5 than were WT, V45L, and P108A OXTR (0.8 nm versus 1.2-1.4 nm, **Figure 3.7A**). When we superimposed the bound β-arrestin and G protein structures, we saw that this collapsed state caused a steric clash with the G

protein but not with β -arrestin (**Figure 3.7B**). This finding suggests that V281M disrupted the binding of Gq to OXTR without affecting β -arrestin recruitment.

3.5 Conclusions

Our data indicate that *OXTR* variants found in the global human population significantly altered OXTR function. Specifically, these variants altered oxytocin response by changing OXTR localization to the cell membrane, decreasing oxytocin-induced Ca²⁺ signaling, altering β-arrestin recruitment and signaling, or a combination of these effects. The variants P108A and L206V increased the percentage of OXTR on the cell membrane, whereas V281M and E339K caused OXTR to be retained inside the cell. V281M and E339K also decreased Ca²⁺ signaling. Three variants (V45L, P108A, and E339K) impaired OXTR desensitization and OXTR internalization upon exposure to oxytocin. Our molecular dynamics simulations predict that both V45L and P108A introduce an extra helical turn at the end of TM1, which may explain the impaired coupling to β-arrestin seen *in vitro*.

Our results from the V281M and E339K variants highlight the importance of efficient membrane trafficking for receptor function. These intracellularly retained variants were the only two variants studied that decreased oxytocin-induced Ca²⁺ signaling (**Figure 3.1B, 3.1C**). In contrast, P108A and L206V, which increased the number of OXTR on the cell surface, did not increase maximal Ca²⁺ signaling. This may be because Ca²⁺ signaling becomes saturated at a certain concentration of receptors per cell. Because Gq signaling amplifies through the signaling pathway that leads to Ca²⁺ mobilization, intracellular Ca²⁺ is not a one-to-one readout of Gq activation. A more direct measurement of Gq activation may show that maximal Gq activation

correlates with surface OXTRs, but this may not translate directly to the activation of downstream pathways important for myometrial contractions.

Unlike maximal Ca^{2+} signaling, maximal recruitment of β -arrestin measured in the bioluminescence resonance energy transfer screen closely matched the number of OXTRs on the cell membrane. P108A and L206V, which increased cell surface OXTR, caused higher maximal recruitment (E_{max}), whereas V281M and E339K, which decreased cell surface OXTR, caused lower maximal recruitment (**Figure 3.1E, 3.1G**). Changes in E_{max} in our bioluminescence resonance energy transfer assays seemed to reflect a change in the number of receptors available to recruit β -arrestin, but did not always correspond to functional changes in receptor desensitization or internalization (**Figure 3.4**). For example, the L206V and V281M variants had the largest effects on E_{max} for β -arrestin recruitment but did not alter receptor desensitization or internalization. In contrast, increases in the EC50 for β -arrestin recruitment corresponded to right shifts in desensitization and internalization curves. Whereas OXTR desensitization and internalization can occur by several mechanisms, our results suggest that changes in β -arrestin recruitment EC50 translate to functional differences in desensitization and internalization.

To complement our *in vitro* assays, we used an *in silico* method to model the behavior of variant OXTRs. Our *in vitro* assays showed that V45L and P108A caused rightward shifts in the dose-response curves for β-arrestin recruitment, OXTR desensitization, and OXTR internalization but not oxytocin-induced Ca²⁺ signaling. We used the deep-learning approach DiffNets to identify structural changes that were common to V45L and P108A OXTRs but not present in OXTRs with normal internalization and desensitization. Importantly, the DiffNet required no input of information about OXTR/GPCR structure/function relationships to identify locations in OXTR that appear to be associated with β-arrestin binding. This discovery-based

approach yielded predictions that correspond with our *in vitro* data as well as published work on the mechanism of β-arrestin binding in other GPCRs. ⁴⁸ The structural differences shown in **Figure 3.6** suggest one mechanism by which OXTR can bind to and activate G proteins without activating β-arrestin. However, further work is necessary to validate these predictions and determine the mechanism of β-arrestin binding to OXTR. In the future, these findings may guide the design of biased agonists, as recently demonstrated by Suomuoviri *et al.* for the angiotensin II type 1 receptor. ⁵¹ Novel uterotonics that mimic the effects of V45L and P108A may preferentially activate OXTR signaling through Gq with less β-arrestin activation, thus decreasing the risk of adverse effects associated with OXTR internalization and desensitization.

We used a similar approach to identify conformational changes associated with V281M, a variant that decreased OXTR activation (oxytocin-induced Ca²⁺ signaling) but had no effect on desensitization or internalization. Our Markov State Models predicted conformational changes in V281M OXTR consistent with steric hindrance of G protein binding (**Figure 3.7**). Importantly, these changes would not hinder binding of β-arrestin and thus present a possible mechanism by which V281M altered Ca²⁺ signaling without altering desensitization or internalization. However, the changes caused by V281M were also likely due, at least in part, to inefficient cell membrane localization of V281M OXTR (**Figure 3.2**). Therefore, further *in vitro* studies are necessary to determine whether V281M OXTR displays decreased binding to Gq and thus validate the predictions from our molecular dynamics simulations.

Our findings add to two previous *in vitro* studies examining human *OXTR* variants. First, Ma *et al.* showed that R376G, a variant associated with autism spectrum disorder, increased the rate of OXTR internalization and recycling to the cell surface after treatment with oxytocin.⁵² It is unclear whether the small changes in β -arrestin recruitment seen in our screening assays

(Tables A.2.2 and A.2.3) explain the differences in OXTR internalization and recycling observed by Ma *et al.* Second, Kim *et al.* characterized three missense *OXTR* variants, including P108A, that they identified in patients who experienced premature labor. ⁵³ These authors reported that P108A decreased oxytocin binding but did not significantly affect Gq activation as measured by inositol phosphate production, which was consistent with our results. Furthermore, our findings show that P108A impaired OXTR desensitization, meaning that some OXTR Gq activation occurred unopposed. This could result in premature initiation of uterine contractions and thus explain an association between P108A and premature labor. Future studies are needed to determine whether P108A – and V45L, which we found to have similarly impaired desensitization and structural changes – predispose patients to preterm labor.

Understanding how genetic variants alter receptor function is an important step towards personalized drug dosing. Our functional annotation of the 11 most prevalent variants of unknown significance in *OXTR* helped us to prioritize the variants most likely to affect OXTR function for further study. These variants caused EC50 changes in the two- to four-fold range, consistent with effects caused by other naturally-occurring GPCR variants linked to disease risk and drug response. 54-57 Additionally, our data indicate that the two most prevalent missense variants, A218T and A238T, are unlikely to appreciably affect OXTR function.

Both activation and desensitization of the Ca²⁺ signaling pathway play an important role in determining clinical response to oxytocin. Currently, most oxytocin dosing protocols for labor induction call for providers to increase the oxytocin infusion rate at steady intervals, which compensates for a given amount of OXTR desensitization over time.⁵⁸ Imbalance between these processes, also known as signaling bias, may therefore have clinical consequences, as shown in other GPCRs.^{19, 57, 59} In our study, we identified three variants that may cause signaling bias: 1)

V45L and P108A impaired OXTR desensitization but not activation and 2) V281M decreased OXTR activation but not desensitization. However, further studies are necessary to determine whether these changes represent signaling bias between β-arrestin and Gq. Our data indicate that individuals who carry the V281M allele may be less responsive to oxytocin but still susceptible to the potential adverse effects that result from OXTR desensitization during labor (i.e., post-partum hemorrhage, uterine atony). These individuals may require higher doses of oxytocin to achieve labor induction and thus may have increased risk of these adverse events. Furthermore, oxytocin may be less effective in preventing postpartum hemorrhage in these individuals. In contrast, patients with V45L or P108A variants may be less susceptible to the adverse effects that result from OXTR desensitization but more susceptible to uterine hyperstimulation as a result of induction with oxytocin. Finally, patients with the E339K variant, which impairs OXTR activation and desensitization to roughly the same extent, may require higher oxytocin doses to achieve clinical effects.

Our studies indicate that individuals who carry the V45L, P108A, V281M, or E339K variants may benefit from personalized oxytocin dosing protocols or alternative methods of labor induction. P108A is found in 0.3% of the Finnish population, V281M is found in 0.7% of the Swedish population, and E339K is found in 1.5% of the Ashkenazi Jewish population. ^{18, 60} Further studies in these populations are necessary to determine the utility of genetic analyses in developing precision medicine approaches to oxytocin dosing.

L206V ^{5.44} E339K	TM5 C-terminus	551 308	0.39 0.22
$V172A^{4.61}$	TM4	1613	1.14
A238T	ICL3	5067	3.87
$A218T^{5.56}$	TM5	41562	27.09
Variant	Location	Allele count in gnomAD	Affected (%)

G221S ^{5.59}	ICL3	215	0.15
G252A	ICL3	178	0.14
$V281M^{6.41}$	TM6	107	0.08
$V45L^{1.38}$	TM1	91	0.09
R376G	C-terminus	89	0.06
P108A	ECL1	74	0.05

Table 3.1 OXTR variants for study.Affected (%): Percent of gnomAD participants with sequencing coverage at that locus who were homozygous or heterozygous for that variant. ECL: extracellular loop. ICL: intracellular loop. TM: transmembrane domain. Ballesteros-Weinstein numbering⁶¹ is shown for TM residues (superscripts).

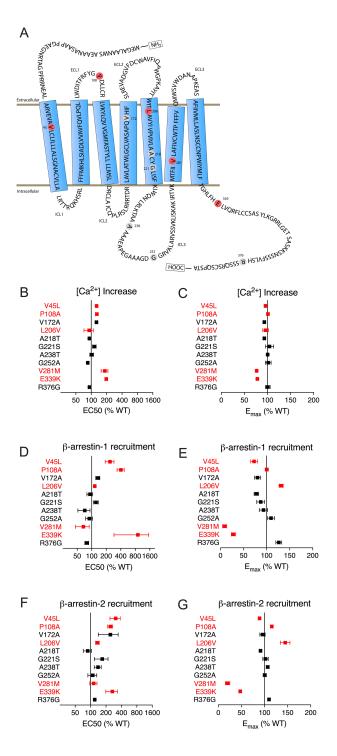


Figure 3.1 Screen identifies OXTR variants that alter oxytocin response in Ca^{2^+} assays and β -arrestin recruitment assays.

(A) Variant residues within OXTR. ICL: intracellular loop. ECL: extracellular loop. (A–G) Plots show EC50 (B, D, F) and E_{max} (C, E, G) for dose-response curves for each variant, relative to WT value (100%). Variants shown in red were chosen for further study on the basis of large effect size and statistical significance (see **Tables A.2.1**, A.2.2, and A.2.3). Error bars show standard error of the mean from N=3 independent experiments with 3-5 technical replicates per experiment.

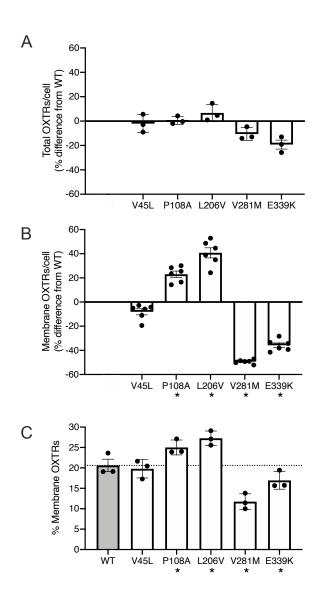


Figure 3.2 Genetic variants alter quantity of OXTR on the cell membrane.

(A) Total number of OXTRs, (B) the number of OXTRs on the cell surface, and (C) the percentage of OXTRs on the cell surface in HEK293T cells transfected with plasmids encoding wild type (WT) and variant HA-OXTR-GFP. For (A) and (B), values for variants are shown as % difference from the WT OXTR value. Error bars show standard error from N=3-6 independent experiments with 15000 cells across 3 technical replicates per experiment. * indicates variant value differs from 0 with P<0.01 in one-sample t-test (B), or differs from WT with P<0.05 in one-way repeated measures ANOVA with post-hoc Dunnet multiple comparisons test (omnibus P=0.0024 (C).

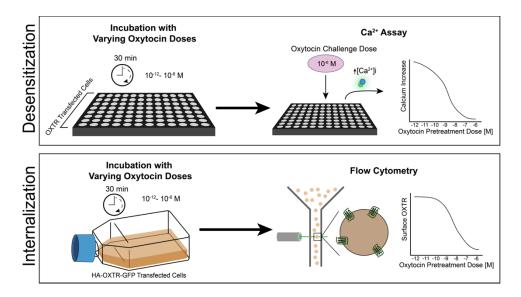


Figure 3.3 Method and data processing for desensitization and internalization assays.

For desensitization assays, cells were incubated with indicated oxytocin doses for 30 minutes, then challenged with 1 μ M oxytocin. Ca²⁺ increase in response to 1 μ M challenge is shown. For internalization assays, cells were incubated with indicated oxytocin doses for 30 minutes, then analyzed by quantitative flow cytometry to measure surface OXTR.

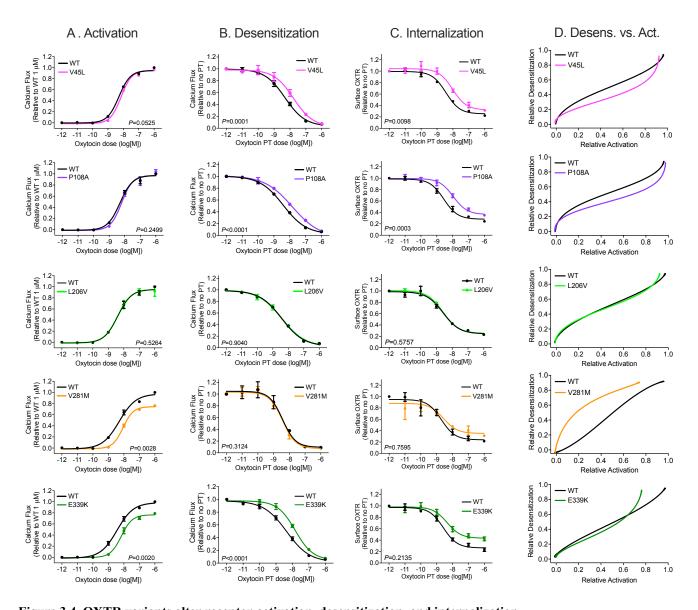


Figure 3.4 OXTR variants alter receptor activation, desensitization, and internalization. (A) Activation: increase in intracellular Ca^{2+} concentration in HEK293T cells transfected with wild type (WT) or variant OXTR and treated with oxytocin. Results are normalized to WT value at highest oxytocin concentration. (B) Desensitization: increase in intracellular Ca^{2+} concentration in cells treated with 1 μ M oxytocin after pretreatment (PT) with the indicated oxytocin concentration. Results are normalized to response without PT. (C) Internalization of OXTR from the cell surface after PT with indicated oxytocin concentration. (D) Bias plot showing relative activation (y values from regression in A) and relative desensitization (regression of 1-y from B). See also Figure A.2.4. *P*-values for difference in log(EC50) or log(IC50) between WT and variant are shown (extra sum-of-squares F test, see also Tables A.2.1 and A.2.4). Error bars show standard error of the mean from N=3 independent experiments.

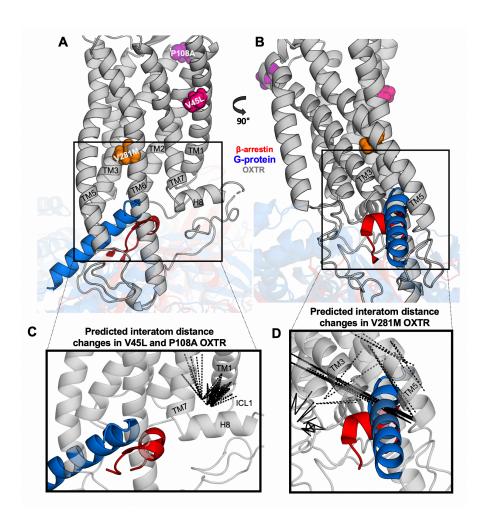


Figure 3.5 DiffNets identify distances associated with V45L, P108A, and V281M OXTR. (**A-B**) Homology model for OXTR showing the location of V45L, P108A, and V281M. Structures for β-arrestin-1 (red, PDB: 6pwc⁴⁸) and Gαs (blue, PDB: $3sn6^{49}$) are superimposed on the OXTR structure. (**C**) Dotted lines show the 100 interatom distance changes most associated with DiffNet label (V45L and P108A vs. WT and V281M). (**D**) Distance changes most associated with DiffNet label (V281M vs. WT, V45L, P108A). TM: transmembrane domain. ICL1: intracellular loop 1. H8: helix 8.

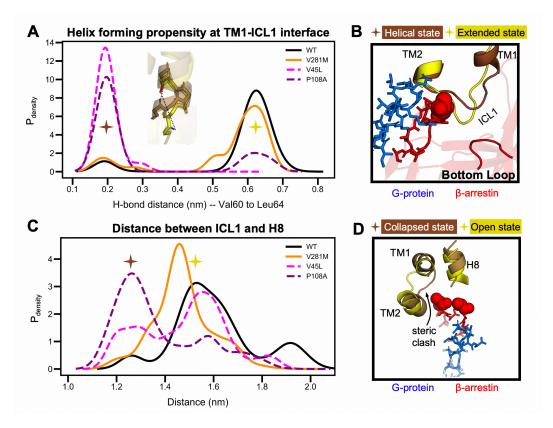


Figure 3.6 Potential mechanism for altered β-arrestin function in V45L and P108A OXTR.

(A) Distribution of the probability-weighted density for the hydrogen bond distance between the most C-terminal TM1 helix i, i+4 residue pair (Val 60 and Leu 64). β -arrestin-impaired variants (V45L and P108A) have a high probability of having a tight helix, whereas OXTRs with normal desensitization and internalization (WT and V281M) are more likely to lack this hydrogen bond. (B) Representative structures from each peak in (A). The β -arrestin-1 "bottom loop" (red), which is involved in binding to ICL1, is closer to ICL1 when ICL1 is extended. (C) Distribution of the probability-weighted density for ICL1-H8 distances that each OXTR variant occupies. V45L and P108A have strong, left-shifted peaks indicating a collapse between ICL1 and H8. (D) Representative structures of ICL1-H8 at collapsed distances (brown) and open distances (yellow). In the collapsed position, there is a steric clash between ICL1 and β -arrestin. TM: transmembrane domain. ICL1: intracellular loop 1. H8: helix 8.

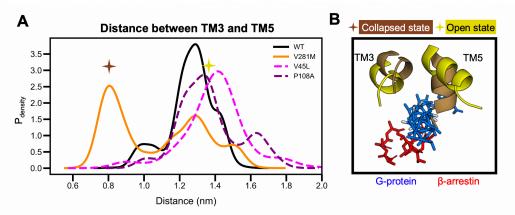


Figure 3.7 Conformational changes in V281M OXTR may reduce G protein binding.
(A) Histogram showing a probability-weighted distribution of TM3-TM5 distances that each OXTR variant occupies. V281M OXTR is highly likely to adopt a collapsed state that would sterically hinder G-protein binding.
(B) Representative structures of TM3 and TM5 at collapsed distances (brown) and open distances (yellow). The collapsed position sterically clashes with the G protein (blue) but not β-arrestin (red). TM: transmembrane domain.

Bibliography

- [1] Martin, J. A., Hamilton, B. E., Osterman, M. J. K., and Driscoll, A. K. (2021) Births: Final Data for 2019.

 National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System 70, 1-51.
- [2] Bulletins-Obstetrics, C. o. P. (2017) Practice Bulletin No. 183: Postpartum Hemorrhage. *Obstet Gynecol 130*, e168-e186. 10.1097/AOG.000000000002351.
- [3] Frey, H. A., Tuuli, M. G., England, S. K., Roehl, K. A., Odibo, A. O., Macones, G. A., and Cahill, A. G. (2015) Factors associated with higher oxytocin requirements in labor. *J Matern Fetal Neonatal Med* 28, 1614-1619. 10.3109/14767058.2014.963046.
- [4] Grotegut, C. A., Lewis, L. L., Manuck, T. A., Allen, T. K., James, A. H., Seco, A., and Deneux-Tharaux, C. (2018) The Oxytocin Product Correlates with Total Oxytocin Received during Labor: A Research Methods Study. *Am J Perinatol* 35, 78-83. 10.1055/s-0037-1606119.
- [5] Cahill, A. G., Waterman, B. M., Stamilio, D. M., Odibo, A. O., Allsworth, J. E., Evanoff, B., and Macones, G. A. (2008) Higher maximum doses of oxytocin are associated with an unacceptably high risk for uterine rupture in patients attempting vaginal birth after cesarean delivery. *Am J Obstet Gynecol* 199, 32 e31-35. 10.1016/j.ajog.2008.03.001.
- [6] Grotegut, C. A., Paglia, M. J., Johnson, L. N., Thames, B., and James, A. H. (2011) Oxytocin exposure during labor among women with postpartum hemorrhage secondary to uterine atony. *Am J Obstet Gynecol* 204, 56 e51-56. 10.1016/j.ajog.2010.08.023.
- [7] Grotegut, C. A., Gilner, J., Brancazio, L., James, A., Swamy, G. (2015) The maximal oxytocin infusion rate in labor is associated with uterine atony, In *Society for Maternal-Fetal Medicine: 2015 35th Annual Meeting: The Pregnancy Meeting*, p S86, American Journal of Obstetrics and Gynecology, San Diego, California.
- [8] Frolova, A. I., Raghuraman, N., Woolfolk, C.L., Lopez, J.D., Macones, G.A., Cahill, A.G. (2019) Effect of oxytocin maximum dose and duration of exposure on postpartum hemorrhage following vaginal delivery, In Society for Maternal-Fetal Medicine 2019: 39th Annual Meeting: The Pregnancy Meeting, pp S213-S214, American Journal of Obstetrics and Gynecology, Las Vegas, Nevada.
- [9] Hammad, I. A., Chauhan, S. P., Magann, E. F., and Abuhamad, A. Z. (2014) Peripartum complications with cesarean delivery: a review of Maternal-Fetal Medicine Units Network publications. *J Matern Fetal Neonatal Med* 27, 463-474. 10.3109/14767058.2013.818970.
- [10] Arrowsmith, S., and Wray, S. (2014) Oxytocin: its mechanism of action and receptor signalling in the myometrium. *J Neuroendocrinol* 26, 356-369. 10.1111/jne.12154.
- [11] Oakley, R. H., Laporte, S. A., Holt, J. A., Barak, L. S., and Caron, M. G. (2001) Molecular determinants underlying the formation of stable intracellular G protein-coupled receptor-beta-arrestin complexes after receptor endocytosis*. *J Biol Chem* 276, 19452-19460. 10.1074/jbc.M101450200.
- [12] Smith, M. P., Ayad, V. J., Mundell, S. J., McArdle, C. A., Kelly, E., and López Bernal, A. (2006) Internalization and Desensitization of the Oxytocin Receptor Is Inhibited by Dynamin and Clathrin Mutants in Human Embryonic Kidney 293 Cells. *Molecular Endocrinology* 20, 379-388. 10.1210/me.2005-0031.
- [13] Grotegut, C. A., Feng, L., Mao, L., Heine, R. P., Murtha, A. P., and Rockman, H. A. (2011) beta-Arrestin mediates oxytocin receptor signaling, which regulates uterine contractility and cellular migration. *Am J Physiol Endocrinol Metab* 300, E468-477. 10.1152/ajpendo.00390.2010.
- [14] Hasbi, A., Devost, D., Laporte, S. A., and Zingg, H. H. (2004) Real-time detection of interactions between the human oxytocin receptor and G protein-coupled receptor kinase-2. *Mol Endocrinol* 18, 1277-1286. 10.1210/me.2003-0440.
- [15] Reinl, E. L., Goodwin, Z. A., Raghuraman, N., Lee, G. Y., Jo, E. Y., Gezahegn, B. M., Pillai, M. K., Cahill, A. G., de Guzman Strong, C., and England, S. K. (2017) Novel oxytocin receptor variants in laboring women requiring high doses of oxytocin. *Am J Obstet Gynecol* 217, 214 e211-214 e218. 10.1016/j.ajog.2017.04.036.
- [16] Grotegut, C. A., Ngan, E., Garrett, M. E., Miranda, M. L., Ashley-Koch, A. E., and Swamy, G. K. (2017) The association of single-nucleotide polymorphisms in the oxytocin receptor and G protein-coupled receptor kinase 6 (GRK6) genes with oxytocin dosing requirements and labor outcomes. *Am J Obstet Gynecol 217*, 367.e361-367.e369. 10.1016/j.ajog.2017.05.023.

- [17] Füeg, F., Santos, S., Haslinger, C., Stoiber, B., Schäffer, L., Grünblatt, E., Zimmermann, R., and Simões-Wüst, A. P. (2019) Influence of oxytocin receptor single nucleotide sequence variants on contractility of human myometrium: an in vitro functional study. *BMC Med Genet 20*, 178. 10.1186/s12881-019-0894-8.
- [18] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Neale, B. M., Daly, M. J., and MacArthur, D. G. (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human proteincoding genes. *bioRxiv*, 531210. 10.1101/531210.
- [19] Hauser, A. S., Chavali, S., Masuho, I., Jahn, L. J., Martemyanov, K. A., Gloriam, D. E., and Babu, M. M. (2018) Pharmacogenomics of GPCR Drug Targets. *Cell* 172, 41-54.e19. 10.1016/j.cell.2017.11.033.
- [20] Koehbach, J., O'Brien, M., Muttenthaler, M., Miazzo, M., Akcan, M., Elliott, A. G., Daly, N. L., Harvey, P. J., Arrowsmith, S., Gunasekera, S., Smith, T. J., Wray, S., Göransson, U., Dawson, P. E., Craik, D. J., Freissmuth, M., and Gruber, C. W. (2013) Oxytocic plant cyclotides as templates for peptide G protein-coupled receptor ligand design. *Proc Natl Acad Sci U S A 110*, 21183-21188. 10.1073/pnas.1311183110.
- [21] Imoukhuede, P. I., and Popel, A. S. (2011) Quantification and cell-to-cell variation of vascular endothelial growth factor receptors. *Exp Cell Res* 317, 955-965. 10.1016/j.yexcr.2010.12.014.
- [22] Hall, D. A., and Langmead, C. J. (2010) Matching models to data: a receptor pharmacologist's guide. *Br J Pharmacol* 161, 1276-1290. 10.1111/j.1476-5381.2010.00879.x.
- [23] Meddings, J. B., Scott, R. B., and Fick, G. H. (1989) Analysis and comparison of sigmoidal curves: application to dose-response data. *Am J Physiol* 257, G982-989. 10.1152/ajpgi.1989.257.6.G982.
- [24] Zhang, J., Yang, J., Jang, R., and Zhang, Y. (2015) GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome. *Structure 23*, 1538-1549. 10.1016/j.str.2015.06.007.
- [25] Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008) CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem 29*, 1859-1865. 10.1002/jcc.20945.
- [26] Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., Wei, S., Buckner, J., Jeong, J. C., Qi, Y., Jo, S., Pande, V. S., Case, D. A., Brooks, C. L., MacKerell, A. D., Klauda, J. B., and Im, W. (2016) CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput* 12, 405-413. 10.1021/acs.jctc.5b00935.
- [27] Wu, E. L., Cheng, X., Jo, S., Rui, H., Song, K. C., Dávila-Contreras, E. M., Qi, Y., Lee, J., Monje-Galvan, V., Venable, R. M., Klauda, J. B., and Im, W. (2014) CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *J Comput Chem* 35, 1997-2004. 10.1002/jcc.23702.
- [28] Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009) CHARMM: the biomolecular simulation program. *J Comput Chem 30*, 1545-1614. 10.1002/jcc.21287.
- [29] Jo, S., Cheng, X., Islam, S. M., Huang, L., Rui, H., Zhu, A., Lee, H. S., Qi, Y., Han, W., Vanommeslaeghe, K., MacKerell, A. D., Roux, B., and Im, W. (2014) CHARMM-GUI PDB manipulator for advanced modeling and simulations of proteins containing nonstandard residues. *Adv Protein Chem Struct Biol* 96, 235-265. 10.1016/bs.apcsb.2014.06.002.
- [30] Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem 24*, 1999-2012. 10.1002/jcc.10349.
- [31] Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., and MacKerell, A. D. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* 14, 71-73. 10.1038/nmeth.4067.

- [32] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX 1-2*, 19-25. 10.1016/j.softx.2015.06.001.
- [33] Hess, B. (2008) P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J Chem Theory Comput 4*, 116-122. 10.1021/ct700200b.
- [34] Evans, D. J., and Holian, B. L. (1985) The Nose–Hoover thermostat. *The Journal of Chemical Physics 83*, 4069-4074. 10.1063/1.449071.
- [35] Ivanova, N., and Ivanova, A. (2018) Testing the limits of model membrane simulations-bilayer composition and pressure scaling. *J Comput Chem* 39, 387-396. 10.1002/jcc.25117.
- [36] Zimmerman, M. I., and Bowman, G. R. (2015) FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J Chem Theory Comput 11*, 5747-5757. 10.1021/acs.jctc.5b00737.
- [37] Zimmerman, M. I., and Bowman, G. R. (2016) How to Run FAST Simulations. *Methods Enzymol* 578, 213-225. 10.1016/bs.mie.2016.05.032.
- [38] Zimmerman, M. I., Hart, K. M., Sibbald, C. A., Frederick, T. E., Jimah, J. R., Knoverek, C. R., Tolia, N. H., and Bowman, G. R. (2017) Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS Cent Sci* 3, 1311-1321. 10.1021/acscentsci.7b00465.
- [39] Ward, M. D., Zimmerman, M. I., Meller, A., Chung, M., Swamidass, S. J., and Bowman, G. R. (2021) Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets. *Nat Commun* 12, 3023. 10.1038/s41467-021-23246-1.
- [40] (2014) An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation, Vol. 797, Springer, Netherlands: Dordrecht.
- [41] Chodera, J. D., and Noé, F. (2014) Markov state models of biomolecular conformational dynamics. *Curr Opin Struct Biol* 25, 135-144. 10.1016/j.sbi.2014.04.002.
- [42] Schütte, C., and Sarich, M. (2013) Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches.
- [43] Porter, J. R., Zimmerman, M. I., and Bowman, G. R. (2019) Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *J Chem Phys* 150, 044108. 10.1063/1.5063794.
- [44] Shrake, A., and Rupley, J. A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* 79, 351-371. 10.1016/0022-2836(73)90011-9.
- [45] McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L. P., Lane, T. J., and Pande, V. S. (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* 109, 1528-1532. 10.1016/j.bpj.2015.08.015.
- [46] Gonzalez, T. F. (1985) Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science* 38, 293-306. 10.1016/0304-3975(85)90224-5.
- [47] Zimmerman, M. I., Porter, J. R., Sun, X., Silva, R. R., and Bowman, G. R. (2018) Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *J Chem Theory Comput 14*, 5459-5475. 10.1021/acs.jctc.8b00500.
- [48] Yin, W., Li, Z., Jin, M., Yin, Y. L., de Waal, P. W., Pal, K., Yin, Y., Gao, X., He, Y., Gao, J., Wang, X., Zhang, Y., Zhou, H., Melcher, K., Jiang, Y., Cong, Y., Edward Zhou, X., Yu, X., and Eric Xu, H. (2019) A complex structure of arrestin-2 bound to a G protein-coupled receptor. *Cell Res* 29, 971-983. 10.1038/s41422-019-0256-2.
- [49] Rasmussen, S. G., DeVree, B. T., Zou, Y., Kruse, A. C., Chung, K. Y., Kobilka, T. S., Thian, F. S., Chae, P. S., Pardon, E., Calinski, D., Mathiesen, J. M., Shah, S. T., Lyons, J. A., Caffrey, M., Gellman, S. H., Steyaert, J., Skiniotis, G., Weis, W. I., Sunahara, R. K., and Kobilka, B. K. (2011) Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* 477, 549-555. 10.1038/nature10361.
- [50] Gundry, J., Glenn, R., Alagesan, P., and Rajagopal, S. (2017) A Practical Guide to Approaching Biased Agonism at G Protein Coupled Receptors. *Front Neurosci* 11, 17. 10.3389/fnins.2017.00017.
- [51] Suomivuori, C. M., Latorraca, N. R., Wingler, L. M., Eismann, S., King, M. C., Kleinhenz, A. L. W., Skiba, M. A., Staus, D. P., Kruse, A. C., Lefkowitz, R. J., and Dror, R. O. (2020) Molecular mechanism of biased signaling in a prototypical G protein-coupled receptor. *Science* 367, 881-887. 10.1126/science.aaz0326.
- [52] Ma, W. J., Hashii, M., Munesue, T., Hayashi, K., Yagi, K., Yamagishi, M., Higashida, H., and Yokoyama, S. (2013) Non-synonymous single-nucleotide variations of the human oxytocin receptor gene and autism spectrum disorders: a case-control study in a Japanese population and functional analysis. *Mol Autism 4*, 22. 10.1186/2040-2392-4-22.
- [53] Kim, J., Stirling, K. J., Cooper, M. E., Ascoli, M., Momany, A. M., McDonald, E. L., Ryckman, K. K., Rhea, L., Schaa, K. L., Cosentino, V., Gadow, E., Saleme, C., Shi, M., Hallman, M., Plunkett, J., Teramo, K. A.,

- Muglia, L. J., Feenstra, B., Geller, F., Boyd, H. A., Melbye, M., Marazita, M. L., Dagle, J. M., and Murray, J. C. (2013) Sequence variants in oxytocin pathway genes and preterm birth: a candidate gene association study. *BMC Med Genet 14*, 77. 10.1186/1471-2350-14-77.
- [54] Koole, C., Wootten, D., Simms, J., Valant, C., Miller, L. J., Christopoulos, A., and Sexton, P. M. (2011) Polymorphism and ligand dependent changes in human glucagon-like peptide-1 receptor (GLP-1R) function: allosteric rescue of loss of function mutation. *Mol Pharmacol* 80, 486-497. 10.1124/mol.111.072884.
- [55] Ringholm, A., Klovins, J., Rudzish, R., Phillips, S., Rees, J. L., and Schioth, H. B. (2004) Pharmacological characterization of loss of function mutations of the human melanocortin 1 receptor that are associated with red hair. *J Invest Dermatol* 123, 917-923. 10.1111/j.0022-202X.2004.23444.x.
- [56] Costa, E. M., Bedecarrats, G. Y., Mendonca, B. B., Arnhold, I. J., Kaiser, U. B., and Latronico, A. C. (2001) Two novel mutations in the gonadotropin-releasing hormone receptor gene in Brazilian patients with hypogonadotropic hypogonadism and normal olfaction. *J Clin Endocrinol Metab* 86, 2680-2686. 10.1210/jcem.86.6.7551.
- [57] Gorvin, C. M., Babinsky, V. N., Malinauskas, T., Nissen, P. H., Schou, A. J., Hanyaloglu, A. C., Siebold, C., Jones, E. Y., Hannan, F. M., and Thakker, R. V. (2018) A calcium-sensing receptor mutation causing hypocalcemia disrupts a transmembrane salt bridge to activate β-arrestin-biased signaling. *Sci Signal* 1110.1126/scisignal.aan3714.
- [58] Daly, D., Minnie, K. C. S., Blignaut, A., Blix, E., Vika Nilsen, A. B., Dencker, A., Beeckman, K., Gross, M. M., Pehlke-Milde, J., Grylka-Baeschlin, S., Koenig-Bachmann, M., Clausen, J. A., Hadjigeorgiou, E., Morano, S., Iannuzzi, L., Baranowska, B., Kiersnowska, I., and Uvnäs-Moberg, K. (2020) How much synthetic oxytocin is infused during labour? A review and analysis of regimens used in 12 countries. *PLoS One 15*, e0227941. 10.1371/journal.pone.0227941.
- [59] Lotta, L. A., Mokrosiński, J., Mendes de Oliveira, E., Li, C., Sharp, S. J., Luan, J., Brouwers, B., Ayinampudi, V., Bowker, N., Kerrison, N., Kaimakis, V., Hoult, D., Stewart, I. D., Wheeler, E., Day, F. R., Perry, J. R. B., Langenberg, C., Wareham, N. J., and Farooqi, I. S. (2019) Human Gain-of-Function MC4R Variants Show Signaling Bias and Protect against Obesity. *Cell* 177, 597-607.e599. 10.1016/j.cell.2019.03.044.
- [60] Ameur, A., Dahlberg, J., Olason, P., Vezzi, F., Karlsson, R., Martin, M., Viklund, J., Kähäri, A. K., Lundin, P., Che, H., Thutkawkorapin, J., Eisfeldt, J., Lampa, S., Dahlberg, M., Hagberg, J., Jareborg, N., Liljedahl, U., Jonasson, I., Johansson, Å., Feuk, L., Lundeberg, J., Syvänen, A. C., Lundin, S., Nilsson, D., Nystedt, B., Magnusson, P. K., and Gyllensten, U. (2017) SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet* 25, 1253-1260. 10.1038/ejhg.2017.130.
- [61] Ballesteros, J. A., and Weinstein, H. (1995) [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors, In *Methods in Neurosciences* (Sealfon, S. C., Ed.), pp 366-428, Academic Press.

Chapter 4

SARS-CoV-2 Simulations Go Exascale to Predict Dramatic Spike Opening and Cryptic Pockets Across the Proteome

4.1 Preamble

This chapter is adapted from the following article: Zimmerman, M.I., Porter, J.R., Ward, M.D., Singh S., Vithani N., Meller, A., Mallimadugula, U.L., Kuhn, C.E., Borowsky, J.H., Wiewiora, R.P., Hurley, M.F.D, Harbison, A.M., Fogarty, C.A., Coffland, J.E., Fadda, E., Voelz, V.A., Chodera, J.D., and Bowman, G.R (2021). "SARS-CoV-2 Simulations Go Exascale to Predict Dramatic Spike Opening and Cryptic Pockets Across the Proteome", *Nature Chemistry* 13, 651-659.

4.2 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus that poses an imminent threat to global human health and socioeconomic stability. With estimates of the basic reproduction number at ~3-4 and a case fatality rate for coronavirus disease 2019 (COVID-19) ranging from ~0.1-12% (high temporal variation), SARS-CoV-2/COVID-19 has spread quickly and currently endangers the global population. As of September 12th, 2020, there have been over 29 million confirmed cases and over 925,000 fatalities, globally. Quarantines and social distancing are effective at slowing the rate of transmission; however, they cause significant social and economic disruption. Taken together, it is crucial that we find immediate therapeutic interventions.

A structural understanding of the SARS-CoV-2 proteins could accelerate the discovery of new therapeutics by enabling the use of rational design.⁷ Towards this end, the structural biology

community has made heroic efforts to rapidly build models of SARS-CoV-2 proteins and the complexes they form. 8–16 However, it is well established that a protein's function is dictated by the full range of conformations it can access; many of which remain hidden to experimental methods. Mapping these conformations for SARS-CoV-2 proteins will provide a clearer picture of how they enable the virus to perform diverse functions, such as infecting cells, evading a host's immune system, and replicating. Such maps may also present new therapeutic opportunities, such as 'cryptic' pockets that are absent in experimental snapshots but provide novel targets for drug discovery.

Molecular dynamics simulations have the ability to capture the full ensemble of structures a protein adopts but require significant computational resources. Such simulations capture an all-atom representation of the range of motions a protein undergoes. Modern datasets often consist of a few microseconds of simulation for a single protein, with a few noteworthy examples reaching millisecond timescales. However, many important processes occur on slower timescales. Moreover, simulating every protein that is relevant to SARS-CoV-2 for biologically relevant timescales would require compute resources on an unprecedented scale.

To overcome this challenge, more than a million citizen scientists from around the world have donated their computer resources to simulate SARS-CoV-2 proteins. This massive collaboration was enabled by the Folding@home distributed computing platform, which has crossed the exascale computing barrier and is now the world's largest supercomputer. Using this resource, we constructed quantitative maps of the structural ensembles of over two dozen proteins and complexes that pertain to SARS-CoV-2 from milliseconds of simulation data generated for each system. Together, we have run an unprecedented 0.1 s of simulation. Our data uncover the mechanisms of conformational changes that are essential for SARS-CoV-2's

replication cycle and reveal a multitude of new therapeutic opportunities. The data are supported by a variety of experimental observations and are being made publicly available (https://covid.molssi.org/ and https://cov

4.3 Results

4.3.1 To the exascale and beyond!

Folding@home (http://foldingathome.org) is a community of citizen scientists, researchers, and tech organizations dedicated to applying their collective computational and intellectual resources to understand the role of proteins' dynamics in their function and dysfunction, and to aid in the design of new proteins and therapeutics.²¹ It enables anyone with a computer and an internet connection to contribute to biomedical research by volunteering to run small chunks of simulation, called "work units," that are used to build maps of protein dynamics. The project has provided insight into diverse topics, ranging from protein folding to signaling mechanisms.^{22–24} to the connection between phenotype and genotype.^{25–27} Translational applications have included new means to combat antimicrobial resistance, Ebola virus, and SFTS virus.^{28–30}

In response to the COVID-19 pandemic, Folding@home quickly pivoted to focus on SARS-CoV-2 and the host factors it interacts with. Many people found the opportunity to take action at a time when they were otherwise feeling helpless alluring. In less than three months, the project grew from ~30,000 active devices to over a million devices around the globe (Fig. 4.1A and 1B).

We conservatively estimate the peak performance of Folding@home hit 1.01 exaFLOPS. This performance was achieved at a point when ~280,000 GPUs and 4.8 million CPU cores were performing simulations. As explained in the Methods, to be conservative about our claims, we

assume that each GPU/CPU has worse performance than a card released before 2015. For reference, the aggregate 1 exaFLOPS performance we report for Folding@home is 5-fold greater than the peak performance of the world's fastest traditional supercomputer at the time, called Summit (Fig. 4.1C). It is also more than the top 100 supercomputers combined. Prior to Folding@home, the first exascale supercomputer was not scheduled to come online until the end of 2021.

4.3.2 Extreme spike opening reveals cryptic epitopes

The Spike complex (S) is a prominent vaccine target that is known to undergo substantial conformational changes as part of its function. ^{10,14,31} Structurally, S is composed of three interlocking proteins, with each chain having a cleavage site separating an S1 and S2 fragment. S resides on the virion surface, where it waits to engage with an angiotensin-converting enzyme 2 (ACE2) receptor on a host cell to trigger infection. ^{32,33} The fact that S is exposed on the virion surface makes it an appealing vaccine target. However, it has a number of effective defense strategies. First, S is decorated extensively with glycans that aid in immune evasion by shielding potential antigens. ^{34,35} S also uses a conformational masking strategy, wherein it predominantly adopts a closed conformation (often called the down state) that buries the receptor-binding domains (RBDs) to evade immune surveillance mechanisms. To engage with ACE2, S must somehow expose the conserved binding interface of the RBDs. Characterizing the full range of S opening is important for understanding pathogenesis and could provide insights into novel therapeutic options.

To capture S opening, we employed our goal-oriented adaptive sampling algorithm, FAST, in conjunction with Folding@home. The FAST method^{36,37} iterates between running a batch of simulations, building a map of conformational space called a Markov state model

(MSM)^{38,39} from all the data generated so far, ranking the conformational states of this MSM based on how likely starting a new simulation from that state is to yield useful data, and starting a new batch of simulations from the top ranked states. The ranking function is designed to balance between favoring structures with a desired geometric feature (in this case opening of S) and broad exploration of conformational space. By balancing exploration-exploitation tradeoffs, FAST often captures conformational changes with orders of magnitude less simulation time than alternative methods. Broadly distributed structures from our FAST simulations were then used as starting points for extensive Folding@home simulations, totaling over 1 millisecond of data for SARS-CoV-2 S, enabling us to obtain a statistically sound final model.

Our SARS-CoV-2 S protein simulations predict extreme opening of S and substantial conformational heterogeneity in the open state (Fig. 4.2). Capturing opening of S is an impressive technical feat. Other large-scale simulations have provided valuable insight into aspects of S, but were unable to capture this essential event for the initiation of infection. ^{28,32,33} For example, Casalino *et al.* performed ~10 microseconds of simulation to show that one of the glycans helps stabilize a partially open state and Turonová *et al.* performed 2.5 microseconds of simulation that revealed three hinges in the stalk. ^{35,40} However, the shorter timescale of these simulations prevented the authors from capturing the opening process at all. With our milliseconds of sampling, we successfully ^{35,40,41}captured this rare event for both glycosylated and unglycosylated S and find that glycosylation slightly increases the population of the open state, but the difference between glycosylated and unglycosylated S is smaller than that between different spike variants (Fig. A.3.1). The closed state is more probable than the open state, explaining the experimental observation that full-length S has a lower affinity for ACE2 than an isolated RBD. ⁴² Intriguingly, we find that opening occurs only for a single RBD at a time, akin to

the up state observed in cryoEM structures.⁴³ Moreover, we predict that the scale of S opening is often substantially larger than has been observed in experimental snapshots in the absence of binding partners (Fig. A.3.2).

The dramatic opening we discover predicts that antibodies, and other therapeutics, can bind to regions of S that are deeply buried and seemingly inaccessible in existing experimental snapshots. 9,13,44,45 Consistent with this prediction, the cryptic epitope for the antibody CR3022 is buried in up and down cryoEM structures, but is clearly exposed in our conformational ensemble (Fig. 4.2C). Indeed, our ensemble captures the exposure of many known epitopes, despite their occlusion in apo experimental snapshots (Fig. 4.2D). Our models also provide a quantitative estimate of the probability that different epitopes are exposed, is consistent with experimental measures of dynamics, and can be used to determine the most suitable regions for the design of neutralizing antibodies.

Our results suggest that S binds ACE2 and many antibodies via a conformational selection mechanism, wherein S first opens and then binds to its partners. Previous work based on examining the up and down structures observed by cryoEM also proposed a role for conformational selection, hypothesizing that an S RBD may bind CR3022 by first adopting an up conformation and then twisting to expose the cryptic epitope. To test this hypothesis, we projected the free energy landscape and the highest-flux pathway for S opening onto two order parameters: the angle of RBD opening and the twist of the RBD (Fig. A.3.3). We find that the RBD simultaneously twists and peels off of the spike complex as it transitions from the closed to open conformation. Furthermore, the motion we observe predicts the exposure of other epitopes that would not be exposed by the mechanism proposed by Yuan et al. These additional epitopes have now been corroborated by work on the binding sites of other antibodies (Fig. 4.2D).

To understand the potential role of conformational masking in determining the lethality and infectivity of different coronaviruses, we also simulated the opening of S proteins from two related viruses: SARS-CoV-1 and HCoV-NL63. These viruses were selected because they also bind the ACE2 receptor but are associated with varying mortality rates. SARS-CoV-1 caused an outbreak in 2003 with a high case fatality rate but has not become a pandemic. NL63 was discovered the following year and continues to spread around the globe, although it is significantly less lethal than either SARS virus. The significantly less lethal than either SARS virus.

We hypothesized that phenotypic differences between coronaviruses may be partially explained by changes to the S conformational ensemble, particularly the probability of spike opening. Specifically, we propose mutations or other perturbations can increase the S-ACE2 affinity by increasing the probability that S adopts an open conformation or by increasing the affinity between an exposed RBD and ACE2. In contrast, the affinity of S for ACE2 (or antibodies that bind cryptic epitopes) can be reduced by stabilizing the closed state or decreasing the affinity between an exposed RBD and its binding partner(s).

As expected, the three S complexes have very different propensities to adopt an open state and bind ACE2. Structures from each ensemble were classified as competent to bind ACE2 if superimposing an ACE2-RBD structure on S did not result in any steric clashes between ACE2 and the rest of the S complex. We find that SARS-CoV-1 has the highest population of conformations that can bind to ACE2 without steric clashes, followed by SARS-CoV-2, while opening of NL63 is sufficiently rare that we did not observe ACE2-binding competent conformations in our simulations (Fig. 4.2B). Interestingly, S proteins that are more likely to adopt structures that are competent to bind ACE2 are also more likely to adopt highly open structures (Fig 4.2C).

We also predict a number of interesting correlations between the conformational masking, lethality, and infectivity of different coronaviruses. First, more deadly coronaviruses have S proteins with less conformational masking. Second, there is an inverse correlation between S opening and the affinity of an isolated RBD for ACE2 (RBD-ACE2 affinities of ~35 nM, ~44 nM, and ~185 nM for HCoV-NL63, SARS-CoV-2, and SARS-CoV-1, respectively). 48,49

These observations suggest a tradeoff wherein stabilizing the closed spike enables immune evasion but hampers cell entry, requiring a higher affinity between an exposed RBD and ACE2 to reliably infect a host cell. We propose that the NL63 S complex is probably best at evading immune detection but is not as infectious as the SARS viruses because strong conformational masking reduces the overall affinity for ACE2. In contrast, the SARS-CoV-1 S complex adopts open conformations more readily but is also more readily detected by immune surveillance mechanisms. Finally, SARS-CoV-2 balances conformational masking and the RBD-ACE2 affinity in a manner that allows it to evade an immune response while maintaining its ability to infect a host cell.

Our atomically detailed model of glycosylated S can facilitate structure-based vaccine antigen design through identification of regions minimally protected by conformational masking or the glycan shield.⁵⁰ To identify these potential epitopes, we calculated the probability that each residue in S could be exposed to therapeutics (e.g. not shielded by a glycan or buried by conformational masking), as shown in Fig. 4.3A. Visualizing these values on the protein reveals a few patches of protein surface that are exposed through the glycan shielding (Fig. 4.3B). However, another important factor when targeting an antigen is picking a region with a conserved sequence to yield broader and longer lasting efficacy. Not surprisingly, many of the

exposed regions do not have a strongly conserved sequence. Promisingly, though, we do find a conserved area with a larger degree of solvent exposure (Fig. 4.3C). This region was recently found to be an effective site for neutralizing antibodies.⁵¹ Another possibility for antigen design is to exploit the opening motion. A number of residues surrounding the receptor binding motif (RBM) of the RBD show an increase in exposure by ~30% in ACE2 binding competent structures (Fig. 4.3C). Consistent with immunoassays and cryoEM structures, these regions are hotspots for neutralizing antibody binding.^{9,52,53}

4.3.3 Cryptic pockets and functional dynamics are present throughout the proteome

Every protein in SARS-CoV-2 remains a potential drug target. So, to understand their role in disease and help progress the design of antivirals, we unleashed the full power of Folding@home to simulate dozens of systems related to pathogenesis. While we are interested in all aspects of a proteins' functional dynamics, expanding on the number of antiviral targets is of immediate value. Towards this end, we seeded Folding@home simulations from our FAST-pockets adaptive sampling to aid in the discovery of cryptic pockets. We briefly discuss two illustrative examples, out of 36 datasets.

Nonstructural protein number 5 (NSP5, also named the main protease, 3CL^{pro}, or as we will refer to it, M^{pro}) is an essential protein in the lifecycle of coronaviruses, cleaving polyprotein 1a into functional proteins, and is a major target for the design of antivirals. 11 It is highly conserved between coronaviruses and shares 96% sequence identity with SARS-CoV-1 M^{pro}; it cleaves polyprotein 1a at no fewer than 11 distinct sites, placing significant evolutionary constraint on its active site. M^{pro} is only active as a dimer, however it exists in a monomer-dimer equilibrium with estimates of its dissociation constant in the low µM range. 54 Small molecules

targeting this protein to inhibit enzymatic activity, either by altering its active site or favoring the inactive monomer state, would be promising broad-spectrum antiviral candidates.⁵⁵

Our simulations predict two novel cryptic pockets on M^{pro} that expand our current therapeutic options. These are shown in Fig. 4.4A, which projects states from our MSM onto the solvent exposure of residues that make up the pockets. The first cryptic pocket is an expansion of NSP5's catalytic site. We predict that the loop bridging domains II and III is highly dynamic and can fully undock from the rest of the protein. This motion may impact catalysis—i.e. by sterically regulating substrate binding—and is similar to motions we have observed previously for the enzyme β-lactamase.⁵⁶ Owing to its location, a small molecule bound in this pocket is likely to prevent catalysis by obstructing polypeptide association with catalytic residues. The second pocket is a large opening between domains I/II and domain III. Located at the dimerization interface, this pocket offers the possibility to find small molecules or peptides that favor the inactive monomer state. We have repeated these calculations and found that the discovery of cryptic pockets is robust to the choice of forcefield (Fig. A.3.4).

In addition to cryptic pockets, our data captures many potentially functionally relevant motions within the SARS-CoV-2 proteome. We illustrate this with the SARS-CoV-2 nucleoprotein. The nucleoprotein is a multifunctional protein responsible for major lifecycle events such as viral packaging, transcription, and physically linking RNA to the envelope. ^{57,58} As such, we expect the protein to accomplish these goals through a highly dynamic and rich conformational ensemble, akin to context-dependent regulatory modules observed in Ebola virus nucleoprotein. ^{59,60} Investigating the RNA-binding domain, we predict both cryptic pockets and an incredibly dynamic beta-hairpin, which hosts the RNA binding site, referred to as a "positive finger" (Fig. 4.4C-D). Our observed conformational heterogeneity of the positive finger is

consistent with a structural ensemble determined using solution-state nuclear magnetic resonance (NMR) spectroscopy.⁶¹ Our simulations also capture numerous states of the putative RNA binding pose, where the positive finger curls up to form a cradle for RNA. These states can provide a structural basis for the design of small molecules that would compete with RNA binding, preventing viral assembly.

The data we present in this paper represents the single largest collection of all-atom simulations. Table 4.1 is a comprehensive list of the systems we have simulated. Systems span various oligomerization states, include important complexes, and include representation from multiple coronaviruses. We also include human proteins that are targets for supportive therapies and preventative treatments. To accelerate the discovery of new therapeutics and promote open science, our MSMs and structures of cryptic pockets are available online (https://covid.molssi.org/ and https://osf.io/fs2yv/). For each system analyzed, we provide a detailed Markov model and relevant analysis. For cryptic pockets, we provide two directories, model and cryptic pockets, as well as a README.dat that details all hyperparameters used for model construction. The model directory contains the following files: full centers.xtc (GROMACS binary of cluster centers), populations.npy (numpy binary file of equilibrium populations), prot masses.pdb (PDB topology file), tcounts.npy (numpy binary of the transition count matrix), and tprobs.npy (numpy binary of the transition probability matrix). For each cryptic pocket 'X' that we characterize, there exists a cryptic pockets/pocketX resis.dat and cryptic pockets/pocketX rankings.dat, which details the residues that are present in the cryptic pocket and a list of states with cryptic pockets ranked from most open to most closed. Other contemporary works are already building on these data, providing new insight into

multiple systems (i.e. NSP16, Spike protein, and nucleoprotein) and making new connections with experiments. 59,62,63

4.4 Conclusions

The Folding@home community has created one of the largest computational resources in the world to tackle a global threat. Over a million citizen scientists have pooled their computer resources to help understand and combat COVID-19, generating more than 0.1 seconds of simulation data. The unprecedented scale of these simulations has helped to characterize crucial stages of infection. We predict that Spike proteins have a strong trade-off between making ACE2 binding interfaces accessible to infiltrate cells and conformationally masking epitopes to subvert immune responses. SARS-CoV-2 represents a more optimal tradeoff than related coronaviruses, which may explain its success in spreading globally. Our simulations also provide an atomically detailed roadmap for designing vaccines and antivirals. For example, we have made a comprehensive atlas and repository of cryptic pockets hosted online to accelerate the development of novel therapeutics. Many groups are already using our data, such as the COVID Moonshot,⁶⁷ an international collaboration between multiple computational and experimental groups working to develop a patent-free inhibitor of the main protease.

Beyond SAR S-CoV-2, we expect this work to aid in a better understanding of the roles of proteins in the *coronaviridae* family. Coronaviruses have been around for millennia, yet many of their proteins are still poorly understood. Because climate change has made zoonotic transmission events more commonplace, it is imperative that we continue to perform basic research on these viruses to better protect us from future pandemics. For each protein system in Table 4.1, an extraordinary amount of sampling has led to the generation of a quantitative map of

its conformational landscape. There is still much to learn about coronavirus function and these conformational ensembles contain a wealth of information to pull from.

While we have aggressively targeted research on SARS-CoV-2, Folding@home is a general platform for running molecular dynamics simulations at scale. Before the COVID-19 pandemic, Folding@home was already generating datasets that were orders of magnitude greater than from conventional means. With our explosive growth, our compute power has increased around 100-fold. Our work here highlights the incredible utility this compute power has to rapidly understand health and disease, providing a rich source of structural data for accelerating the design of therapeutics. With the continued support of the citizen scientists that have made this work possible, we have the opportunity to make a profound impact on other global health crises such as cancer, neurodegenerative diseases, and antibiotic resistance.

4.5 Methods

4.5.1 System preparation

All simulations were prepared using Gromacs 2020.⁶⁸ Initial structures were placed in a dodecahedral box that extends 1.0 nm beyond the protein in any dimension. Systems were then solvated and energy minimized with a steepest descents algorithm until the maximum force fell below 100 kJ/mol/nm using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor list, Coulomb interactions, and van der Waals interactions. The AMBER03 force field was used for all systems except Spike protein with glycans, which used CHARMM36.^{69,70} We chose to use the AMBER03 forcefield for the discovery of cryptic pockets since we have had extensive success experimentally confirming predictions based on simulations using this forcefield on other systems.⁷¹ We have also found that AMBER03 gives comparable results to other force fields given sufficient sampling.⁷² Furthermore, we find that discovery of cryptic

pockets on NSP5 is robust to the choice of forcefield (Fig. A.3.4). All simulations were simulated with explicit TIP3P solvent.⁷³

Systems were then equilibrated for 1.0 ns, where all bonds were constrained with the LINCS algorithm and virtual sites were used to allow a 4 fs time step.⁷⁴ Cutoffs of 1.1 nm were used for the neighbor list with 0.9 for Coulomb and van der Waals interactions. The particle mesh ewald method was employed for treatment of long-range interactions with a fourier spacing of 0.12 nm. The Verlet cutoff scheme was used for the neighbor list. The stochastic velocity rescaling (*v*-rescale) thermostat was used to hold the temperature at 300 K.⁷⁵

4.5.2 Adaptive sampling simulations

The FAST algorithm was employed for each protein in Table 4.1 to enhance conformational sampling and quickly explore dominant motions. The procedure for FAST simulations is as follows: 1) run initial simulations, 2) build MSM, 3) rank states based on FAST ranking, 4) restart simulations from the top ranked states, 5) repeat steps 2-4 until ranking is optimized. For each system, MSMs were generated after each round of sampling using a *k*-centers clustering algorithm based on the RMSD between select atoms. Clustering continued until the maximum distance of a frame to a cluster center fell within a predefined cutoff. In addition to the FAST ranking, a similarity penalty was added to promote conformational diversity in starting structures, as has been described previously.⁷⁶ The code used to run FAST simulations can be found online (https://github.com/bowman-lab/fast).

FAST-distance simulations of all Spike proteins were run at 310 K on the Microsoft Azure cloud computing platform. The FAST-distance ranking favored states with greater RBD openings using a set of distances between atoms. Each round of sampling was performed with 22 independent simulations that were 40 ns in length (0.88 µs aggregate sampling per round), where

the number of rounds totaled 13 (11.44 μ s), 22 (19.36 μ s), and 17 (14.96 μ s), for SARS-CoV-1, SARS-CoV-2, and HCoV-NL63, respectively.

For all other proteins, FAST-pocket simulations were run at 300 K for 6 rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 µs aggregate simulation). The FAST-pocket ranking function favored restarting simulations from states with large pocket openings. Pocket volumes were calculated using the LIGSITE algorithm.⁷⁷

4.5.3 Folding@home simulations

For each adaptive sampling run, a conformationally diverse set of structures was selected to be run on Folding@home. Structures came from the final k-centers clustering of adaptive sampling, as is described above. Simulations were deployed using a simulation core based on either GROMACS 5.0.4 or OpenMM 7.4.1. 68,78

Estimating the aggregate compute power of Folding@home is non-trivial due to factors like hardware heterogeneity, measures to maintain volunteers' anonymity, and the fact that volunteers can turn their machines on and off at-will. Furthermore, volunteers' machines only communicate with the Folding@home servers at the beginning and end of a work unit, with the intervening time taking anywhere from tens of minutes to a few days depending on the volunteer's hardware and the protein to simulate. Therefore, we chose to estimate the performance by counting the number of GPUs and CPUs that participated in Folding@home during a three-day window and making a conservative assumption about the computational performance of each device. We note that a larger time window has been used on our website for historical reasons. We make the conservative assumption that each CPU core performs at 0.0127 TFLOPS and each GPU at 1.672 native TFLOPS (or 3.53 X86-equivalent TFLOPS), as explained in our long-standing performance estimate (https://stats.foldingathome.org/os). For

reference, a GTX 980 (which was released in 2014) can achieve 5 native TFLOPS (or 10.56 X86-equivalent TFLOPS). An Intel Core i7 4770K (released in 2013) can achieve 0.046 TFLOPS/core. We report x86-equivalent FLOPS.

4.5.4 Markov state models

A Markov state model is a network representation of a free energy landscape and is a key tool for making sense of molecular dynamics simulations. 39,79 All MSMs were built using our python package, enspara (https://github.com/bowman-lab/enspara). Each system was clustered with the combined FAST and Folding@home datasets. In the case of Spike proteins, states were defined geometrically based on the RMSD between backbone C_a coordinates. States were generated as the top 3000 centers from a k-centers clustering algorithm. All other proteins were clustered based on the Euclidean distance between the solvent accessible surface area of residues, as is described previously. Select systems generated either 2500, 5000, 7500, or 10000 cluster centers from a k-centers clustering algorithm. Select systems were refined with 1-10 k-medoid sweeps. Transition probability matrices were produced by counting transitions between states, adding a prior count of $1/n_{states}$, and row-normalizing, as is described previously. Equilibrium populations were calculated as the eigenvector of the transition probability matrix with an eigenvalue of one.

4.5.5 Spike/ACE2 binding competency

To determine Spike protein binding competency to ACE2 the following structures of the RBD bound to ACE2 were used: 3D0G, 6M0J, and 3KBH, for SARS-CoV-1, SARS-CoV-2, and HCoV-NL63, respectively. The RBD of the bound complex was superimposed onto each RBD for structures in our MSM. Steric clashes were then determined between backbone atoms on the

ACE2 molecule and the rest of the spike protein. If any of the structures had a superposition that resulted in no clashes, it was deemed binding competent. The final population of binding competent states was determined as the sum of state populations that were deemed binding competent. Error bars were obtained from bootstrapping the MSM equilibrium populations, as implemented in enspara.

4.5.6 Cryptic pockets and solvent accessible surface area

For ease of detecting cryptic pockets and other functional motions, we employed our exposon analysis method.⁵⁶ This method correlates the solvent exposure between residues to find concerted motions that tend to represent cryptic pocket openings. Solvent accessible surface area calculations were computed using the Shrake-Rupley algorithm as implemented in the python package MDTraj.⁸² For all proteins and complexes, a solvent probe radius of 0.28 nm was used, which has been shown to produce a reasonable clustering and exposon map.⁵⁶

Spike protein solvent accessible surface areas for SARS-CoV-2 were computed with glycan chains modeled onto each cluster center. Multiple glycan rotamers were sampled for each state and accessible surface areas for each residue were weighted based on MSM equilibrium populations.

4.5.7 Sequence conservation

Sequence conservation of spike proteins was calculated using the Uniprot database.⁸³ Sequences between 30% - 90% were pulled and aligned with the Muscle algorithm.⁸⁴ The entropy at each position was calculated to quantify variability of amino acids. Conservation was defined as one minus the entropy.

4.5.8 Data availability

The datasets generated and/or analyzed during the current study are available at https://covid.molssi.org/ and https://csf.io/fs2yv/.

4.5.9 Code availability

GROMACS (https://github.com/gromacs/gromacs), OpenMM (https://github.com/openmm/openmm), our FAST adaptive sampling method (https://github.com/bowman-lab/fast), mdtraj (https://github.com/mdtraj/mdtraj), and our enspara code (https://github.com/bowman-lab/enspara) are all open source.

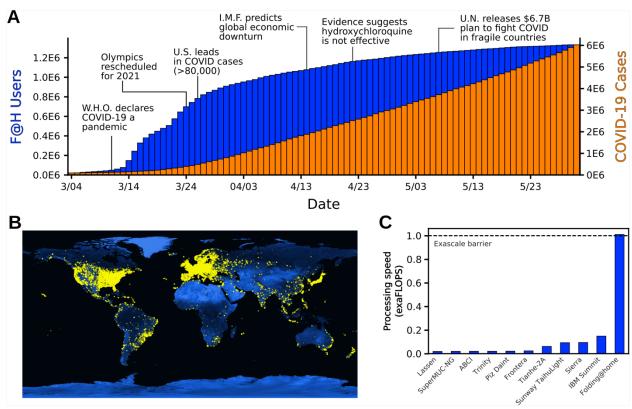


Figure 4.1 Summary of Folding@home's computational power. A) The growth of Folding@home (F@H) in response to COVID-19. The cumulative number of users is shown in blue and COVID-19 cases are shown in orange. **B)** Global distribution of Folding@home users. Each yellow dot represents a unique IP address contributing to Folding@home. **C)** The processing speed of Folding@home and the next 10 fastest supercomputers, in exaFLOPS (one exaFLOPS is 10¹⁸ floating point operations per second).

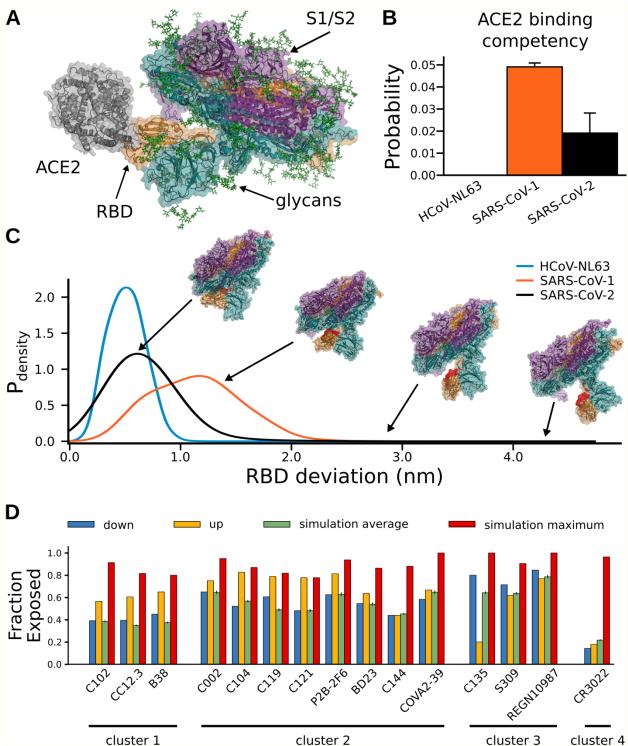


Figure 4.2 Structural characterization of Spike opening and conformational masking for three Spike homologues.

A) An example structure of SARS-CoV-2 Spike protein from our simulations that is fully compatible with receptor binding, as shown by superimposing ACE2 (gray). The three chains of Spike are illustrated with a cartoon and transparent surface representation (orange, teal, and purple), and glycans are shown as sticks (green). **B)** Three Spike homologues have very different probabilities of adopting ACE2 binding competent conformations, likely modulating their affinities for both ACE2 and antibodies that engage the ACE2-binding interface. HCoV-NL63, SARS-CoV-1, and SARS-CoV-2 are shown as light-blue, orange, and black, respectively. **C)** The probability

distribution of Spike opening for each homologue. Opening is quantified in terms of how far the center of mass of an RBD deviates from its position in the closed (or down) state. The cryptic epitope for the antibody CR3022 (red) is only accessible to antibody binding in extremely open conformations. **D)** Our simulations capture exposure of cryptic epitopes that are buried in the up and down cryoEM structures. The fraction of residues within different epitopes that are exposed to a 0.5 nm radius probe for the down structure (blue), up structure (yellow), the ensemble average from our simulations (green), and the maximum value we observe in our simulations (red). Epitopes are determined as the residues that contact the specified antibody and are clustered by their binding location on the RBD. ¹³

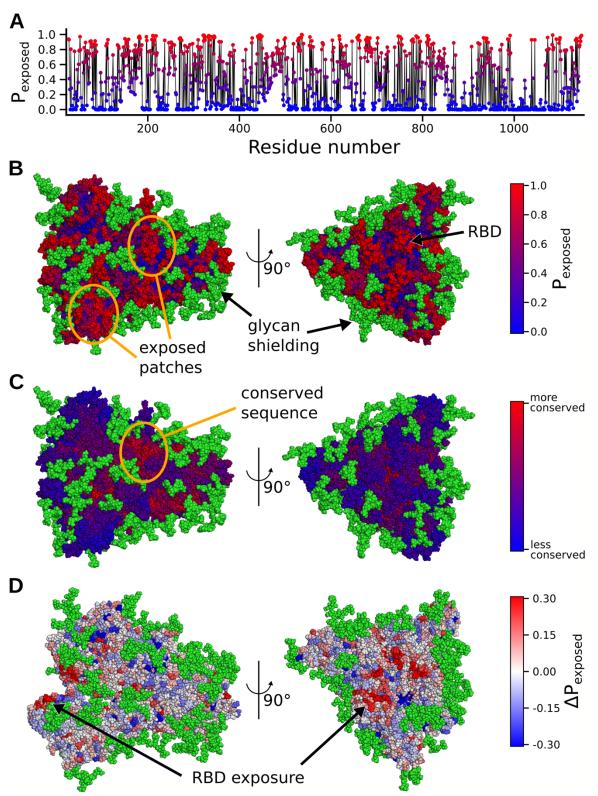


Figure 4.3 Effects of glycan shielding and conformational masking on the accessibility of different parts of the Spike to potential therapeutics.

A) The probability that a residue is exposed to potential therapeutics, as determined from our structural ensemble. Red indicates a high probability of being exposed and blue indicates a low probability of being exposed. **B)** Exposure probabilities colored on the surface of the Spike protein. Exposed patches are circled in orange. Red

residues have a higher probability of being exposed, whereas blue residues have a lower probability of being exposed. Green atoms denote glycans. **C)** Sequence conservation score colored onto the Spike protein. A conserved patch on the protein is circled in orange. Red residues have higher conservation, whereas blue residues have lower conservation. **D)** The difference in the probability that each residue is exposed between the ACE2-binding competent conformations and the entire ensemble. Red residues have a higher probability of being exposed upon opening, whereas blue residues have a lower probability of being exposed. Exposure data can be found online at https://osf.io/fs2yv/ under the SARS-CoV-2 spike project in the analysis folder.

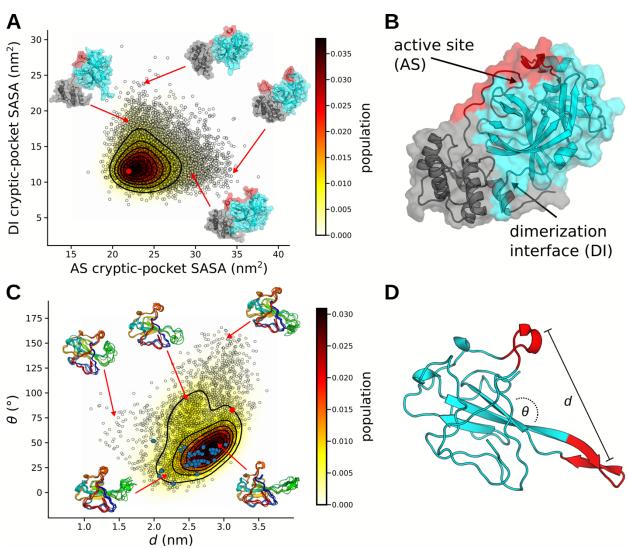


Figure 4.4 Examples of cryptic pockets and functionally-relevant dynamics.

A-B) Conformational ensemble of M^{pro} (monomeric) predict cryptic pockets near the active site (AS) and domain interface (DI). Conformational states (black circles) are projected onto the solvent accessible surface areas (SASAs) of residues surrounding either the active-site or dimerization interface. The starting structure for simulations (6Y2E) is shown as a red dot. Representative structures are depicted with cartoon and transparent surface. Domains I and II are colored cyan and domain III is colored gray. The loop of domain III, which covers the active-site residues and is seen to be highly dynamic, is colored red. **C-D)** The conformational ensemble from our simulations of nucleoprotein is similar to the distribution of structures seen experimentally. Conformational states are projected onto the distance and angle between the positive finger and a nearby loop. Angles were calculated between vectors that point along each red segment in panel D and distances were calculated between their centers of mass. Cluster centers are represented as black circles, the starting structure for simulations (6VYO) is shown as a red dot, and NMR structures are shown with solid blue dots. Representative structures are shown as cartoons.

Table 4.1 Summary of protein systems we have simulated on Folding@home, organized by viral strain.

^{***}Missing residues were modeled using CHARMM-GUI. 65,66

System name	Oligomerization	Initial structure	Residues	Atoms in system	Aggregate simulation time (µs)	Cryptic pockets discovered
SARS-CoV-2						
NSP3 (Macrodomain "X")	Monomer	6W02	167	23907	10,906	-
NSP3 (Papain-like protease 2, PL2 ^{pro})	Monomer	3E9S**	306	97285	731	2
NSP5 (main protease, M ^{pro} , 3CL ^{pro})	Monomer	6Y2E	306	64791	6,405	2
NSP5 (main protease, M ^{pro} , 3CL ^{pro})	Dimer	6Y2E	612	77331	2,902	2
NSP7	Monomer	5F22**	79	20094	3,722	3
NSP8	Monomer	2AHM**	191	156282	1,776	3
NSP9	Dimer	6W4B*	226	49885	8,939	2
NSP10	Monomer	6W4H*	131	29560	6,141	2
NSP12 (polymerase)	Monomer	6NUR**	891	186622	3,330	3
NSP13 (helicase)	Monomer	6JYT**	596	129368	3,407	3
NSP14	Monomer	5C8S**	527	216380	2,384	2
NSP15	Monomer	6VWW	347	67345	3,674	4
NSP15	Hexamer	6VWW	2082	230339	4,270	-
NSP16	Monomer	6W4H*	298	45672	2,382	5
Nucleoprotein (RBD)	Monomer	6VYO	173	29125	9,493	3
Nucleoprotein Dimerization Domain	Monomer	6YUN*	118	34905	6,782	-
Nucleoprotein Dimerization Domain	Dimer	6YUN*	236	72733	1,458	2
Spike	Trimer	6VXX***	3363	442881	1,109	-
NSP7 / NSP8 / NSP12	Trimer complex	6NUR**	1184	215694	1,686	-
NSP10 / NSP14	Dimer complex	5C8S**	688	226672	689	3
NSP10 / NSP16	Dimer complex	6W4H*	429	63752	3,463	2
SARS-CoV-1						
NSP3 (Macrodomain "X")	Monomer	2FAV	172	33117	659	-
NSP9	Dimer	1QZ8*	226	49599	7,763	-
NSP15	Monomer	2H85	345	67345	4,734	-
NSP15	Hexamer	2H85	2070	230339	1,130	-
Nucleoprotein RBD	Monomer	2OFZ	174	29125	4,088	-
Nucleoprotein Dimerization Domain	Monomer	2GIB	370	34905	1,626	-
Nucleoprotein Dimerization Domain	Dimer	2GIB	740	72733	4,221	-
Spike	Trimer	5X58***	3261	375851	741	-
NSP10 / NSP16	Dimer complex	6W4H**	425	69589	518	-
Human						
IL6	Monomer	1ALU	166	26855	1,593	2
IL6-R	Monomer	1N26	299	149764	196	5
ACE2	Monomer	6LZG	596	75787	664	2

^{*}Missing residues were modeled using Swiss model.⁶⁴

^{**}Structural model was generated from a homologous sequence using Swiss model.⁶⁴

MERS							
NSP13	Monomer	5WWP	596	121134	719	-	
NSP10 / NSP16	Dimer Complex	6W4H**	424	69127	518	-	
HCoV-NL63							
Spike	Trimer	5SZS***	3606	453348	651	-	

Bibliography

- 1. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
- 2. Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine* **27**, (2020).
- 3. Sorci, G., Faivre, B. & Morand, S. Why Does COVID-19 Case Fatality Rate Vary Among Countries? *SSRN Electronic Journal* (2020) doi:10.2139/ssrn.3576892.
- 4. Khafaie, F. R. M. A. Cross-Country Comparison of Case Fatality Rates of COVID-19/SARS-COV-2. *Osong Public Health and Research Perspectives* **11**, 74–80 (2020).
- 5. Mahase, E. Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. *BMJ* **368**, m641 (2020).
- 6. Onder, G., Rezza, G. & Brusaferro, S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA* **323**, 1775–1776 (2020).
- 7. Ferreira, L. G., Santos, R. N. D., Oliva, G. & Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **20**, 13384–13421 (2015).
- 8. Yuan, M. *et al.* A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* **368**, 630–633 (2020).
- 9. Zhou, T. *et al.* A pH-dependent switch mediates conformational masking of SARS-CoV-2 spike. *bioRxiv* **16**, 2020.07.04.187989 (2020).
- 10. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
- 11. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412 (2020).
- 12. Lu, M. *et al.* Real-time Conformational Dynamics of SARS-CoV-2 Spikes on Virus Particles. *bioRxiv* **581**, 2020.09.10.286948 (2020).
- 13. Barnes, C. O. *et al.* Structural classification of neutralizing antibodies against the SARS-CoV-2 spike receptor-binding domain suggests vaccine and therapeutic strategies. *bioRxiv* **584**, 2020.08.30.273920 (2020).
- 14. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

- 15. Benton, D. J. *et al.* Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* **588**, 327–330 (2020).
- 16. Cai, Y. et al. Distinct conformational states of SARS-CoV-2 spike protein. Science **369**, 1586–1592 (2020).
- 17. Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1–39). *Journal of the American Chemical Society* **132**, 1526–1528 (2010).
- 18. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **334**, 517–520 (2011).
- 19. Stodden, V. Enabling Reproducible Research: Open Licensing for Scientific Innovation. (2009).
- 20. Amaro, R. E. & Mulholland, A. J. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *Journal of Chemical Information and Modeling* **60**, 2653–2656 (2020).
- 21. Shirts, M. & Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **290**, 1903–1904 (2000).
- 22. Kohlhoff, K. J. *et al.* Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature Chemistry* **6**, 15–21 (2014).
- 23. Shukla, D., Meng, Y., Roux, B. & Pande, V. S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature Communications* **5**, 1–11 (2014).
- 24. Sun, X., Singh, S., Blumer, K. J. & Bowman, G. R. Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. *eLife* 7, 19 (2018).
- 25. Hart, K. M., Ho, C. M. W., Dutta, S., Gross, M. L. & Bowman, G. R. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nature Communications* 7, 1–10 (2016).
- 26. Chen, S. *et al.* The dynamic conformational landscape of the protein methyltransferase SETD8. *eLife* **8**, 213 (2019).
- 27. Porter, J. R., Meller, A., Zimmerman, M. I., Greenberg, M. J. & Bowman, G. R. Conformational distributions of isolated myosin motor domains encode their mechanochemical properties. *eLife* **9**, 19 (2020).
- 28. Hart, K. M. *et al.* Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLOS ONE* **12**, e0178678 (2017).

- 29. Cruz, M. A. *et al.* Discovery of a cryptic allosteric site in Ebola's 'undruggable' VP35 protein using simulations and experiments. *bioRxiv* 17, 2020.02.09.940510 (2020).
- 30. Wang, W. *et al.* The Cap-Snatching SFTSV Endonuclease Domain Is an Antiviral Target. *Cell Reports* **30**, 153-163.e5 (2020).
- 31. Kirchdoerfer, R. N. *et al.* Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Scientific Reports* **8**, 1–11 (2018).
- 32. Zhang, H., Penninger, J. M., Li, Y., Zhong, N. & Slutsky, A. S. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Medicine* **46**, 586–590 (2020).
- 33. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
- 34. Watanabe, Y. *et al.* Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nature Communications* **11**, 1–10 (2020).
- 35. Casalino, L. *et al.* Shielding and Beyond: The Roles of Glycans in SARS-CoV-2 Spike Protein. **9**, 221–27 (2020).
- 36. Zimmerman, M. I. & Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *Journal of Chemical Theory and Computation* **11**, 5747–5757 (2015).
- 37. Zimmerman, M. I. & Bowman, G. R. How to Run FAST Simulations. *Methods in Enzymology* **578**, 213–225 (2016).
- 38. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52**, 99–105 (2010).
- 39. Wang, X., Unarta, I. C., Cheung, P. P.-H. & Huang, X. Elucidating molecular mechanisms of functional conformational changes of proteins via Markov state models. *Curr Opin Struc Biol* **67**, 69–77 (2021).
- 40. Turoňová, B. *et al.* In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science* eabd5223 (2020) doi:10.1126/science.abd5223.
- 41. Sikora, M. *et al.* Map of SARS-CoV-2 spike epitopes not shielded by glycans. *bioRxiv* 2020.07.03.186825 (2020) doi:10.1101/2020.07.03.186825.
- 42. Shang, J. et al. Cell entry mechanisms of SARS-CoV-2. Proceedings of the National Academy of Sciences 117, 11727–11734 (2020).

- 43. Yuan, Y. *et al.* Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nature Communications* **8**, 1–9 (2017).
- 44. Guo, L. *et al.* Engineered Trimeric ACE2 Binds and Locks "Three-up" Spike Protein to Potently Inhibit SARS-CoVs and Mutants. *bioRxiv* 2020.08.31.274704 (2020) doi:10.1101/2020.08.31.274704.
- 45. Huo, J. *et al.* Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *SSRN Electronic Journal* (2020) doi:10.2139/ssrn.3613273.
- 46. Zhong, N. S. *et al.* Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet* **362**, 1353–1358 (2003).
- 47. Hoek, L. van der *et al.* Identification of a new human coronavirus. *Nature Medicine* **10**, 368–373 (2004).
- 48. Wu, K., Li, W., Peng, G. & Li, F. Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *Proceedings of the National Academy of Sciences* **106**, 19970–19974 (2009).
- 49. Shang, J. et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
- 50. Graham, B. S., of, M. G. A. review & 2019. Structure-based vaccine antigen design. *annualreviews.org*.
- 51. Li, Y. *et al.* Linear epitopes of SARS-CoV-2 spike protein elicit neutralizing antibodies in COVID-19 patients. *medRxiv* 2020.06.07.20125096 (2020) doi:10.1101/2020.06.07.20125096.
- 52. Brouwer, P. J. M. *et al.* Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science* **38**, eabc5902 (2020).
- 53. Hansen, J. *et al.* Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* eabd0827 (2020) doi:10.1126/science.abd0827.
- 54. Graziano, V., McGrath, W. J., Yang, and L. & Mangel, W. F. *SARS CoV Main Proteinase: The Monomer–Dimer Equilibrium Dissociation Constant.* vol. 45 (American Chemical Society, 2006).
- 55. Goyal, B. & Goyal, D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Combinatorial Science* **22**, 297–305 (2020).
- 56. Porter, J. R. *et al.* Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. *Biophysical Journal* **116**, 818–830 (2019).

- 57. McBride, R., Zyl, M. V. & Fielding, B. C. The Coronavirus Nucleocapsid Is a Multifunctional Protein. *Viruses* **6**, 2991–3018 (2014).
- 58. Masters, P. S. Coronavirus genomic RNA packaging. Virology 537, 198–207 (2019).
- 59. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *bioRxiv* **53**, 171–39 (2020).
- 60. Su, Z. *et al.* Electron Cryo-microscopy Structure of Ebola Virus Nucleoprotein Reveals a Mechanism for Nucleocapsid-like Assembly. *Cell* **172**, 966-978.e12 (2018).
- 61. Dinesh, D. C., Chalupska, D., Silhan, J., Veverka, V. & Boura, E. Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. **73**, 213–13 (2020).
- 62. Kucherova, A., Strango, S., Sukenik, S. & Theillard, M. Modeling the Opening SARS-CoV-2 Spike: an Investigation of its Dynamic Electro-Geometric Properties. *Biorxiv* 2020.10.29.361261 (2020) doi:10.1101/2020.10.29.361261.
- 63. Vithani, N. *et al.* SARS-CoV-2 Nsp16 activation mechanism and a cryptic pocket with pancoronavirus antiviral potential. *Biorxiv* 2020.12.10.420109 (2020) doi:10.1101/2020.12.10.420109.
- 64. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* **46**, W296–W303 (2018).
- 65. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. 1859–1865 (2008).
- 66. Lee, J. *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation* **12**, 405–413 (2016).
- 67. Chodera, J., Lee, A. A., London, N. & Delft, F. von. Crowdsourcing drug discovery for pandemics. *Nature Chemistry* **12**, 581–581 (2020).
- 68. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multilevel parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
- 69. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry* **24**, 1999–2012 (2003).
- 70. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of computational chemistry* **34**, 2135–2145 (2013).

- 71. Knoverek, C. R., Amarasinghe, G. K. & Bowman, G. R. Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities. *Trends Biochem Sci* **44**, 351–364 (2018).
- 72. Bowman, G. R. Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *J Comput Chem* **37**, 558–566 (2016).
- 73. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935 (1983).
- 74. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 116–122 (2008).
- 75. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007).
- 76. Zimmerman, M. I. *et al.* Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS central science* **3**, 1311–1321 (2017).
- 77. Hendlich, M., Rippmann, F. & Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of molecular graphics & modelling* **15**, 359-63–389 (1997).
- 78. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).
- 79. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* **140**, 2386–2396 (2018).
- 80. Porter, J. R., Zimmerman, M. I. & Bowman, G. R. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *The Journal of Chemical Physics* **150**, 044108 (2019).
- 81. Zimmerman, M. I., Porter, J. R., Sun, X., Silva, R. R. & Bowman, G. R. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *Journal of Chemical Theory and Computation* **14**, 5459–5475 (2018).
- 82. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**, 1528–1532 (2015).
- 83. Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506–D515 (2019).

84. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–1797 (2004).

Chapter 5

SARS-CoV2 Nsp16 activation mechanism and a cryptic pocket with pan-coronavirus antiviral potential

5.1 Preamble

This chapter is adapted from the following article: Vithani N.,* Ward, M.D.,* Zimmerman, M.I., Novak, B., Borowsky, J.H., Singh S., and Bowman, G.R (2021). "SARS-CoV2 Nsp16 activation mechanism and a cryptic pocket with pan-coronavirus antiviral potential", *Biophysical Journal*. 120, 14, 2880-2889.

5.2 Introduction

With the coronavirus 2019 (COVID-19) pandemic ravaging communities across the globe there is a massive ongoing effort to understand the molecular machinery of coronaviruses, which may provide insight into therapeutic opportunities (1–3). The severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) virus responsible for COVID-19 disease has infected over sixty million and killed over 1.5 million people globally to date (4). Additionally, coronaviruses have caused several past epidemics including severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) which had fatality rates of ~10% and ~34%, respectively (5, 6). Therefore, there is likely to be evolution and outbreaks of additional zoonotic coronaviruses in the future (7). While vaccine trials for COVID-19 are successfully wrapping up, there are still no approved antivirals that reduce mortality to coronavirus infections (8–10). Taken together, there is strong incentive to understand the fundamental mechanisms of how these coronaviruses

^{*}These authors contributed equally to the work

operate in hopes of discovering effective therapeutics. Biophysical studies can provide these details, and a tremendous amount of biophysical work has already been done to understand the virus' twenty-nine proteins. So far, the spike protein, positioned on the outside of the viral envelope, has proven to be a good vaccine candidate (11). Beyond the spike, the sixteen "nonstructural" (i.e. accessory) proteins carry out the majority of the virus' essential processes, making them good targets for antiviral therapeutics (12, 13).

Among the nonstructural proteins (Nsp's), Nsp16 is particularly important to the viral replication cycle as it is essential to coronavirus' immune evasion (14–16). Nsp16 is a 2'-O-Methyltransferase (2'-O-MTase) that forms part of the replication-transcription complex (17). It mimics the human protein Cap-specific mRNA (nucleoside-2'-O-)-methyltransferase (CMTr1) to perform a crucial step in capping transcribed mRNA (18). Specifically, Nsp16 facilitates the transfer of a methyl group from its S-adenosylmethionine (SAM) cofactor to the 2' hydroxyl of ribose sugar of viral mRNA (18, 19). This methylation both improves translation efficiency and camouflages the mRNA so that it is not recognized by intracellular pathogen recognition receptors, such as IFIT and RIG-I (15, 20). Importantly, inhibiting or knocking out 2'-O-MTase activity severely attenuates viral replication and infectivity of coronaviruses (13, 20). Thus, developing small molecules inhibitors of Nsp16 is a promising therapeutic strategy.

Interestingly, while all other 2'-O-MTases (eukaryotic and viral) are active as monomers, Nsp16 requires a binding partner, Nsp10, to be active (16–18, 21–23). In fact, Nsp16 does not even bind its ligands (SAM and RNA) in the absence of Nsp10. In the experimentally-derived structures of the Nsp16/Nsp10 complex, Nsp10 does not form any direct interaction with either ligand (Fig. 5.1a), suggesting that Nsp10 may allosterically regulate Nsp16 to enable substrate binding (18, 19, 24–27). Given that there is significant structural variation in the RNA-binding

loops of different crystal structures of Nsp16 (Fig. 5.1b) and structures of monomeric Nsp16 have not been solved, we hypothesized that Nsp16 is highly dynamic in solution, and Nsp10 acts by stabilizing the active state. In contrast, we anticipate that human CMTr1 would be less dynamic as it doesn't require a binding partner for substrate binding and has been crystalized in its monomeric state. Often, dynamics of proteins reveal allosteric pockets that remain hidden in their crystal structures (i.e., cryptic pockets). If monomeric Nsp16 is more dynamic than CMTr1, it may adopt inactive configurations that reveal allosteric cryptic pockets, which can be targeted by small-molecule inhibitors for its selective inhibition.

Here, we use computer simulations to understand the activation mechanism of Nsp16 and identify cryptic pockets that may be valuable antiviral targets. Active site inhibitors, such as Sinefungin, have been shown to outcompete SAM binding and render Nsp16 catalytically inactive (28, 29). However, there are more than 200 human proteins with known or putative methyltransferase activity that use SAM as a cofactor (30). Therefore, it may be difficult to design antivirals that target the SAM (or RNA) binding sites of Nsp16 without eliciting off-target effects by also binding human methyltransferases. For example, Sinefungin has been shown to occupy the SAM-binding pocket of human N7 methyltransferase in a crystal structure (PDB: 3epp). Targeting the Nsp16/Nsp10 interface could be an alternative means to selectively inhibit Nsp16 since CMTr1 lacks a homologous binding partner. Towards this, peptide-based inhibitors that mimic Nsp10 to compete for interactions at the Nsp10/Nsp16 interface have been shown to inhibit Nsp16 activity (31, 32). While this approach seems promising, peptide-based inhibitors face challenges including limited stability and shelf-life, the possibility of adverse immunogenic responses, and the high cost of production (33). To expand the therapeutic opportunities, we search for other ways to inactivate Nsp16. First, we compare the structure and dynamics of

SARS-CoV2 Nsp16 in the presence and absence of Nsp10 to understand Nsp16's activation. Specifically, we use over one millisecond of molecular dynamics simulation data (2) to characterize how Nsp10 binding shifts Nsp16's conformational ensemble to activate Nsp16. After showing that the resulting model is consistent with a variety of experimental observations, we use it to hunt for cryptic pockets that may provide a means to inhibit Nsp16. Finally, we extend our simulations to SARS-CoV1, MERS, and human CMTr1 to determine if targeting such a pocket could provide an opportunity to develop pan-coronavirus antivirals.

5.3 Methods

5.3.1 System Preparation

The systems were prepared starting from crystal structures 6w4h, 3r24, 5ynf and 4n49, for SARS-CoV2, SARS-CoV1, MERS, and CMTr1, respectively. All ligands, solutes, and water molecules from the crystal structures were removed. For monomeric Nsp16 simulations, Nsp10 was also removed. In the coronavirus homologs, two zinc ions were retained, and the coordinating residues were modified accordingly (CYS->CYM and HIS->HID). Missing residues in the crystal structure of CMTr1 were modeled using the Modeller package (34). All systems were solvated in TIP3P water (35) in a rhombic dodecahedral box with periodic boundary conditions and Na⁺ and Cl⁻ ions added to neutralize the system. Then, systems were energy minimized with a steepest descent algorithm until the maximum force fell below 100 kJ/mol/nm using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor list, Coulomb interactions, and van der Waals interactions.

Systems were equilibrated for 1.0 ns in *NPT* simulations, with all bonds constrained using the LINCS algorithm (36) and virtual sites were used to allow a 4 fs time step. Cutoffs of 1.1 nm were used for the neighbor list with 0.9 for Coulomb and van der Waals interactions. The

particle mesh Ewald method (37) was employed for treatment of long-range interactions with a Fourier spacing of 0.12 nm. The Verlet cutoff scheme was used for the neighbor list. Berendsen barostat was used to control the pressure during the equilibration.(38) The stochastic velocity rescaling (v-rescale) thermostat was used to control the temperature at 300 K (39).

5.3.2 Adaptive sampling simulations

The FAST algorithm (40, 41) was employed for all four homologs for a total of five FAST simulations (SARS-CoV2 FAST simulations were performed on both monomeric Nsp16 and the Nsp10/Nsp16 complex). FAST was used here to generally enhance conformational sampling and also to quickly explore cryptic pockets. The procedure for FAST simulations is as follows: 1) run initial simulations, 2) build MSM, 3) rank states based on FAST ranking, 4) restart simulations from the top ranked states, 5) repeat steps 2-4 until ranking is optimized. For each system, MSMs were generated after each round of sampling using a k-centers clustering algorithm based on the RMSD between select atoms. Clustering continued until the maximum distance of a frame to a cluster center fell within a predefined cutoff. In addition to the FAST ranking, a similarity penalty was added to promote conformational diversity in starting structures, as has been described previously (42).

For SARS-CoV2 monomeric Nsp16 and Nsp16/Nsp10, the simulation data was generated in a previous manuscript published by our group. Briefly, FAST-pocket simulations were run at 300 K for 6 rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 µs aggregate simulation for each system). The FAST-pocket ranking function favored restarting simulations from states with large pocket openings. Pocket volumes were calculated using the LIGSITE algorithm (43). From these simulations, a conformationally diverse set of structures was selected to be run on Folding@home based on the k-centers

clustering algorithm mentioned above. A total of 283 microseconds and 770 microseconds of aggregate simulation time was collected for the Nsp10/Nsp16 complex and monomeric Nsp16, respectively.

FAST-distance simulations were used for SARS-CoV1 Nsp16, MERS Nsp16, and CMTr1 to sample the β3-β4 pocket identified from SARS-CoV2 simulations. FAST-distance simulations were run at 300 K for 15 rounds, with 10 simulations per round, where each simulation was 40 ns in length (6.0 μs aggregate simulation for each system). The FAST-distance ranking favored stated with greater distances between the alpha carbons of β3 and β4.

5.3.3 DiffNets

We used DiffNets, a deep learning-based dimensionality reduction algorithm developed by our group, to highlight biochemically relevant differences between datasets. (44) We trained a DiffNet to compare and contrast structure ensembles of monomeric Nsp16 and the Nsp16/Nsp10 complex to find features that discriminate them, highlighting the structural determinants of Nsp16 activation. First, we subsampled the data by a factor of 25 and 68 for the Nsp16/Nsp10 complex and monomeric Nsp16 data, respectively to have an equal amount of data. Then, we converted simulation data to DiffNet input following the data normalization procedure from the original manuscript. Briefly, XYZ atom coordinates from simulations were mean-shifted to zero, and then multiplied by the inverse of the square root of a covariance matrix, which was calculated from simulations. For all DiffNet training and analysis, we used a split architecture (as described previously) where the classification task was focused on all atoms within 1nm of SAM or RNA-cap based on 6wks crystal structure. This atom selection was chosen to guide DiffNets to find differences in the active site region of Nsp16, which is inherently linked to its activation.

For training, simulation frames are classified as "Nsp16 inactive" or "Nsp16 active" based on initial classification labels of 0 (i.e. Nsp16 inactive) for all monomeric Nsp16 frames, and labels of 1 (i.e. Nsp16 active) for all frames from the Nsp10/Nsp16 complex. These labels were iteratively updated in a self-supervised manner described in the original manuscript where we choose expectation maximization bounds of [0.1-0.4] for monomeric Nsp16 and [0.6-0.9] for the Nsp10/Nsp16 complex. This allows for more coherent classification labels as monomeric Nsp16 may sometimes adopt structural poses associated with Nsp16 activation and vice-versa for the Nsp10-Nsp16 complex. Additionally, we used 30 latent variables, 10 training epochs where we subsampled the data by a factor of 10 in each epoch, a batch size of 32, and a learning rate of 0.0001.

To analyze the DiffNet output, we calculated 10 representative structures that span from "Nsp16 inactive" states to "Nsp16 active" states (i.e. structures with classification labels spanning 0 to 1). After training, the DiffNet learns a low-dimensional representation of each simulation frame (i.e. a latent vector) and outputs a classification label for every simulation frame. We binned the structures into 10 equally spaced bins based on their classification labels, which span from 0-1. The, we calculated the mean latent vector for each bin and used the DiffNet to reconstruct a structure based on each latent vector. These structures were used as representative structures for each bin. All training and analysis were performed using the open-source package https://github.com/bowman-lab/diffnets.

5.3.4 Markov State Models

A Markov State Model (MSM) is a statistical framework for analyzing molecular dynamics simulations that provides a network representation of a free energy landscape. (45–47) To quantify cryptic pocket opening across the homologs and changes between monomeric Nsp16

and the Nsp10/Nsp16 complex, we performed several measurements that rely on MSMs built based on the simulation data. We built a separate MSM for each system using all simulation data available for that system. All MSMs were constructed with the Enspara python package (48). First, the solvent accessible surface area (SASA) of each residue side-chain was calculated using the Shrake-Rupley algorithm (49) implemented in MDTraj (50) using a drug-sized probe (2.8 Å sphere).

Then, we clustered the data using a hybrid clustering algorithm. First, we used a k-centers algorithm (51, 52) to cluster the data. Next, we applied sweeps of k-medoids update steps (3 for SARS-CoV2 data, 2 for other homologs) which refined the cluster centers to be in the densest regions of conformational space (53). We clustered the simulation data based on the residuelevel SASA. For SARS-CoV2 Nsp16 and Nsp10/Nsp16 complex for which we had massive datasets from Folding@home (283 microseconds and 770 microseconds), we used 5000 cluster centers. For SARS-CoV1 Nsp16, MERS Nsp16 and human CMTr1 (6 microseconds of FAST adaptive sampling data per system), we used 1500 cluster centers. We validated that these produced Markovian models by plotting the implied timescales, from which we chose a lag time of 5 ns (see appendix, Fig. A.4.1). To further ensure robustness of the MSMs, we also built models based on alternative clusterings and confirmed that they gave similar results. Specifically, for SARS-CoV2 Nsp16 and Nsp10/Nsp16 complex we built MSMs using (1) 5.2 nm² cluster-radius cut-off and (2) 5.5 nm² cluster-radius. For SARS-CoV1 Nsp16, MERS Nsp16 and human CMTr1 we built MSMs using (1) 4.0 nm² cluster-radius and (2) 4.5 nm² clusterradius. All MSMs were Markovian (see Fig. A.4.1). Moreover, we used these MSMs to recreate the distributions in Fig's 5.2-5.4 and we find that the results are robust across all MSMs (see Fig. A.4.2, A.4.3, A.4.4). A Markov time of 5 ns was selected for based on the implied timescales to

build a Markov state model (MSM) for each homolog. To build the MSMs, transition probability matrices were produced by counting transitions between states (i.e. clusters), adding a prior count of $\frac{1}{N_{states}}$ and row-normalizing, as is described previously (54). Equilibrium populations were calculated as the eigenvector of the transition probability matrix with an eigenvalue of one. For all histograms shown, we calculated the order parameter of distance (e.g. distance between $\beta 3-\beta 4$) using cluster centers (i.e. representative structure of the cluster) and weighted the order parameter by the corresponding equilibrium population calculated with the MSM. We also resampled the equilibrium populations 100 times by bootstrapping the MSM, which provided error bars for computing the fraction of SAM and RNA compatible states adopted by monomeric Nsp16 and the Nsp16/10 complex.

5.3.5 Distance and SASA calculations

Figures 5.2, 5.3, and 5.4 include distance and SASA measurements that are explained in more detail here. In Figure 5.2 we measure the distance between gate loop 1 and gate loop 2 as the distance between Gln28 and Lys141 since these residues are known to undergo significant changes for RNA binding. We measure the distance between SAM binding loop 2 and gate loop 2 as the average distance between (Met131, Tyr132, Asp133, Pro134) and (Asp99, Leu100, Asn101, Asp102) as these are key residues that cradle SAM in the bound state. All SASA measurements are performed using Ala79, Thr82, Ala83, Leu86, Thr93, Leu94, Leu95, Val96, Asp97, Ala98 and Asp99 as this is the main component that gets exposed during cryptic pocket opening.

5.3.6 Cryptic pocket detection

Cryptic pockets in SARS-CoV2 Nsp16 were identified using our previously established approach called Exposons analysis (55). This analysis was performed using the cluster centers and the equilibrium probabilities derived from the MSMs built on the residue level SASA described above. The center of each cluster was taken as an exemplar of that conformational state, and residues were classified as exposed if their SASA exceeded 2.0 Ų and buried otherwise. The mutual information between the exposure/burial of each residue-pair was calculated based on the MSM, by treating the SASA values in the cluster centers as samples and weighting them by the equilibrium probability of the representative state. The mutual information was computed using the following equation:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Finally, cryptic pockets (Exposons) were identified as groups of residues undergoing cooperative change in SASA, by clustering the matrix of pairwise mutual information using affinity propagation.

The β3-β4 cryptic pocket identified in SARS-CoV2 Nsp16 consists of residues Ala79, Thr82, Ala83, Leu86, Thr93, Leu94, Leu95, Val96, Asp97, Ala98 and Asp99. Total SASA of these residues/homologous residues was measured for detecting cryptic pocket opening in all homologs of Nsp16 (SARS-CoV2, SARS-CoV1 and MERS). For measuring equivalent cryptic pocket in CMTR1, total SASA of structurally homologous residues (Gly141, Ser144, Glu145, Val148, Ala155, Lys156, Gly157, His158, Gly159, Met160, Thr161) was calculated.

5.3.7 Sequence Conservation

135

Protein sequences of Nsp16 from SARS-CoV2 (YP_009725311.1), SARS-CoV1 (Uniprot ID: P0C6X7), MERS (Uniprot ID: K0BWD0), NL63 (AFD64750.1), HKU1 (YP_460023.1), Turkey CoV (YP_001941189.1), Bat CoV (YP_008439226.1), Murine hepatitis virus (YP_209243.1) were used for multiple sequence alignment. Sequences alignment was performed on Clustal Omega server (56). Sequence alignment was visualized, and the sequence conservation score was generated using Jalview 2 software (57).

For sequence comparison of SARS-CoV2, SARS-CoV1, MERS and human CMTr1 shown in Fig. 5.4, structure-based sequence alignment was performed using UCSF Chimera package (58). For the structure-based sequence alignment, we first aligned the structures of these homologs (PDB: 6wks (SARS-CoV2), 3r24 (SARS-CoV1), 5ynf (MERS) and 4n49 (CMTr1). Then, the sequences were aligned based on the structural alignment of the backbone atoms.

5.4 Results

5.4.1 Nsp10 promotes opening of Nsp16's SAM- and RNA-binding pockets

While experimental studies have demonstrated that Nsp16 requires Nsp10 to be functionally active, the structural determinants of Nsp16's activation remain unknown (17, 18, 23). Chen et. al. proposed that Nsp10's stimulatory effects are rooted in its ability to assist Nsp16 in binding SAM and RNA, which is supported by data showing that Nsp16 alone cannot bind SAM or RNA (18). They also propose that Nsp10 manages this by stabilizing or changing the conformation of the SAM binding pocket based on the fact that Nsp10 contacts SAM binding loops in their crystal structure (and numerous other structures). However, without assessing Nsp10-Nsp16 complex's dynamics and comparing it to monomeric Nsp16, this hypothesis is left wanting. It has also been proposed that Nsp10 assists in RNA binding by directly contacting RNA (59). However, a recent crystal structure with RNA bound (PDB: 7jyy) contains a stretch

of nucleotides long enough to contact Nsp10, but the RNA curls off into solution instead of interacting with Nsp10. Another recent study compared an RNA and SAM bound Nsp10/16 complex structure to one with only SAM bound and found a major opening of RNA binding gate loops suggesting that the dynamics of these loops might be important for Nsp16 activation (25). However, it is not clear if Nsp10 plays a role in those dynamics. Altogether, there is strong evidence that Nsp10 modulates Nsp16's structure and dynamics to assist it in binding SAM and RNA, but the mechanism of these structural changes is unclear.

To explore how Nsp10 activates Nsp16, we analyzed simulations of Nsp16 in the presence and absence of Nsp10 using DiffNets. Recently, our group combined the sampling powers of the FAST-pockets adaptive sampling algorithm (40) and the computational resources of Folding@home to accumulate more than one millisecond of simulation data between simulations of monomeric Nsp16 and the Nsp16/Nsp10 complex (see methods) (2). Here, we compare these simulations using a deep learning-based dimensionality reduction algorithm called DiffNets (44). DiffNets has been shown to accurately capture the structural determinants of biochemical differences between protein variants. While we are not considering protein variants, our problem is similar since Nsp16 has different biochemical properties when in the presence/absence of Nsp10 (i.e. active/inactive). Therefore, we trained a DiffNet to learn the structural determinants of Nsp16 activation by learning differences between Nsp16's ensemble when in the presence and absence of Nsp10. For each simulation frame, the DiffNet learns a low dimensional projection of the protein structure and classifies the structure with a label between 0 and 1 that indicates the likelihood that the structure is associated with Nsp16 being active.

Analysis of the DiffNet suggests that Nsp10 shifts Nsp16's conformational ensemble to stabilize more open SAM- and RNA-binding pockets. Using the DiffNet classification labels, we

identified ten structures that are representative of the progression from Nsp16 inactive states to active states (see methods and Fig. 5.2). We noticed that RNA gate loop 2 moves away from RNA gate loop 1, making for a more open RNA binding pocket in active states compared to inactive states (Fig. 5.2A). Additionally, the SAM-binding pocket also opens up in the active states relative to the inactive states. RNA-binding gate loop 2 and SAM-binding loop 2 move away from each other in the active state, which widens the pocket creating space for SAM. (Fig. 5.2A). Strikingly, the structure associated with the highest label (i.e. most strongly associated with Nsp16 activation) matches well to a recently solved crystal structure that is bound to both RNA and SAM (Fig. 5.2B) (25). Specifically, when we align the predicted active structure to 6WKS, then measure the root-mean squared deviation (RMSD) of gate loop 2, we find a deviation on par with the typical resolution of crystal structures (1.40 Å). When we perform this calculation for the predicted inactive structure, the RMSD is much higher (3.24 Å). The predicted inactive structure adopts a more collapsed gate loop 2, similar to known structures with SAM, but not RNA, bound (i.e. PDB: 6w4h & 7c2i – see Fig. 5.2B) (26). This result implies that the DiffNet learned that Nsp10 activates Nsp16, in part, by rearranging the RNA gate loop into an RNA binding competent pose. Though it is known that this RNA gate loop needs to open to bind RNA, this is the first evidence, to our knowledge, to suggest that Nsp10 may activate Nsp16 through increasing its propensity to form a more open RNA-binding pocket. Altogether, these results suggest that Nsp10's presence increases the propensity for both SAM- and RNA- binding pockets to be open.

To quantify the effect of Nsp10 on the SAM- and RNA-binding pockets, we built MSMs for both the complex and monomeric Nsp16. MSMs are a statistical framework for analyzing molecular dynamics simulation data that provide (among other things) a discrete map of

structural configurations, an equilibrium population value that corresponds to the proportion of time a protein spends in any given configuration, and the probability of transitioning between any pair of configurations (45). We constructed MSMs for Nsp16 simulations both in the presence and absence of Nsp10.

Our MSMs reveal that Nsp10 binding stabilizes open structures of both the SAM- and RNA-binding pockets that are competent to bind their respective substrates. We first found that the presence of Nsp10 results in a substantial reduction of flexibility in important binding components including both SAM binding loops and RNA gate loops (see Fig. A.4.5). This result is somewhat surprising since gate loop 2, which contacts both SAM and RNA, is not in direct contact with Nsp10, suggesting strong allosteric communication. Next, we calculated the distribution of distances for opening and closing of the SAM and RNA binding pockets (Fig. 5.2C,D). From these histograms it is clear that both of these binding pockets have an increased propensity to open when Nsp10 is present. We considered pockets as SAM/RNA binding competent when the distance between loops in a pocket is at least as open as in the crystal structure that binds both ligands (PDB: 6wks). From this analysis, Nsp16 adopts binding competent states with higher probability when Nsp10 is present vs when Nsp10 is absent for both SAM $(0.70 \pm 0.04 \text{ vs } 0.46 \pm 0.04)$ and RNA $(0.48 \pm 0.04 \text{ vs } 0.27 \pm 0.03)$. Altogether, our data suggest that Nsp10 aids SAM and RNA binding by preventing the collapse of SAM and RNA binding gate loops. Our analysis also provides structural snapshots of what inactive states look like, which may be useful in targeting Nsp16 with therapeutics.

5.4.2 A cryptic pocket in Nsp16 is a potential therapeutic target

A traditional approach to drug development involves molecules designed to target binding cavities observed in singular structural snapshots of a protein, but this approach often misses "cryptic" pockets that can form in proteins due to thermal fluctuations. Often times the active site of an enzyme is targeted for drug development to design an inhibitor that can outcompete substrate binding. However, active sites are often conserved among functional homologs. In the case of Nsp16, its human homolog (CMTr1) shares the same overall fold and binds the same substrates. Though there are significant sequence and structural differences in the active site, specificity may be more easily achieved by targeting a less functionally relevant region of the protein. Cryptic pockets can provide both a new target for drug development and the potential to achieve specificity. For example, cryptic pockets that remain closed and invisible in the crystal structure, but open in solution due to thermal fluctuations (55), can present unique potential binding sites due to differences in the dynamics of subsets of homologs (e.g. open in coronavirus homologs, but closed in human CMTr1). Therefore, it may be easier to achieve specificity by targeting a cryptic pocket. Importantly, the cryptic pocket must communicate with functional sites in order for it to be an effective therapeutic target. Here, we explore if Nsp16 contains any cryptic pockets that, when open, would stabilize the inactive state identified with DiffNets.

To find cryptic pockets, we applied "Exposons", an algorithm (55) that identifies residues with cooperative changes in solvent exposure, to Nsp16 simulation data. Using this method, we found that residues in the $\beta 3$ strand and $\alpha 3$ helix transition between closed states and open states (i.e. low to high solvent accessible surface area) (Figure 5.3A). Specifically, the $\beta 4$ strand curls up to form an α -helical structure, which results in surface exposure of $\beta 3$ and residues from $\alpha 3$ (Fig. 5.3A). The opening motion of $\alpha 4$ shifts the adjacent SAMBL2 against gate loop 2 to collapse the SAM binding pocket in a closed conformation (Fig. 5.3A,B). This agrees with the DiffNet prediction that the $\alpha 4$ strand moving away from $\alpha 3$ is associated with inactivation (see

Fig. A.4.6). Further, several residues forming this cryptic pocket directly contact Nsp10 in crystal structures of the Nsp16/Nsp10 complex (see Fig. A.4.7). The β 3- β 4 pocket opening displaces these Nsp10 binding residues, which could inhibit Nsp16's association with Nsp10 (see Fig. A.4.7). The Nsp16/Nsp10 binding interface has also been targeted with peptide-based inhibitor design (31, 32). While this flat surface may be amenable to peptide inhibitors, it is a challenging target for small molecules. In contrast, the concave shape of the cryptic pocket identified in this work presents a more viable target for small molecule inhibitors. Finally, we find that this open pocket structure is commonly visited as part of monomeric Nsp16's conformational ensemble, as measured with MSM equilibrium populations (Fig. 5.3B). Taken together, we propose that targeting the β 3- β 4 pocket with a small molecule could inhibit Nsp16's activity by preventing SAM binding or preventing association with Nsp10.

5.4.3 Conservation of the cryptic pocket in Nsp16 makes it a promising target for broad-spectrum inhibitors

To explore the possibility of targeting the cryptic pocket for broad-spectrum inhibition of coronaviruses, we evaluated the conservation of cryptic pocket opening in Nsp16 homologs. Ideally, a therapeutic developed to treat SARS-CoV2 would also work against other coronaviruses like MERS, SARS-CoV1, and potentially future outbreaks. Additionally, the therapeutic target should be sufficiently dissimilar from human CMTr1 such that it would not cause unwanted, off-target effects. While we identified a promising cryptic pocket in SARS-CoV2, we wanted to investigate if this pocket is specific to SARS-CoV2, or specific to coronaviruses in general, or if it is common across homologs including CMTr1.

First, we analyze cryptic pocket conservation by comparing sequence features and structural features based on the native, folded state. We find that the β3-β4 pocket residues are

100% conserved between SARS-CoV2 and SARS-CoV1 (Fig. 5.4B). Additionally, of the eleven residues that form the pocket, there are only two non-conservative mutations between SARS-CoV2 and MERS. Based on the sequence similarity, we expect that, if the cryptic pocket forms in all homologs, it may be possible to develop small-molecule therapeutics that targets all three. Further, we find substantial sequence differences between SARS-CoV2 and CMTr1. Eight out of the eleven pocket residues are non-conservative mutations relative to SARS-CoV2. Based on sequence differences alone, we reason that selective inhibition could be achieved even if the cryptic pocket is adopted by CMTr1. Moreover, the sequences and structure of SARS-CoV2. Nsp16 and human CMTr1 are sufficiently different in the β3-β4 pocket region that the human protein may not even have the cryptic pocket (see Fig. A.4.8). Based on these sequence and structural differences, combined with the lack of requirement of a stabilizing binding partner, we hypothesized that cryptic pocket opening is not likely to be conserved in CMTr1.

To explore cryptic pocket opening across homologs, we performed FAST-pocket simulations of monomeric Nsp16 for SARS-CoV1 and MERS, as well as, for human CMTr1. Then, we built an MSM for each homolog and measured opening of the β 3- β 4 pocket by measuring the equilibrium weighted solvent exposure of the pocket residues we previously used to define the pocket (i.e. Fig. 5.3). In these simulations, we find that the β 3- β 4 pocket opens with high probability in both SARS-CoV1 and MERS Nsp16 (Fig. 5.4). The timescales for transitioning between the open and closed states of the pocket are given in Supporting Information (Table A.4.1). Encouragingly, we find that the β 3- β 4 pocket has a substantially lower probability of opening in CMTr1. Taken together, features of the β 3- β 4 cryptic pocket in coronavirus homologs of Nsp16 appear sufficiently similar to each other and dissimilar to CMTr1 to make for a promising target for pan-coronavirus inhibitors.

5.5 Conclusions

Our work provides mechanistic insight into how Nsp16 is activated and reveals a new opportunity for inhibiting this essential viral component that could provide a target for pancoronavirus antivirals. First, we elucidate the activation mechanism of Nsp16 by comparing its dynamics in the presence and absence of its activator, Nsp10. Our results are consistent with previous experimental findings that Nsp16 cannot bind its substrates SAM or RNA in the absence of Nsp10 (18). We provide a structural rationale for this observation by elucidating the structural dynamics of Nsp16 in its monomeric state, which has remained inaccessible to experimental studies, and comparing it to the structural dynamics of the Nsp16/Nsp10 complex. Here, we find that Nsp10 activates Nsp16 by opening its SAM and RNA binding loops, allowing them to accommodate their respective ligands. Guided by this activation mechanism, we identify structural states of Nsp16 that are incompatible with substrate binding and also contain potential drug binding sites. Specifically, we find a pocket formed between β3 and β4 of Nsp16 that collapses the SAM binding pocket when open. The region of the pocket has overlap with where Nsp10 binds to Nsp16, so targeting this cryptic pocket could inhibit both substrate (SAM) and Nsp10 binding. Therefore, this cryptic site is a promising target for small-molecule inhibitor development. Further, we find that this cryptic pocket is conserved in MERS and SARS-CoV1 Nsp16, but not in the human homolog CMTr1, suggesting its potential for development of a pancoronavirus, broad-spectrum inhibitor that may be efficacious against COVID19 and yet unseen coronavirus outbreaks.

Bibliography

- 1. Zhou, P., X. Lou Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, H.R. Si, Y. Zhu, B. Li, C.L. Huang, H.D. Chen, J. Chen, Y. Luo, H. Guo, R. Di Jiang, M.Q. Liu, Y. Chen, X.R. Shen, X. Wang, X.S. Zheng, K. Zhao, Q.J. Chen, F. Deng, L.L. Liu, B. Yan, F.X. Zhan, Y.Y. Wang, G.F. Xiao, and Z.L. Shi. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 579: 270–273.
- 2. Zimmerman, M.I., J.R. Porter, M.D. Ward, S. Singh, N. Vithani, A. Meller, U.L. Mallimadugula, C.E. Kuhn, J.H. Borowsky, R.P. Wiewiora, M.F.D. Hurley, A.M. Harbison, C.A. Fogarty, J.E. Coffland, E. Fadda, V.A. Voel, J.D. Chodera, and G.R. Bowman. 2020. SARS-CoV-2 Simulations Go Exascale to Capture Spike Opening and Reveal Cryptic Pockets Across the Proteome. bioRxiv, doi 10.1101/2020.06.27.175430 (preprint posted Oct. 07, 2020).
- 3. Wu, A., Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, J. Sheng, L. Quan, Z. Xia, W. Tan, G. Cheng, and T. Jiang. 2020. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe. 27: 325–328.
- 4. Johns Hopkins Coronavirus Resource Center. 2020. COVID-19 Map. Johns Hopkins Coronavirus Resour. Center-Johns Hopkins Univ. Med.
- 5. Chan-Yeung, M., and R.H. Xu. 2003. SARS: Epidemiology. Respirology. 8 Suppl: S9-14.
- 6. Zaki, A.M., S. van Boheemen, T.M. Bestebroer, A.D.M.E. Osterhaus, and R.A.M. Fouchier. 2012. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. N. Engl. J. Med. 367: 1814–1820.
- 7. Chan, J.F.W., K.K.W. To, H. Tse, D.Y. Jin, and K.Y. Yuen. 2013. Interspecies transmission and emergence of novel viruses: Lessons from bats and birds. Trends Microbiol. 21: 544–55.
- 8. Belete, T.M. 2020. A review on Promising vaccine development progress for COVID-19 disease. Vacunas. 21: 121–128.
- 9. Callaway, E. 2020. COVID vaccine excitement builds as Moderna reports third positive result. Nature. 587: 337–338.
- 10. Jackson, L.A., E.J. Anderson, N.G. Rouphael, P.C. Roberts, M. Makhene, R.N. Coler, M.P. McCullough, J.D. Chappell, M.R. Denison, L.J. Stevens, A.J. Pruijssers, A. McDermott, B. Flach, N.A. Doria-Rose, K.S. Corbett, K.M. Morabito, S. O'Dell, S.D. Schmidt, P.A. Swanson, M. Padilla, J.R. Mascola, K.M. Neuzil, H. Bennett, W. Sun, E. Peters, M. Makowski, J. Albert, K. Cross, W. Buchanan, R. Pikaart-Tautges, J.E. Ledgerwood, B.S. Graham, and J.H. Beigel. 2020. An mRNA Vaccine against SARS-CoV-2 Preliminary Report. N. Engl. J. Med. 383: 1920–1931.
- 11. Huang, Y., C. Yang, X. feng Xu, W. Xu, and S. wen Liu. 2020. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol. Sin. 41: 1141–1149.
- 12. da Silva, S.J.R., C.T. Alves da Silva, R.P.G. Mendes, and L. Pena. 2020. Role of nonstructural proteins in the pathogenesis of SARS-CoV-2. J. Med. Virol. 92: 1427–1429.
- 13. Snijder, E.J., E. Decroly, and J. Ziebuhr. 2016. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. Adv. Virus Res. 96: 59–126.

- 14. Ramanathan, A., G.B. Robb, and S.H. Chan. 2016. mRNA capping: Biological functions and applications. Nucleic Acids Res. 44: 7511–7526.
- 15. Daffis, S., K.J. Szretter, J. Schriewer, J. Li, S. Youn, J. Errett, T.Y. Lin, S. Schneller, R. Zust, H. Dong, V. Thiel, G.C. Sen, V. Fensterl, W.B. Klimstra, T.C. Pierson, R.M. Buller, M. Gale Jr, P.Y. Shi, and M.S. Diamond. 2010. 2'-O methylation of the viral mRNA cap evades host restriction by IFIT family members. Nature. 468: 452–456.
- 16. Decroly, E., I. Imbert, B. Coutard, M. Bouvet, B. Selisko, K. Alvarez, A.E. Gorbalenya, E.J. Snijder, and B. Canard. 2008. Coronavirus Nonstructural Protein 16 Is a Cap-0 Binding Enzyme Possessing (Nucleoside-2'O)-Methyltransferase Activity. J. Virol. 82: 8071–8084.
- 17. Sawicki, S.G., D.L. Sawicki, D. Younker, Y. Meyer, V. Thiel, H. Stokes, and S.G. Siddell. 2005. Functional and genetic analysis of coronavirus replicase-transcriptase proteins. PLoS Pathog. 1: e39.
- 18. Chen, Y., C. Su, M. Ke, X. Jin, L. Xu, Z. Zhang, A. Wu, Y. Sun, Z. Yang, P. Tien, T. Ahola, Y. Liang, X. Liu, and D. Guo. 2011. Biochemical and structural insights into the mechanisms of sars coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex. PLoS Pathog. 7: e1002294.
- 19. Decroly, E., C. Debarnot, F. Ferron, M. Bouvet, B. Coutard, I. Imbert, L. Gluais, N. Papageorgiou, A. Sharff, G. Bricogne, M. Ortiz-Lombardia, J. Lescar, and B. Canard. 2011. Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2'-o-methyltransferase nsp10/nsp16 complex. PLoS Pathog. 7: e1002059.
- 20. Menachery, V.D., K. Debbink, and R.S. Baric. 2014. Coronavirus non-structural protein 16: Evasion, attenuation, and possible treatments. Virus Res. 194: 191–199.
- 21. Smietanski, M., M. Werner, E. Purta, K.H. Kaminska, J. Stepinski, E. Darzynkiewicz, M. Nowotny, and J.M. Bujnicki. 2014. Structural analysis of human 2'-O-ribose methyltransferases involved in mRNA cap structure formation. Nat. Commun. 5: 3004.
- 22. Hodel, A.E., P.D. Gershon, and F.A. Quiocho. 1998. Structural basis for sequence-nonspecific recognition of 5'-Capped mRNA by a cap-modifying enzyme. Mol. Cell. 1: 443–447.
- 23. Bouvet, M., C. Debarnot, I. Imbert, B. Selisko, E.J. Snijder, B. Canard, and E. Decroly. 2010. In vitro reconstitution of sars-coronavirus mRNA cap methylation. PLoS Pathog. 6: e1000863.
- 24. Rosas-Lemus, M., G. Minasov, L. Shuvalova, N. Inniss, O. Kiryukhina, G. Wiersum, Y. Kim, R. Jedrzejczak, N. Maltseva, M. Endres, L. Jaroszewski, A. Godzik, A. Joachimiak, and K. Satchell. 2020. The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. bioRxiv, doi 10.1101/2020.04.17.047498 (preprint posted Apr 26, 2020).
- 25. Viswanathan, T., S. Arya, S.H. Chan, S. Qi, N. Dai, A. Misra, J.G. Park, F. Oladunni, D. Kovalskyy, R.A. Hromas, L. Martinez-Sobrido, and Y.K. Gupta. 2020. Structural basis of RNA cap modification by SARS-CoV-2. Nat. Commun. 11: 3718.
- 26. Lin, S., H. Chen, F. Ye, Z. Chen, F. Yang, Y. Zheng, Y. Cao, J. Qiao, S. Yang, and G. Lu. 2020. Crystal structure of SARS-CoV-2 nsp10/nsp16 2'-O-methylase and its implication on antiviral drug design. Signal Transduct. Target. Ther. 5: 5–8.
- 27. Debarnot, C., I. Imbert, F. Ferron, L. Gluais, I. Varlet, N. Papageorgiou, M. Bouvet, J. Lescar, E. Decroly, and B. Canard. 2011. Crystallization and diffraction analysis of the SARS coronavirus nsp10-nsp16 complex. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 67: 404–408.

- 28. Krafcikova, P., J. Silhan, R. Nencka, and E. Boura. 2020. Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. Nat. Commun. 11: 3717.
- 29. Khan, R.J., R.K. Jha, G.M. Amera, M. Jain, E. Singh, A. Pathak, R.P. Singh, J. Muthukumaran, and A.K. Singh. 2020. Targeting SARS-CoV-2: a systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2'-O-ribose methyltransferase. J. Biomol. Struct. Dyn.: 1–14.
- 30. Petrossian, T.C., and S.G. Clarke. 2011. Uncovering the human methyltransferasome. Mol. Cell. Proteomics. 10: M110.000976.
- 31. Wang, Y., Y. Sun, A. Wu, S. Xu, R. Pan, C. Zeng, X. Jin, X. Ge, Z. Shi, T. Ahola, Y. Chen, and D. Guo. 2015. Coronavirus nsp10/nsp16 Methyltransferase Can Be Targeted by nsp10-Derived Peptide In Vitro and In Vivo To Reduce Replication and Pathogenesis. J. Virol. 89: 8416–8427.
- 32. Ke, M., Y. Chen, A. Wu, Y. Sun, C. Su, H. Wu, X. Jin, J. Tao, Y. Wang, X. Ma, J.A. Pan, and D. Guo. 2012. Short peptides derived from the interaction domain of SARS coronavirus nonstructural protein nsp10 can suppress the 2'-O-methyltransferase activity of nsp10/nsp16 complex. Virus Res. 167: 322–328.
- 33. Lee, A.C.L., J.L. Harris, K.K. Khanna, and J.H. Hong. 2019. A comprehensive review on current advances in peptide drug development and design. Int. J. Mol. Sci. 20: 2383.
- 34. Šali, A., and T.L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234: 779–815.
- 35. Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79: 926–935.
- 36. Hess, B. 2008. P-LINCS: A parallel linear constraint solver for molecular simulation. J. Chem. Theory Comput. 4: 116–122.
- 37. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. J. Chem. Phys. 98: 10089–10092.
- 38. Berendsen, H.J.C., J.P.M. Postma, W.F. Van Gunsteren, A. Dinola, and J.R. Haak. 1984. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81: 3684–3690.
- 39. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. J. Chem. Phys. 126: 014101.
- 40. Zimmerman, M.I., and G.R. Bowman. 2015. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. J. Chem. Theory Comput. 11: 5747–5757.
- 41. Zimmerman, M.I., and G.R. Bowman. 2016. How to Run FAST Simulations. Methods Enzymol. 578: 213–225.
- 42. Zimmerman, M.I., K.M. Hart, C.A. Sibbald, T.E. Frederick, J.R. Jimah, C.R. Knoverek, N.H. Tolia, and G.R. Bowman. 2017. Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. ACS Cent. Sci. 3: 1311–1321.
- 43. Hendlich, M., F. Rippmann, and G. Barnickel. 1997. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. J. Mol. Graph. Model. 15: 359–363.
- 44. Ward, M.D., M.I. Zimmerman, S.J. Swamidass, and G.R. Bowman. 2020. DiffNets: Self-supervised deep learning to identify the mechanistic basis for biochemical differences between protein variants. bioRxiv, doi 10.1101/2020.07.01.182725 (preprint posted July 02, 2020).

- 45. Bowman, G.R., V.S. Pande, and F. Noé. 2014. An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Springer.
- 46. Chodera, J.D., and F. Noé. 2014. Markov state models of biomolecular conformational dynamics. Curr. Opin. Struct. Biol. 25: 135–144.
- 47. Schütte, C., and M. Sarich. 2015. A critical appraisal of Markov state models. Eur. Phys. J. Spec. Top. 224: 2445–2462.
- 48. Porter, J.R., M.I. Zimmerman, and G.R. Bowman. 2019. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. J. Chem. Phys. 150: 044108.
- 49. Shrake, A., and J.A. Rupley. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79: 351–371.
- 50. McGibbon, R.T., K.A. Beauchamp, M.P. Harrigan, C. Klein, J.M. Swails, C.X. Hernández, C.R. Schwantes, L.P. Wang, T.J. Lane, and V.S. Pande. 2015. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J. 109: 1528–1532.
- 51. Gonzalez, T.F. 1985. Clustering to minimize the maximum intercluster distance. Theor. Comput. Sci. 38: 293–306.
- 52. Bowman, G.R., X. Huang, and V.S. Pande. 2009. Using generalized ensemble simulations and Markov state models to identify conformational states. Methods. 49: 197–201.
- 53. Gentle, J.E., L. Kaufman, and P.J. Rousseuw. 1991. Finding Groups in Data: An Introduction to Cluster Analysis. Biometrics.
- 54. Zimmerman, M.I., J.R. Porter, X. Sun, R.R. Silva, and G.R. Bowman. 2018. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. J. Chem. Theory Comput. 14: 5459–5475.
- 55. Porter, J.R., K.E. Moeder, C.A. Sibbald, M.I. Zimmerman, K.M. Hart, M.J. Greenberg, and G.R. Bowman. 2019. Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. Biophys. J. 116: 818–830.
- 56. Madeira, F., Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A.R.N. Tivey, S.C. Potter, R.D. Finn, and R. Lopez. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47: W636–W641.
- 57. Waterhouse, A.M., J.B. Procter, D.M.A. Martin, M. Clamp, and G.J. Barton. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics. 25: 1189–1191.
- 58. Pettersen, E.F., T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. 2004. UCSF Chimera A visualization system for exploratory research and analysis. J. Comput. Chem. 25: 1605–1612.
- 59. Joseph, J.S., K.S. Saikatendu, V. Subramanian, B.W. Neuman, A. Brooun, M. Griffith, K. Moy, M.K. Yadav, J. Velasquez, M.J. Buchmeier, R.C. Stevens, and P. Kuhn. 2006. Crystal Structure of Nonstructural Protein 10 from the Severe Acute Respiratory Syndrome Coronavirus Reveals a Novel Fold with Two Zinc-Binding Motifs. J. Virol. 80: 7894–901.

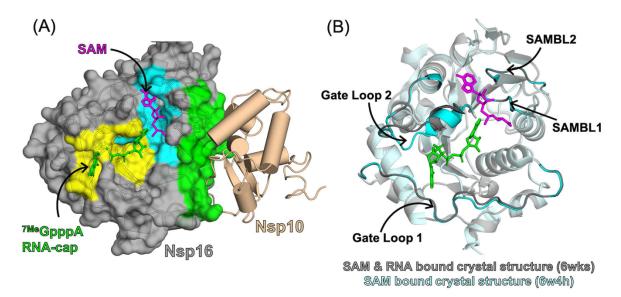


Figure 5.1 Structural view of NSP16.

Substrate binding pockets and Nsp10 binding interface of Nsp16 observed in the crystal structure of the Nsp16/Nsp10 complex (PDB: 6wks). (A) Surface representation of Nsp16 showing the SAM-binding pocket (cyan), RNA-binding pocket (yellow) and Nsp10-binding interface (green). (B) Overlay of Nsp16 structures from structures of the Nsp16/Nsp10 complex with RNA (PDB: 6wks, shown in grey) and without RNA (PDB: 6w4h, shown in cyan), showing structural heterogeneity in the RNA-binding site. Gate loop 1 and Gate loop 2 of the RNA-binding pocket, and SAM-binding loop 1 (SAMBL1) and SAM-binding loop 2 (SAMBL2) lining the SAM-binding pocket are highlighted.

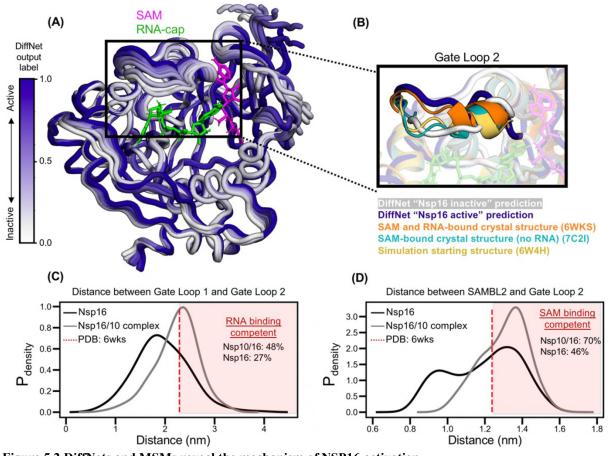


Figure 5.2 DiffNets and MSMs reveal the mechanism of NSP16 activation.

Nsp10 binding shifts Nsp16's conformational ensemble increasing its propensity to adopt structural states that are ligand binding compatible. (A) Ten structures of Nsp16 that represent the DiffNet prediction changing from inactive to active (white to purple). (B) Comparison of the DiffNet predicted active and inactive states (purple + white, respectively) to the starting simulation state (yellow), a known SAM and RNA bound structural state (orange), and a known SAM (but not RNA) bound state (teal). All structures aligned to 6WKS (orange). (C) Probability-weighted distance distribution between RNA-binding gate loops 1 and 2 comparing monomeric Nsp16 (black) to the Nsp10-Nsp16 complex (gray). (D) Probability-weighted distance distribution between SAM-binding loop 2 and gate loop 2, comparing monomeric Nsp16 (black) to the Nsp10-Nsp16 complex (gray). For (C) and (D), the distance for a SAM and RNA bound crystal structure is also plotted (red dotted line).

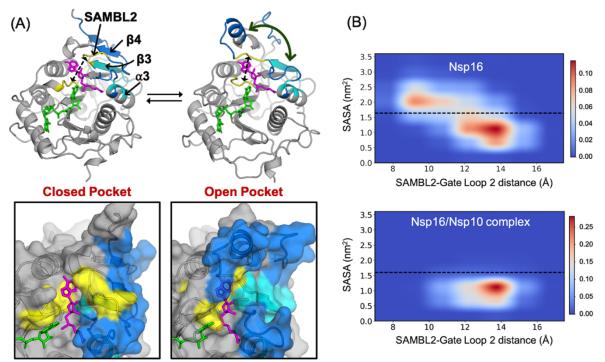


Figure 5.3 Cryptic pocket opening in NSP16.

Cryptic pocket opening in SARS-CoV2 Nsp16. (A) Structural states with the cryptic pocket closed and open. The insets show surface views of the closed and open pocket. Residues exposed upon pocket opening are shown in cyan and the regions undergoing the opening motion are shown in blue. Collapse of the SAM-binding pocket is measured as the distance between SAMBL2 and gate loop 2, shown in yellow. (B) Equilibrium probability weighted 2D histograms of solvent-accessible surface area (SASA) of pocket residues (shown in cyan in A) and the distance between SAMBL2 and gate loop 2 in Nsp16 for monomeric Nsp16 (upper panel) and the Nsp16/Nsp10 complex (lower panel). The black dotted line separates the pocket closed and open states in Nsp16.

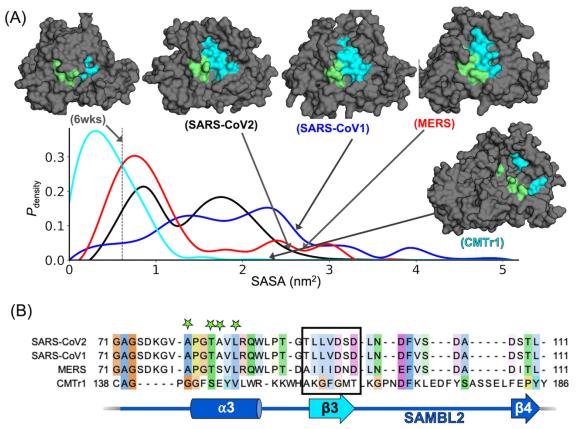


Figure 5.4 Cryptic pocket opening is conserved across coronavirus homologs.

Comparison of cryptic pocket opening in Nsp16 homologs and human CMTr1. (A) Equilibrium probability-weighted distribution of the solvent exposure of pocket forming residues for SARS-CoV2 (black), SARS-CoV1 (blue), MERS (red) and CMTr1 (cyan). Structures representing the open pocket are shown for each homolog with $\beta3$ colored in cyan, and other pocket forming residues from $\alpha3$ colored in green. Black dotted line depicts SASA of pocket residues in the crystal structure of Nsp16/Nsp10 complex (PDB: 6wks). (B) Structure-based sequence alignment of Nsp16 homologs (SARS-CoV2, SARS-CoV1 and MERS) and human CMTr1 is shown for the cryptic pocket forming regions. Residues of $\beta3$ are marked inside the black colored box, and other pocket forming residues from $\alpha3$ are by green colored stars.

Chapter 6

Predicting cryptic pocket opening from protein structures using graph neural networks

6.1 Preamble

This chapter is adapted from the following conference proceeding: Ward, M.D.,* Meller, A.,* Kshirsagar, M., Perhavec, F.O., Borowsy, J.H., Miller. G., Ferres, J.L., and Bowman, G.R (2021). "Predicting cryptic pocket opening from protein structures using graph neural networks", *NeurIPS Machine Learning for Structural Biology workshop.*

6.2 Introduction

A structure of a protein's native, folded state can reveal potential drug binding sites, but leaves us blind to other potential sites that form as the protein structure fluctuates in solution. There are over 100 confirmed examples of these "other" binding sites where a small molecule binds in a pocket on a protein which was not observable from any previously determined structure of that protein (i.e. a "cryptic pocket")¹. Currently, it is challenging to predict these cryptic pockets from the ground state experimental structure, but the ability to do so would come with several benefits. For example, protein structures that lack any obvious binding pockets are often considered undruggable², but they may actually prove to be good drug targets if they have a cryptic pocket that can be targeted. Even if a protein structure already reveals a binding pocket that can be targeted with a small molecule (e.g. an active site), it is useful to know if there are cryptic pockets as targeting cryptic pockets may improve specificity (i.e. reduce off-target effects

^{*}These authors contributed equally to the work

when targeting a family of homologous proteins) or lead to the discovery of allosteric activators^{3,4}.

Current methods for identifying cryptic pockets in proteins are either slow or have low accuracy. Molecular dynamics simulations, which use physics-based force fields to model protein structural fluctuations⁵, are the primary means to identify and sample structural configurations of cryptic pockets but they often consume 100s of GPU hours per protein. Ideally, one could employ an algorithm that quickly and accurately determines if a protein will form a cryptic pocket, then use this result to determine if resources should be deployed to run a costly simulation or an experimental drug screen. Cryptosite is one such machine learning algorithm that predicts which amino acid residues of a protein will form a cryptic pocket with good performance (AUC=0.83)¹. However, this method takes ~1 day to run because it relies on simulation data as input to the algorithm. When simulation features are removed its performance markedly drops (AUC=0.74).

In the current study, we train a graph neural network to accurately determines sites of cryptic pockets from experimental structures. In a previous study, molecular dynamics simulations were performed to identify and sample cryptic pockets across most proteins in the SARS-CoV-2 proteome to uncover ~50 new potential binding sites⁶. Across these simulations there are thousands of events where a cryptic pocket forms. We used these events as training examples to train a graph neural network to classify whether or not a residue is likely to participate in a cryptic pocket given the 3D topology and the chemical environment of its neighborhood.

6.3 Results

6.3.1 Predicting cryptic pockets with Geometric Vector Perceptrons

We hypothesized that the propensity of an amino acid residue to participate in a cryptic pocket is a function of the 3D topology and the chemical environment in which the amino acid resides. On one hand, if a residue is in a tightly packed region of a protein and has extremely strong attractive interactions with its neighbors, it is unlikely to undergo a structural rearrangement that creates a pocket that a ligand might bind. On the other hand, if a residue is in a loosely packed environment and has weak interactions with its neighbors, it may be more likely to be in a region that forms a cryptic pocket. Given this hypothesis we sought to train a model that takes a protein structure as input and outputs a value that indicates the likelihood that each amino acid residue will participate in a cryptic pocket.

Previous work has established that graph neural networks are an efficient way to learn complicated tasks from 3D protein structures. Specifically, a graph neural network architecture that employs a Geometric Vector Perceptron (GVP) has been previously shown to accurately evaluate the model quality of predicted protein structures and also successfully predict feasible protein sequences from 3D structures⁷. Briefly, the GVP-based graph neural network takes a set of node (i.e. an amino acid residue) and edge features that describe geometric and chemical features describing a residue, *i*, and its relation to another residue, *j* (**Fig. 6.1a**). To learn a representation for a given residue, the network uses message passing in which messages from neighboring residues and edges are used to update the residue representation.

Here, we adapt the GVP-based graph neural network to the task of predicting sites of cryptic pockets from a native, folded structure of a protein. As in the original paper, we use node features that include: the type of amino acid residue, *sine* and *cosine* transformations of backbone dihedral angles, an imputed unit vector between the alpha and beta carbon, and a forward and reverse unit vector to the i - 1 and i + 1 neighboring residues. Edge features include

a unit vector between nodes, a distance between nodes, and a *sine* transformation of the distance in the protein's primary sequence. We update a node of interest using its 30 nearest neighbors as done in the original work. In the original work, protein-level predictions were made by taking the mean representation of all residues and making a prediction. Importantly, we do not take a mean of residues, but instead predict on residues individually.

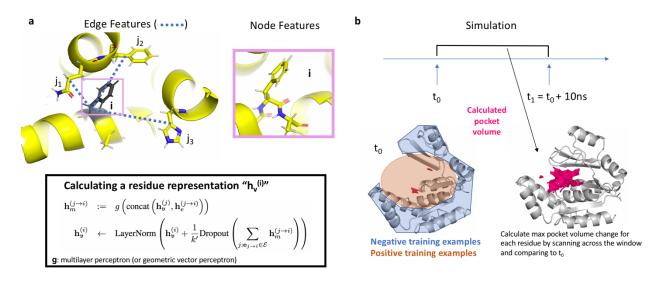


Figure 6.1 Training scheme to predict cryptic pocket opening.

Depiction of how residue level representations are calculated in the graph neural network (a) and how residues are labelled for training (b). Node features include a one-hot encoding of residue type, sine and cosine transformations of 3 different backbone dihedral angles (only 2 shown for clarity), a unit vector capturing the direction of the C-alpha to C-beta bond, and forward and reverse unit vectors (C-alpha to C-alpha from preceding and following residues). Edge features include a unit vector in the direction of the neighbor, the distance to the neighbor, and a sine transformation of the distance in sequence space.

The training data for our model comes from time windows from molecular dynamics simulations. Specifically, we select a structure at some timepoint in simulation (t₀) and the resulting structures within the next 10 nanoseconds of simulation to calculate the maximum pocket volume increase across the protein structures within that time window using the pocket detection algorithm LIGSITE⁸ (**Fig. 6.1b**). LIGSITE outputs sets of "pocket grid points" that indicate cavities on the protein surface (i.e. points that are surrounded by protein on all sides).

For all residues, we calculate how many pocket grid points are within 5 Angstroms of the residue. We label a residue a positive example if at some point in the time window that residue's assigned pocket volume increases by 40Å^3 (roughly the size of an ADP molecule) relative to its volume at time t_0 . We label a residue as a negative example if the change is less than 10Å^3 , and we do not consider residues with intermediate values.

6.3.2 Graph neural networks accurately predict residue level pocket volume changes from simulation data

We trained and evaluated a model using a simulation dataset of SARS-CoV-2 proteins (and related human proteins) that consisted of 17 proteins. First, we chose 15 proteins randomly (resulting in 1,160 simulation trajectories) and split the trajectories into training and validation sets in a 90:10 split. Then, we held out all trajectories from 2 proteins as a test set. Importantly, since the simulations provide many different structural configurations of the proteins and because each protein has many residues, the training set contained ~1.6M training examples (~176K positive, ~1.4M negative).

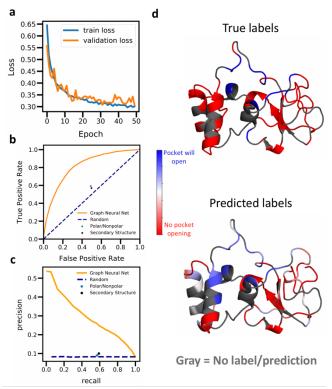


Figure 6.2 Evaluation of graph neural network's ability to predict cryptic pocket opening in simulation. Several metrics evaluating the model relay its ability to accurately predict cryptic pocket opening in simulations. (a) Training and validation loss. (b) ROC curve. (c) Precision-recall curve. (d) Predictions and ground truth labels overlaid on a random structure from simulation.

Our model effectively learned to classify how the pocket volume around a residue will change over the course of 10ns of simulation. First, we show that the model trains stably with the training and validation loss flattening around 25 epochs (Fig. 6.2a). We apply the best model (according to validation loss) to the test set of 2 held out proteins and calculate a ROC-AUC of 0.82 (Fig. 6.2b). Overlaying the ground truth labels and predicted labels onto a random structure selected from simulation also demonstrates the high accuracy of the model (Fig. 6.2d).

Additionally, we have plotted a precision-recall curve and observe good performance (Fig. 6.2c). In both the ROC curve and precision-recall curve our model substantially outperforms a random baseline, a model that classifies residues based on whether they are polar or nonpolar, and a model that classifies residues based on if they have secondary structure.

6.3.3 Graph neural networks accurately identify cryptic pockets from experimental structures

To evaluate if our model can predict known sites of cryptic pockets without the need for simulations, we applied a trained model to a new test set of experimental protein structures with known cryptic pockets. This model was trained identical to the model from section 3.1 except it also included the 2 prior test set proteins. For the test set, we curated a set of 11 protein structure pairs that include an apo protein (no ligand bound) and the corresponding holo protein (ligand bound to cryptic site). We applied our model to the 11 apo protein structures (which were not part of the training) and found that it accurately identified the known cryptic pockets (**Fig. 6.3**, ROC-AUC=0.78). This result is slightly better than the reported result for the related algorithm CryptoSite¹ (ROC-AUC=0.74 when not using features extracted from simulations). However, the test set used in the two studies are different and a follow-up comparison with an identical test set is warranted.

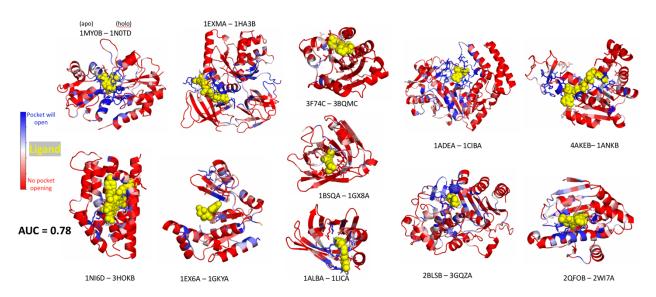


Figure 6.3 Graph neural network prediction on experimentally determined cryptic pockets.

A graph neural network accurately predicts the locations of cryptic sites in 11 crystal structures with known cryptic pockets. *Apo* structures are colored from red to blue depending on the network's prediction of whether pocket opening will occur at that residue. Ligands which bind cryptic sites (shown in yellow) are superposed onto *apo* structures. Residues at the ligand binding site are labelled as positive examples and shown in a stick representation in the *apo* structures. PDB ID's for the apo and holo structures are provided.

6.4 Conclusions

We have shown that a graph neural network trained on protein simulation data can be used to predict sites of cryptic pockets from experimental structures of proteins. First, we showed that we can predict whether or not residues in a protein will undergo structural changes that lead to increased pocket volumes over the course of 10ns of simulation. Next, we showed that this same model can accurately predict sites of cryptic pockets from single structures without the need to run molecular dynamics simulations.

While our model can accurately predict whether or not residues in a protein will undergo structural changes that lead to increased pocket volumes over the course of 10ns of simulation, there are come caveats. The highest precision only reaches ~0.5 meaning there is likely to be one false positive for every true positive at the lowest recall value. One explanation is that pocket formation is stochastic across a 10ns simulation window. Even with the exact same starting structural configuration, a residue can sometimes form a pocket in 10ns, and sometimes not. Therefore, many of the false positives called by our model may actually be residues that do sometimes form cryptic pockets in other 10ns windows.

It may be surprising that our model can predict sites of cryptic pocket formation from crystal structures given that the model was not trained to perform this task. Nonetheless, we find that a model trained to predict how pocket volumes change in protein structures over the course of 10ns of simulation is transferable to the harder task of predicting cryptic pocket formation from experimental structures. Given that success on the former task begets success on the latter

159

task, this suggests that 10ns of simulation time may be a sufficient amount to sample at least partial cryptic pocket openings with molecular dynamics simulations. This is encouraging since, historically, simulations to discover cryptic pockets usually consumed far more resources.

This work represents an encouraging proof-of-concept and should be improved by access to more simulation training data on a larger set of proteins, as well as, better model selection, which can come from a hyperparameter search to find the best model.

Bibliography

- 1. Cimermancic P, Weinkam P, Rettenmaier TJ, et al. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J Mol Biol*. 2016. doi:10.1016/j.jmb.2016.01.029
- 2. Crews CM. Targeting the Undruggable Proteome: The Small Molecules of My Dreams. *Chem Biol.* 2010. doi:10.1016/j.chembiol.2010.05.011
- 3. Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. *Cell.* 2013. doi:10.1016/j.cell.2013.03.034
- 4. Hollingsworth SA, Kelly B, Valant C, et al. Cryptic pocket formation underlies allosteric modulator selectivity at muscarinic GPCRs. *Nat Commun*. 2019. doi:10.1038/s41467-019-11062-7
- 5. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol*. 2002. doi:10.1038/nsb0902-646
- 6. Zimmerman MI, Porter JR, Ward MD, et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat Chem.* 2021. doi:10.1038/s41557-021-00707-0
- 7. Jing B, Eismann S, Suriana P, Townshend RJL, Dror R. Learning from Protein Structure with Geometric Vector Perceptrons. 2020:1-18. http://arxiv.org/abs/2009.01411.
- 8. Hendlich M, Rippmann F, Barnickel G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*. 1997. doi:10.1016/S1093-3263(98)00002-3

Chapter 7

Conclusions

7.1 Main Findings

Applying deep learning approaches to protein biophysics has come with many breakthroughs in recent years, and the goal of this work has been to extend these approaches to areas in protein biophysics that have been relatively untouched. In particular, while there has been an explosion of work predicting static structures of biomolecules^{1–3} there has been less work characterizing structural ensembles. Among the studies at the intersection of deep learning and structural ensembles, most of the work has been around the goal of improving sampling of ensembles.^{4–7} There has been less emphasis on using deep learning with the goal of analyzing structural ensemble data. In this section, I discuss the progress I made pursuing both of these goals.

Chapters 2, 3 and 5 detail the development and application of DiffNets, a deep learning based approach for analyzing structural ensembles to highlight differences between datasets. One feature that makes DiffNets stand out compared to previous methods is that it includes a specific mechanism to highlight differences between datasets. Namely, the low dimensional representation of data is constrained to predict which dataset each structure comes from. Taken together with the task of reconstructing the structure, the algorithm learns how to sort structures based on their association with one system or the other, which ultimately helps connect structural features to the biochemical properties that distinguish the systems of interest. After demonstrating that DiffNets correctly identifies structural signatures that distinguish protein variants and isoforms across model systems we applied it to new systems. In Chapter 3, we

applied it to naturally occurring genetic variants of the oxytocin receptor and showed how different variants alter signaling through modified interactions with β -arrestin and Gq. In chapter 5, we showed how the SARS-CoV-2 NSP10 protein stimulates the enzymatic activity of NSP16 by increasing NSP16's propensity to adopt structural configurations with active sites open wide enough to be competent for ligand binding.

In chapter 4, I adapted the goals of my thesis work joining in the fight against COVID19⁸ by searching for new druggable sites across SARS-CoV-2 proteins. Structural biologists determined structures of the majority of the SARS-CoV-2 proteome within months of the global pandemic announcement.⁹ These structures revealed ways to inhibit these proteins as a means to cripple SARS-CoV-2. Building on this effort, we simulated how the protein structures fluctuate in solution and characterized more than 50 potential drug binding sites that had not been previously described. This effort required an unfathomable amount of resources; over a million people donated spare CPU/GPU cycles from their personal computers through a project called Folding@Home.¹⁰ The amount of resources required made me realize the discovery of cryptic pockets at scale using MD simulations is not currently feasible. However, MD simulations provide invaluable structural information about the cryptic pockets, so a way to prioritize proteins for cryptic pocket discovery could have tremendous impact.

In chapter 6, we developed a deep learning approach for predicting sites of cryptic pockets from single protein structures. Specifically, we utilized the simulation data from chapter 4 to train a graph neural network to predict pocket opening events that were observed in simulation. This model accurately predicts which residues would participate in cryptic pocket opening over the course of 10 nanoseconds of simulation. This suggests that the topology and

chemical environment around a given residue contains sufficient information to forecast future events on a much longer time horizon than a typical MD timestep (~2 femtoseconds). We also found that this same model accurately determines sites where ligands had bound in known cryptic sites (derived from the PDB). This finding was surprising because it was not obvious if predicting a protein's breathing motions would be a good proxy for predicting regions on a protein that are amenable to ligand binding.

7.2 Future Directions

One aspect of scientific research that is equally exciting and frustrating is that no research project is ever neatly completed; the number of questions seems to grow exponentially the deeper you go. My work certainly has not been without its frustrations and there are many things I would do to improve each project I have undertaken. This section will detail some of the future improvements that could be undertaken, which both excite and agitate me.

While DiffNets has been a useful tool, there are several modifications that could expand the scope of its usefulness. One problem with DiffNets is that the architecture is not translationally, nor rotationally, invariant. This means that if the same exact structure gets translated or rotated, the DiffNet will produce a different output. This does not match the underlying physics; the structure has the exact same properties even if it is translated or rotated. Therefore, DiffNets only works well in situations where structures across the ensemble can be well aligned to a starting structure. For this reason, DiffNets does not work well on highly dynamic proteins. Over the past couple of years there have been several physically-inspired neural networks that are equivariant with respect to translation and rotation and these are good candidates architectures to build future DiffNets.¹¹

DiffNets could also be improved by considering transitions between states. Currently, DiffNets is trained only on individual structures from simulations. This limits the networks to only learn differences between ensembles in terms of their preferences for specific structural states. However, there are cases where two proteins with different behavior might adopt the same exact states the same proportion of time, but differ in how frequently they transition between states. Adding in the ability for DiffNets to train on chunks of trajectories from MD simulations might allow them to hone in on important dynamic transitions between structural states rather than just focusing on individual structural states. There are several well-established deep learning architectures for handling this problem including recurrent neural networks¹² and transformers.¹³

Beyond technical improvements, it would be exciting to see DiffNets used in a way that advances progress in medicine. The studies I carried out in chapters 3 and 5 helped understand the mechanism of how perturbations to a protein (e.g. a mutation or a regulatory partner) affect its behavior. These studies are descriptive in nature, and DiffNets was not used to make any predictions that may serve in developing new therapeutics. However, DiffNets do have the potential to be used in this capacity. For example, one might use DiffNets to compare isoforms within a family of proteins. This could help uncover structural features that can be targeted in one isoform but not others allowing for the development of a drug with high specificity. One might also be able to use DiffNets to help with precision medicine. For example, if a patient with a disease has a genetic variant of unknown significance, it's possible that MD simulations in conjunction with DiffNets could help establish if that variant is similar to other variants, which would inform what type of treatment the patient should receive.

Improving cryptic pocket prediction is another avenue for advancing medicine. Cryptic pocket discovery comes with two major challenges. First, determining if a protein has a cryptic pocket is important to determine if it is worth pursuing the protein as a drug target. In chapter 5, we addressed this challenge with a graph neural network-based solution. Our solution reduces the time needed to make this determination from days, or weeks, to less than a second. While the results are promising, there is room for improvement with the accuracy of the model, which will probably come as more cryptic pocket data emerges in the PDB. The second challenge is structurally characterizing a protein's cryptic pockets. Currently this require weeks (at least) of MD simulation time to sample the relevant structures. I propose that this could be accelerated by a neural network explicitly trained to generate structures of open cryptic pockets. Specifically, one could train a time-lagged autoencoder on pairs of structures that include closed and open pockets. From the closed state, the network would have to predict the structure of the open state. After training, one could apply this model to the native, folded structure of a protein and generate structures with open pockets. These open states could serve as a template for drug design. Altogether, this would provide a rational way to target cryptic pockets after ~seconds of calculation on the native, folded structure.

Bibliography

- 1. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020. doi:10.1038/s41586-019-1923-7
- 2. Townshend RJL, Eismann S, Watkins AM, et al. Geometric deep learning of RNA structure. *Science* (80-). 2021. doi:10.1126/science.abe5650
- 3. Townshend RJL, Bedi R, Suriana PA, Dror RO. End-to-end learning on 3D protein structure for interface prediction. In: *Advances in Neural Information Processing Systems*.; 2019.
- 4. Noé F, Olsson S, Köhler J, Wu H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* (80-). 2019. doi:10.1126/science.aaw1147
- 5. Wang Y, Ribeiro JML, Tiwary P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat Commun*. 2019. doi:10.1038/s41467-019-11405-4
- 6. Wehmeyer C, Noé F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J Chem Phys.* 2018. doi:10.1063/1.5011399
- 7. Hernández CX, Wayment-Steele HK, Sultan MM, Husic BE, Pande VS. Variational encoding of complex dynamics. *Phys Rev E*. 2018. doi:10.1103/PhysRevE.97.062412
- 8. Varadarajan J, Brown AM, Chalkley R. Biomedical graduate student experiences during the COVID-19 university closure. *PLoS One.* 2021. doi:10.1371/journal.pone.0256687
- 9. Wang MY, Zhao R, Gao LJ, Gao XF, Wang DP, Cao JM. SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Front Cell Infect Microbiol*. 2020. doi:10.3389/fcimb.2020.587269
- 10. Shirts M, Pande VS. Screen savers of the world unite. *Science* (80-). 2000. doi:10.1126/science.290.5498.1903
- 11. Jing B, Eismann S, Suriana P, Townshend RJL, Dror R. Learning from Protein Structure with Geometric Vector Perceptrons. 2020:1-18. http://arxiv.org/abs/2009.01411.
- 12. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput.* 1997. doi:10.1162/neco.1997.9.8.1735
- 13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*.; 2017.

Appendices

A.1 Appendix to Chapter 2

A.1.1 Expectation Maximization Algorithm

We hypothesize that it is possible to use EM to learn the association between individual structures and a biochemical property of interest. EM is a statistical method that allows the parameters of a model to be fit, even when the outputs of the model cannot directly be observed in the training data (i.e. when they are hidden). In our case, the hidden variables are the elements of a vector of numbers, associated with every structure in the simulation training data. Each variable should be a 1 if it is associated with the biochemical property and 0 otherwise, but we do not know what the correct value is, they are hidden. First, this vector is initialized to reasonable starting values. Next, during the Maximization step (M-step) we train a neural network to create a mapping between each structure's descriptors (i.e. XYZ coordinates) and the current estimate of the hidden variables. Then, during the Expectation step (E-step), we reestimate our hidden variables using the trained model and the region constraints that specify how many structures we expect to be associated with the biochemical property of interest. Finally, we alternate between the E- and M-steps for a predefined number of steps.

The EM algorithm alternates between E- and M-steps. To initialize the algorithm, we pick an output vector $Y = (y_i)$ such that all values corresponding to simulation frames of one class of variant are assigned 0s, and all other values are assigned 1s. This is our initial guess for our hidden variables, $K = (k_i)$ (Eq. 1). Each element of K is our current estimate of which structures are associated with the biochemical property of interest. Next, the M-step fits a neural network using K as targets (Eq. 2),

$$K_1 \leftarrow Y_{init}$$
 (1)

$$W_1 \text{ and } Y_1 \leftarrow M - step \quad (K_1, D),$$
 (2)

where W_1 is the tuned weights of the neural network and Y_1 is the output of the model using these weights with the data. This output vector, Y_1 , is used in the E-step to compute the next guess for the hidden variables K (Eq. 3). The next iteration repeats the E- and M-steps,

$$K_2 \leftarrow E - step \quad (Y_1)$$
 (3)

$$W_2$$
 and $Y_2 \leftarrow M - step \quad (K_2, D),$ (4)

Subsequent iterations repeat these steps for a predefined number of steps. As the algorithm progresses, both the *K* and *Y* vectors should converge to a value that indicates the extent that a structure is associated with the biochemical property of interest. They should label the structures associated with the property with high probabilities, and the other structures with low probabilities.

The E-step computes the expected values of the hidden variables K from the outputs Y conditioned on constraints defined by the user (e.g. only 0-30% of simulation data is expected to be associated with the property of interest for one class of data, and 40-70% for the other class). The expectation of the hidden variables is the probability-weighted average of all binary realizations of binomial distributions parameterized by Y that assign the right number of structures as being associated with the property of interest. Conceptually, the expectation is computed by, first, enumerating all binary realizations of Y, each denoted as a vector of boldface variables $\mathbf{Y} = (y_i)$. Second, vectors that do not have the right number of structures according to the user-defined constraints are rejected. Third, the remaining vectors are scored by their probability according to Y, and, finally, a probability-weighed average of the binary vectors is computed. This average vector is the expectation, and is assigned to K. A straightforward Python

implementation of this calculation can be found here (https://github.com/bowman-lab/diffnets/blob/master/diffnets/exmax.py) under the function name "expectation_range_EXP".

While conceptually clear, computing K in this way is very slow because there are exponentially many realizations of Y that must be enumerated. Fortunately, the expectation is computable in polynomial time. Here, we treat the structure labels as binary random variables following binomial distributions parameterized by Y. For each class of data, the expectation of these variables is assigned to elements of K. Given the user-defined constraints about the number of structures associated with the property of interest, this update can be derived from Baye's Rule,

$$k_s = E[y_s \mid S_L \le y_r \le S_U] \tag{5}$$

$$= P(y_s \text{ is } 1) * \left(\frac{P(S_L - 1 \le y_r - y_s \le S_U - 1)}{P(S_L \le y_r \le S_U)}\right)$$
 (6)

where y_r is the integer sum of the binary labels associated with the structures of the given class which are associated with the biochemical property of interest, y_s is the binary label of a given structure (site s), $P(y_s \text{ is } 1)$ is the probability that the structure is associated with the biochemical property according to Y, the numerator is the probability that the number of structures associated with the biochemical property (ignoring site s) ranges from $S_L - 1$ to $S_U - 1$, and the denominator is the probability that the number of structures associated with the biochemical property range from S_L to S_U . S_L and S_U are equal to the number of structures in a given class that are associated with the biochemical property of interest according to the user-defined constraints.

A.1.2 Supporting Figures

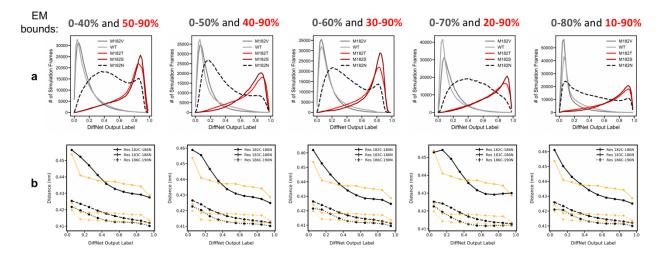


Figure A.1.1 Self-supervised DiffNets are robust across a range of expectation maximization bounds.(a) Histogram showing DiffNet output labels across all simulation frames from M182T and M182S (red – highly stable variants in training set) versus WT and M182V (grey – less stable variants in training set) across a range of expectation maximization bounds. Predictions on a less stable variant not seen during training (M182N) are also shown (black dotted line). (b) Three key hydrogen bond lengths in helix 9 as a function of the DiffNet output label (n=1,300,420 for each plot) (yellow – supervised, black – self-supervised), which ranges from zero for structures associated with low stability to one for structures associated with high stability. The distances are between the carbonyl carbon of the i'th residue and the nitrogen of the (i+4)'th residue. Standard error bars are not visible since the standard error is smaller than scatter points.

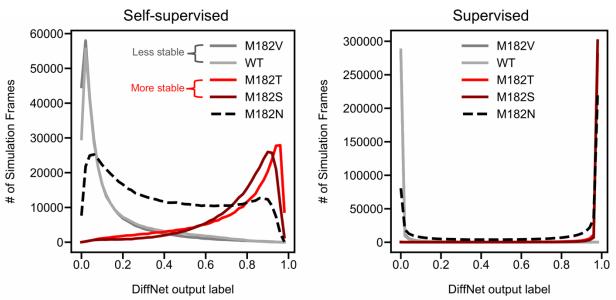


Figure A.1.2 Self-supervised DiffNets improve ability to predict property of a variant outside the training. Histogram of final DiffNet output labels for all simulation data points organized by variant (red – more stable variants, grey – less stable variants, black – less stable variant not seen during training) for a self-supervised DiffNet and a supervised DiffNet.

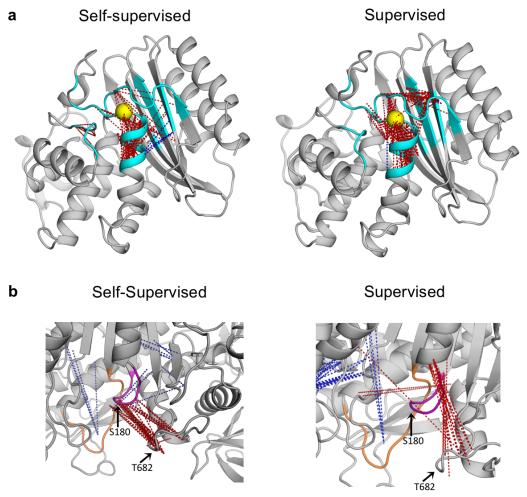


Figure A.1.3 Impact of expectation maximization on what features a DiffNet uses to distinguish variants. Dotted lines indicate distances between two atoms that change in a way that is strongly correlated with an increased DiffNet output label. Red indicates the atoms move closer together as the output label increases, blue indicates atoms moving away from each other. Results for β -lactamase variants and myosin are shown in (a) and (b) respectively. In (a), protein atoms are colored cyan if they are near the mutation, which indicates that they were included in the classification task and considered for the distance correlation calculation. The site of the single point mutation is highlighted with a yellow sphere.

Self-supervised and supervised DiffNets both capture helix 9 compaction as the key feature that distinguishes stability in β -lactamase variants and we observe no qualitative improvement for the self-supervised model. In contrast, a self-supervised DiffNet correctly hones in on the importance of S180 dynamics in determining duty-ratio in myosin isoforms, but a supervised DiffNet does not.

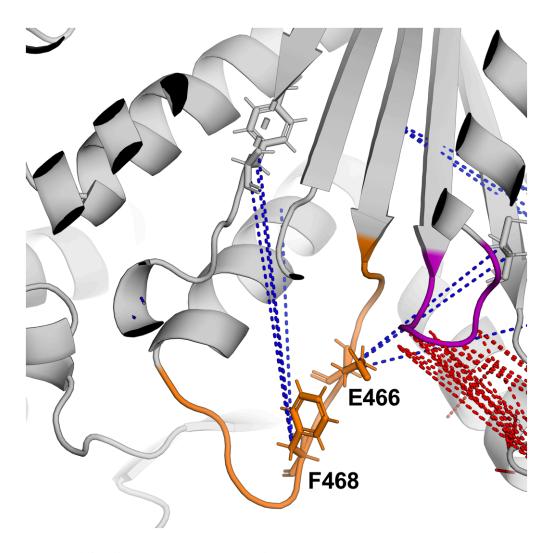


Figure A.1.4 DiffNet analysis suggests conformational changes on switch-II are important for distinguishing high-and low-duty myosin isoforms.

Dotted lines indicate distances between two atoms that change in a way that is strongly correlated with an increased DiffNet output label. Red indicates the atoms move closer together as the output label increases, blue indicates atoms moving away from each other. Switch-II is colored orange and the p-loop is colored purple.

Self-supervised DiffNet predicts that distance changes involving residues on switch-II (F468, E466) distinguish high and low-duty motor myosins. These residues are in close proximity to the p-loop (purple), which has a known role in determining duty-ratio. Moreover, E466 is directly involved in phosphate coordination in phosphate release², which lends support to the DiffNet prediction that changes in this residue are important for determining duty ratio.

Bibliography

- 1. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm . *J R Stat Soc Ser B*. 1977. doi:10.1111/j.2517-6161.1977.tb01600.x
- 2. Llinas P, Isabet T, Song L, et al. How Actin Initiates the Motor Activity of Myosin. *Dev Cell*. 2015. doi:10.1016/j.devcel.2015.03.025

A.2 Appendix to Chapter 3

Table A.2.1 Oxytocin response in Ca^{2+} assays for wild type (WT) and variant OXTRs Log(EC50)

81	/		
	Variant log(EC50) (95% CI)	WT log(EC50) (95% CI)	Significance (P)
V45L	-8.165 (-8.237 to -8.099)	-8.266 (-8.349 to -8.165)	0.0525
P108A	-8.195 (-8.355 to -8.195)	-8.307 (-8.432 to -8.185)	0.2499
V172A	-8.295 (-8.444 to -8.161)	-8.396 (-8.531 to -8.265)	0.2879
L206V	-8.432 (-8.657 to -8.212)	-8.352 (-8.492 to -8.208)	0.5264
A218T	-8.435 (-8.503 to -8.368)	-8.399 (-8.483 to -8.316)	0.4842
G221S	-8.465 (-8.655 to -8.277)	-8.533 (-8.638 to -8.427)	0.5237
A238T	-8.533 (-8.679 to -8.386)	-8.543 (-8.681 to -8.403)	0.9157
G252A	-8.734 (-8.932 to -8.516)	-8.659 (-8.769 to -8.545)	0.5224
V281M	-8.054 (-8.115 to -7.989)	-8.301 (-8.449 to -8.136)	0.0028
E339K	-8.167 (-8.282 to -8.055)	-8.445 (-8.572 to -8.314)	0.0020
R376G	-8.699 (-8.900 to -8.480)	-8.650 (-8.763 to -8.532)	0.6732

\mathbf{E}_{max}

	Variant E _{max} (95% CI)	WT E _{max} (95% CI)	Significance
V45L	93 (89 to 96)	97 (93 to 101)	0.0783
P108A	98 (91 to 107)	96 (91 to 102)	0.7520
V172A	91 (86 to 97)	98 (92 to 103)	0.0868
L206V	93 (85 to 102)	98 (92 to 104)	0.0783
A218T	91 (89 to 94)	98 (95 to 102)	0.0019

G221S	103 (95 to 111)	98 (94 to 102)	0.2751
A238T	100 (94 to 106)	99 (94 to 105)	0.8697
G252A	101 (93 to 109)	99 (95 to 103)	0.7182
V281M	74 (72 to 77)	97 (91 to 105)	< 0.0001
E339K	76 (72 to 81)	98 (94 to 104)	< 0.0001
R376G	98 (91 to 107)	98 (94 to 103)	0.9194

Results shown are point estimates and 95% confidence intervals from dose-response curves generated from three replicate experiments. Variant point estimates are shown next to point estimate from the WT control on the same plate. Statistically significant changes in log(EC50) or E_{max} are shown in bold [(extra sum of squares F test, P < 0.0045 ($\alpha = 0.05$ with Bonferroni correction for 11 comparisons)].

Table A.2.2 Oxytocin-induced β -arrestin-1 recruitment for wild type (WT) and variant OXTRs Log(EC50)

	Variant log(EC50) (95% CI)	WT log(EC50) (95% CI)	Significance (P)
V45L	-6.842 (-7.095 to -6.587)	-7.232 (-7.419 to -7.038)	0.0139
P108A	-6.793 (-6.990 to -6.593)	-7.363 (-7.474 to -7.249)	< 0.0001
V172A	-7.083 (-7.334 to -6.829)	-7.232 (-7.419 to -7.038)	0.3240
L206V	-7.416 (-7.637 to -7.187)	-7.456 (-7.598 to -7.309)	0.7793
A218T	-7.411 (-7.665 to -7.148)	-7.363 (-7.474 to -7.249)	0.7072
G221S	-7.123 (-7.462 to -6.764)	-7.232 (-7.419 to -7.038)	0.5538
A238T	-7.392 (-7.710 to -7.053)	-7.232 (-7.419 to -7.038)	0.3924
G252A	-7.559 (-7.909 to -7.192)	-7.456 (-7.598 to -7.309)	0.5925
V281M	-7.579 (undefined)	-7.363 (-7.474 to -7.249)	0.6644
E339K	-6.750 (-7.443 to -5.984)	-7.363 (-7.474 to -7.249)	0.0478
R376G	-7.566 (-7.729 to -7.400)	-7.456 (-7.598 to -7.309)	0.3221

$\mathbf{E}_{\mathbf{max}}$

	Variant E _{max} (95% CI)	WT E _{max} (95% CI)	Significance (P)
V45L	73 (65 to 82)	99 (93 to 107)	< 0.0001
P108A	101 (92 to 112)	102 (98 to 107)	0.8726
V172A	80 (73 to 89)	99 (93 to 107)	0.0007
L206V	134 (123 to 146)	102 (96 to 108)	<0.0001
A218T	79 (71 to 88)	102 (98 to 107)	<0.0001
G221S	88 (77 to 100)	99 (93 to 107)	0.0751
A238T	94 (83 to 107)	99 (93 to 107)	0.4363
G252A	112 (97 to 127)	102 (96 to 108)	0.2147
V281M	9 (4 to 18)	102 (98 to 107)	0.0174
E339K	26 (18 to 38)	102 (98 to 107)	0.0006
R376G	128 (121 to 137)	102 (96 to 108)	<0.0001

Results shown are point estimates and 95% confidence intervals from dose-response curves generated from three replicate experiments. Variant point estimates are shown next to point estimate from the WT control on the same plate. Statistically significant changes in log(EC50) or E_{max} are shown in bold [(extra sum of squares F test, P < 0.0045 ($\alpha = 0.05$ with Bonferroni correction for 11 comparisons)].

Table A.2.3 Oxytocin-induced β -arrestin-2 recruitment for wild type (WT) and variant OXTRs Log(EC50)

	Variant log(EC50) (95% CI)	WT log(EC50) (95% CI)	Significance (P)
V45L	-7.041 (-7.218 to -6.862)	-7.504 (-7.625 to -7.379)	<0.0001
P108A	-6.893 (-7.048 to -6.735)	-7.295 (-7.485 to -7.099)	0.0017
V172A	-7.206 (-7.378 to -7.030)	-7.504 (-7.625 to -7.379)	0.0050
L206V	-7.533 (-7.880 to -7.167)	-7.655 (-7.850 to -7.457)	0.5756
A218T	-7.349 (-7.538 to -7.153)	-7.295 (-7.485 to -7.099)	0.6795
G221S	-7.352 (-7.588 to -7.103)	-7.504 (-7.625 to -7.379)	0.2376
A238T	-7.392 (-7.603 to -7.169)	-7.504 (-7.625 to -7.379)	0.3463
G252A	-7.601 (-7.829 to -7.364)	-7.655 (-7.850 to -7.457)	0.7065
V281M	-7.287 (-7.612 to -6.940)	-7.295 (-7.485 to -7.099)	0.9767
E339K	-7.074 (-7.287 to -6.857)	-7.295 (-7.485 to -7.099)	0.1975
R376G	-7.570 (-7.740 to -7.395)	-7.655 (-7.850 to -7.457)	0.4945

1	7			
ı	Ľ	m	a	•

	Variant E _{max} (95% CI)	WT E _{max} (95% CI)	Significance (P)
V45L	87 (81 to 94)	98 (94 to 103)	0.0056
P108A	120 (111 to 129)	103 (95 to 111)	0.0039
V172A	93 (87 to 99)	98 (94 to 103)	0.1178
L206V	149 (131 to 169)	103 (95 to 110)	< 0.0001
A218T	94 (87 to 101)	103 (95 to 111)	0.0942
G221S	100 (91 to 109)	98 (94 to 103)	0.7994
A238T	105 (97 to 113)	98 (94 to 103)	0.1488
G252A	104 (95 to 113)	103 (95 to 110)	0.8088
V281M	26 (22 to 30)	103 (95 to 111)	<0.0001
E339K	51 (46 to 56)	103 (95 to 111)	<0.0001
R376G	113 (105 to 120)	103 (95 to 110)	0.0478

Results shown are point estimates and 95% confidence intervals from dose-response curves generated from three replicate experiments. Variant point estimates are shown next to point estimate from the WT control on the same plate. Statistically significant changes in log(EC50) or E_{max} are shown in bold [(extra sum of squares F test, P<0.0045 (α =0.05 with Bonferroni correction for 11 comparisons)].

Table A.2.4 Log(IC50)s for desensitization and internalization curves for wild type (WT) and variant OXTR.

Desensitization

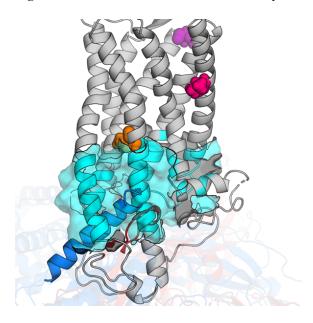
	Variant log(IC50) (95% CI)	WT log(IC50) (95% CI)	Significance (P)
V45L	-7.781 (-7.969 to -7.541)	-8.314 (-8.474 to -8.146)	0.0001
P108A	-7.840 (-8.054 to -7.530)	-8.414 (-8.490 to -8.336)	<0.0001
L206V	-8.414 (-8.609 to -8.200)	-8.427 (-8.541 to -8.310)	0.9040
V281M	-8.532 (-8.734 to -8.328)	-8.338 (-8.683 to -8.002)	0.3124
E339K	-7.828 (-7.976 to -7.654)	-8.355 (-8.518 to -8.180)	< 0.0001

Internalization

	Variant log(IC50) (95% CI)	WT log(IC50) (95% CI)	Significance (P)
V45L	-8.016 (-8.285 to -7.746)	-8.436 (-8.606 to -8.269)	0.0098
P108A	-7.965 (-8.199 to -7.732)	-8.559 (-8.756 to -8.356)	0.0003
L206V	-8.556 (-8.670 to -8.441)	-8.657 (-8.909 to -8.384)	0.4626
V281M	-8.571 (-9.381 to -7.769)	-8.704 (-9.169 to -8.229)	0.7595
E339K	-8.350 (-8.687 to -8.026)	-8.585 (-8.796 to -8.369)	0.2135

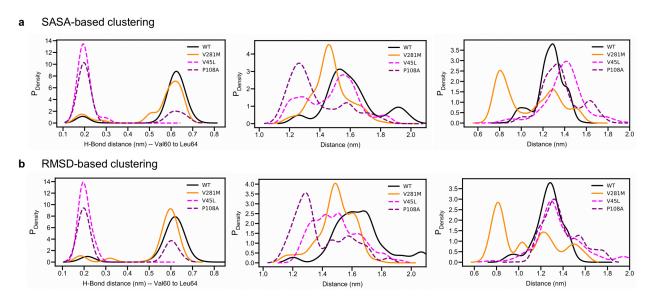
Results shown are point estimates and 95% confidence intervals (CI) from dose-response curves generated from three replicate experiments. Variant parameters are shown next to parameters from the WT control from the same experiment. *P* value shown from extra sum-of-squares F test comparing log(IC50) values between variant and WT.

Figure A.2.1 Atoms included in DiffNets analysis.



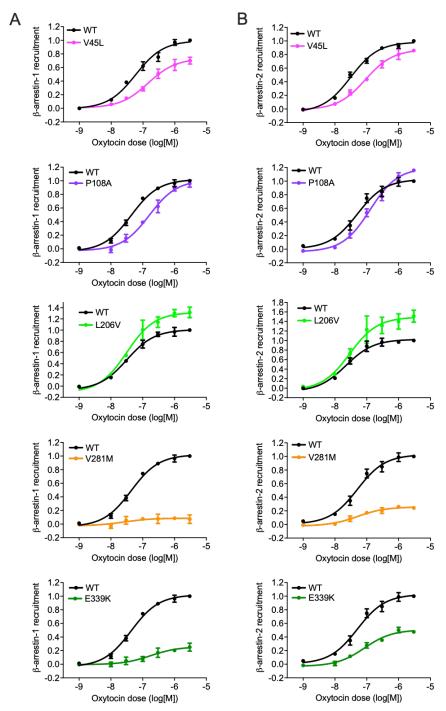
Atoms included in DiffNets analysis (cyan). OXTR homology model (grey) showing variants V45L (magenta), P108A (purple), and V281M (orange). Superimposed structures show β -arrestin-1 (red) and G protein (blue).

Figure A.2.2 Comparison of equilibrium properties calculated from simulations using different clustering methods.



The three distance distributions from Figure 6A, 6C, and 7A are replotted here from left to right using SASA-based clustering (a) and RMSD-based clustering (b). (a) and (b) are highly consistent suggesting that the choice of clustering used prior to MSM construction does not strongly affect the computed equilibrium properties of the system.

Figure A.2.3 Oxytocin-induced β-arrestin recruitment to wild type (WT) and variant OXTRs.



Dose response curves for β-arrestin-1 (A) and β-arrestin-2 (B) recruitment are shown for WT and variant OXTR. Error bars show standard error from N=3 independent experiments.

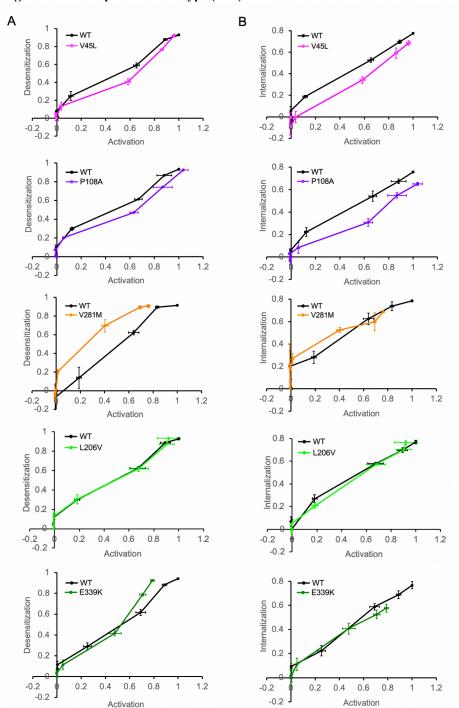


Figure A.2.4 Bias plots for wild type (WT) and variant OXTRs

Each point represents activation and desensitization (A) or internalization (B) for one oxytocin dose (10^{-12} - 10^{-6} M). Error bars are SEM from N=3 independent experiments.

A.3 Appendix to Chapter 4

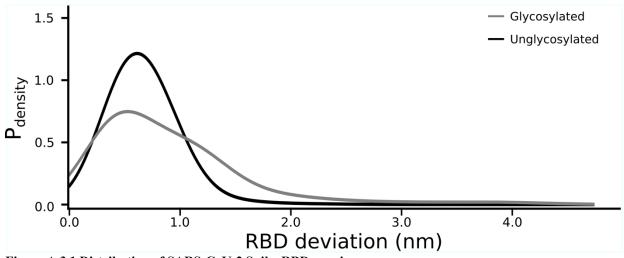


Figure A.3.1 Distribution of SARS-CoV-2 Spike RBD opening.The probability that the center of mass of an RBD deviates from its position in the closed (or down) state for SARS-CoV-2 spike with glycans (gray) and without glycans (black).

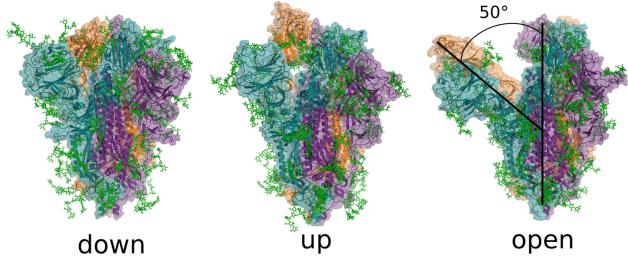


Figure A.3.2 Simulations of the SARS-CoV-2 Spike complex reveal the existence of an "open" state. For reference, three Spike complex snapshots are shown: the "down" state (6VXX), the "up" state (6VSB), and an "open" state from our simulations. Structures are depicted with a cartoon backbone, transparent surface for sidechains, and sticks for glycans. Each chain in the complex has a unique color, orange, purple, or teal, and glycans are colored green.

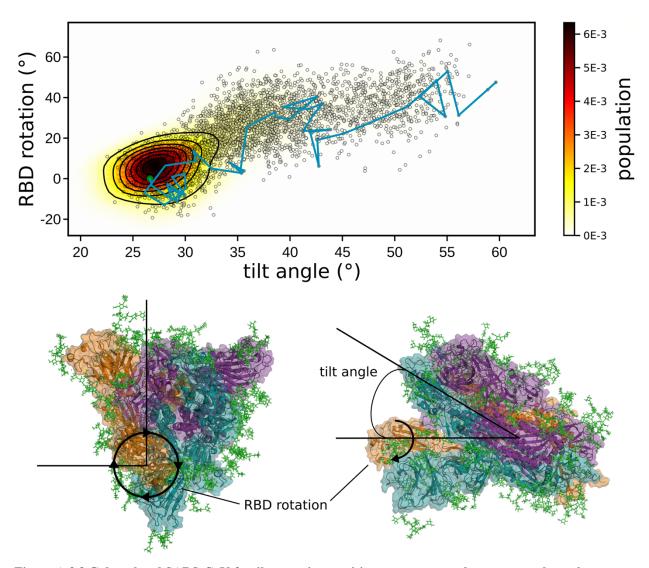
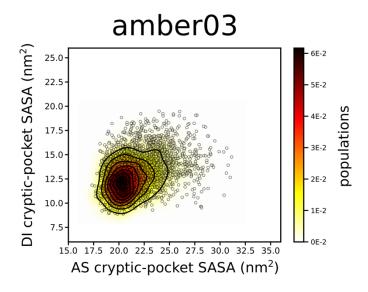
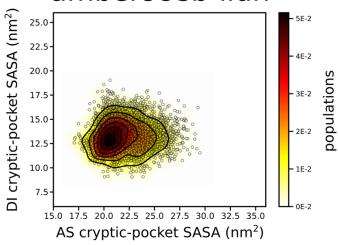


Figure A.3.3 Gylcosylated SARS-CoV-2 spike protein transitions to an extremely open state through simultaneous rotation of the RBD.

The Markov model of the spike protein is projected onto two order parameters: tilt angle and RBD rotation. Tilt angle is determined as the angle between points determined as the center of mass of the RBD, the helical S2 subunit, and the three RBDs when in the down position. The highest-flux transition pathway between the starting state and the state with the largest tilt-angle is shown in cyan.



amber99sb-ildn



charmm36

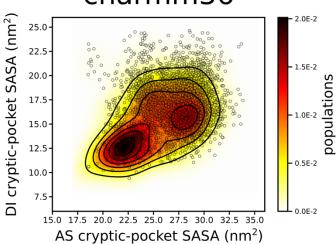


Figure A.3.4 The discovery of cryptic pockets on NSP5 is robust to the choice of forcefield.

FAST simulations of NSP5 were performed using the AMBER03, AMBER99sb-ildn, and CHARMM36 forcefields and projected onto the SASA of the active site and dimerization interface. For all three forcefields, we observe a cryptic pocket at the active site and dimerization interface.

Cryptic pocket highlights for select systems in the SARS-CoV-2 proteome

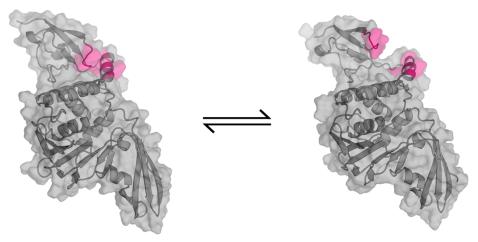


Figure A.3.5 NSP3-PL2Pro domain transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface (gray). The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in pink.

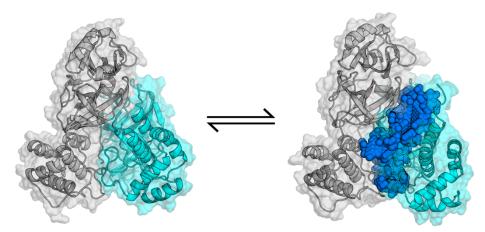


Figure A.3.6 NSP5 (dimer) transition from closed to open state.

Backbone is represented as a cartoon, sidechains are represented with a transparent surface, and pocket volumes are represented as blue spheres. Each molecule in the dimer is identified with a unique color, gray or cyan.

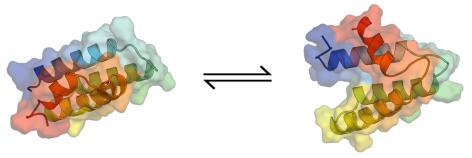


Figure A.3.7 NSP7 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The protein is colored by residue number following a rainbow.

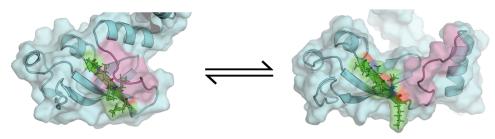


Figure A.3.8 NSP8 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. For reference, two regions that undergo a large conformational transition are highlighted as green and pink.

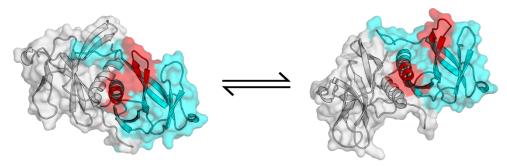


Figure A.3.9 NSP9 (dimer) transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. Each molecule in the dimer is identified with a unique color, gray or cyan. The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in red.

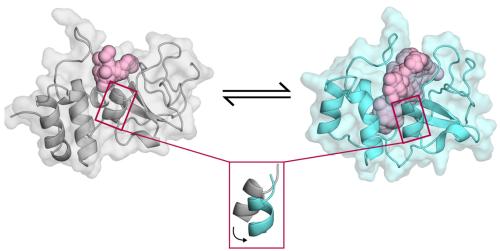


Figure A.3.10 NSP10 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. Pocket volumes are highlighted with pink spheres. Here, an existing pocket is greatly expanded from the swivel of an α -helix.

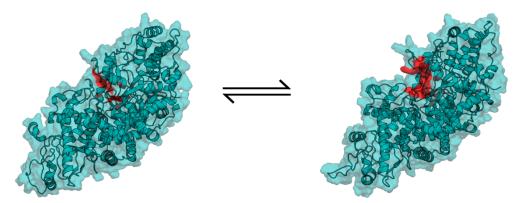


Figure A.3.11 NSP12 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in red.

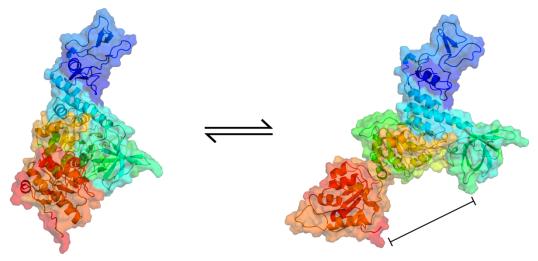


Figure A.3.12 NSP13 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The protein is colored by residue number following a rainbow and highlights the various domains. Here, we observe a large domain motion between domains 1A and 2A, which may be relevant for nucleotide binding.

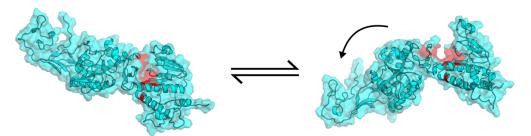


Figure A.3.13 NSP14 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in red.

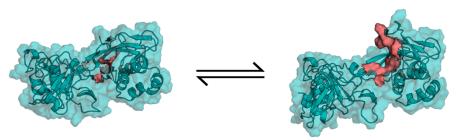


Figure A.3.14 NSP15 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in red.

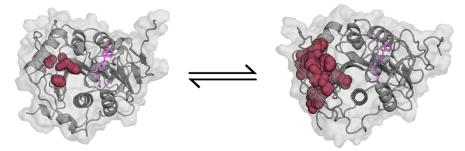


Figure A.3.15 NSP16 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. Pocket volumes are highlighted with maroon spheres. SAM cofactor is shown with pink sticks.

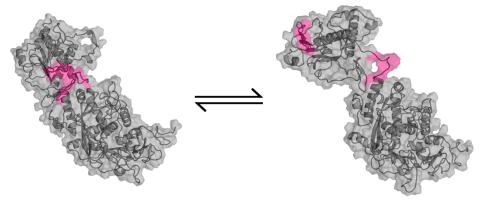


Figure A.3.16 NSP10/NSP14 (complex) transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in pink.

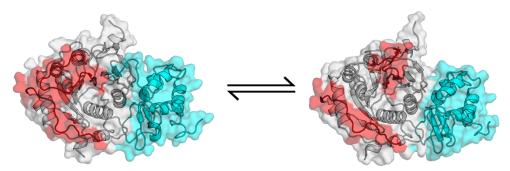


Figure A.3.17 NSP10/NSP16 (complex) transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. Each molecule in the complex is identified with a unique color, gray (NSP16) or cyan (NSP10). The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in red.

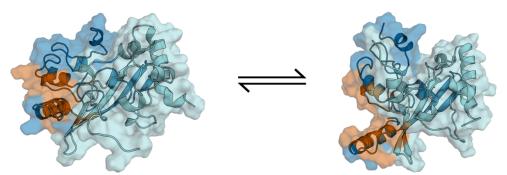


Figure A.3.18 Nucleoprotein dimerization domain transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in orange.

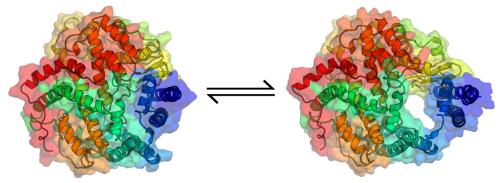


Figure A.3.19 Human ACE2 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The protein is colored by residue number following a rainbow. Pocket is proximal to the region that binds to SARS-CoV-2 spike protein.

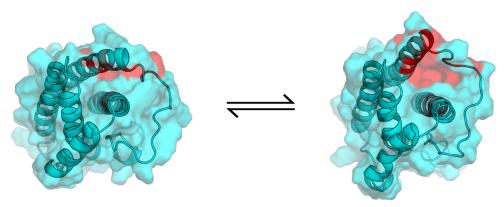


Figure A.3.20 Human IL6 transition from closed to open state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The residues that undergo a large conformational change to expose a cryptic pocket are highlighted in red.

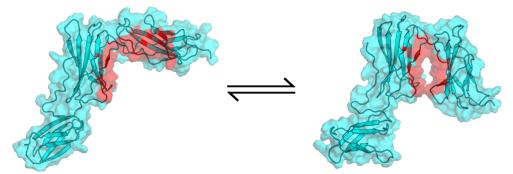


Figure A.3.21 Human IL6-R transition from expanded to closed state.

Backbone is represented as a cartoon and sidechains are represented with a transparent surface. The residues that undergo a large conformational change to reveal a potential druggable site are highlighted in red.

A.4 Appendix to Chapter 5

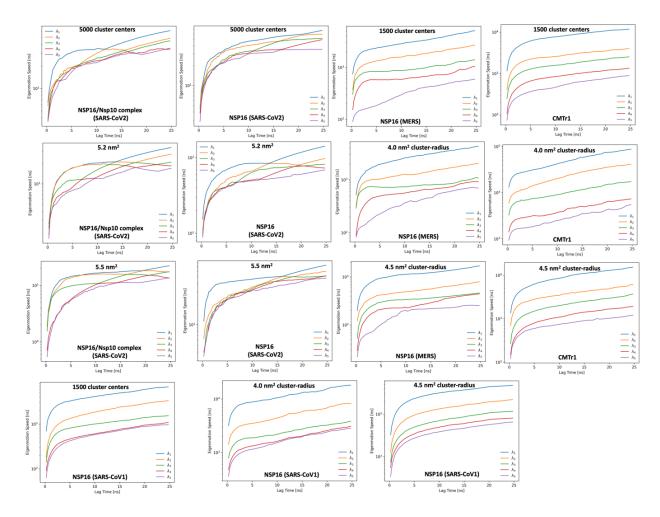


Figure A.4.1 Implied timescales plot.Implied timescales plotted as a function of the lag time for MSMs for Nsp16/Nsp10 complex (SARS-CoV2), Nsp16 homologs (SARS-CoV2, SARS-CoV1, & MERS) and human CMTr1 for different clustering cut-offs. 5 ns lag-time was selected to build the final MSMs used in this study.

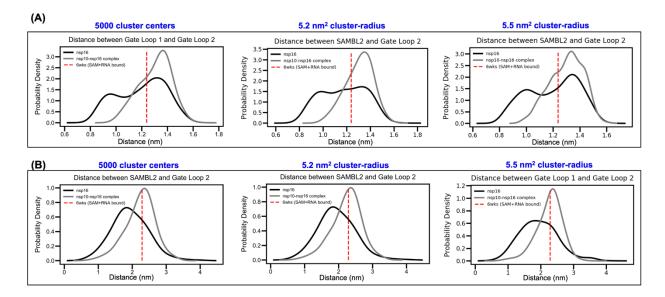


Figure A.4.2 Distance distribution replicates.

(A) Probability-weighted distance distribution between RNA-binding gate loops 1 and 2 comparing monomeric Nsp16 (black) to the Nsp10-Nsp16 complex (gray) are shown for three different clustering cut-offs. (B) Probability-weighted distance distribution between SAM-binding loop 2 and gate loop 2, comparing monomeric Nsp16 (black) to the Nsp10-Nsp16 complex (gray) are shown for three different clustering cut-offs. The distance for a SAM and RNA bound crystal structure is also plotted (red dotted line) in all figures.

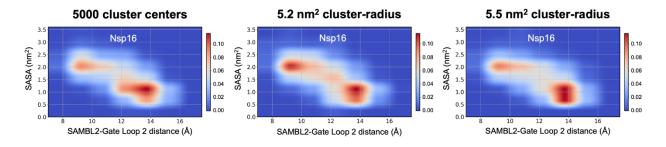


Figure A.4.3 SASA calculation replicates.

Equilibrium probability weighted 2D histograms of solvent-accessible surface area (SASA) of the cryptic pocket residues and the distance between SAMBL2 and gate loop 2 in Nsp16, derived from MSMs built with three different clustering cut-offs (5000 cluster centers, 5.2 nm² cluster radius and 5.5 nm² cluster radius)

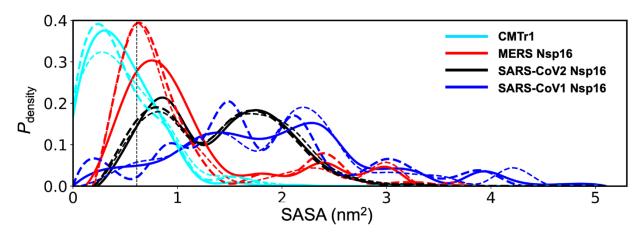


Figure A.4.4 Cryptic pocket opening distribution replicates.

Equilibrium probability-weighted distribution of the solvent exposure of pocket forming residues for SARS-CoV2 (black), SARS-CoV1 (blue), MERS (red) and CMTr1 (cyan). Solid lines show the distributions derived from MSM built on 5000 clusters (for SARS-CoV2 Nsp16) and 1500 clusters (for other homologs). Thick dashed lines show the distributions derived from MSM built on clustering with 5.5 nm² (for SARS-CoV2 Nsp16) and 4.5 nm² clusters (for other homologs). Thin dashed lines show the distributions derived from MSM built on clustering with 5.2 nm² (for SARS-CoV2 Nsp16) and 4.0 nm² clusters (for other homologs). Black dotted line depicts SASA of pocket residues in the crystal structure of Nsp16/Nsp10 complex (PDB: 6wks).

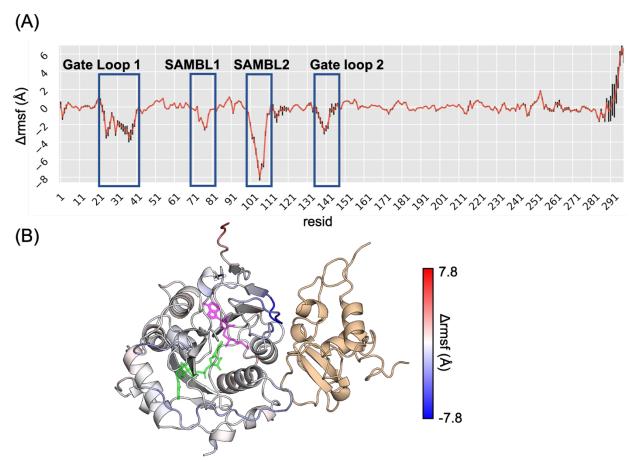


Figure A.4.5 Change in root mean square fluctuation (rmsf) of Nsp16 upon Nsp10 association.

(A) Probability weighted Δrmsf of Nsp16' residues upon Nsp10 binding is plotted. Negative values represent a decrease in rmsf upon Nsp10 binding. RNA binding loops (gate loop 1 and 2) and SAM binding loops (SAMBL1 and 2) are highlighted by the blue colored boxes. (B) Probability weighted Δrmsf of Nsp16 is mapped on its structure, with negative values shown in blue and positive values in red.

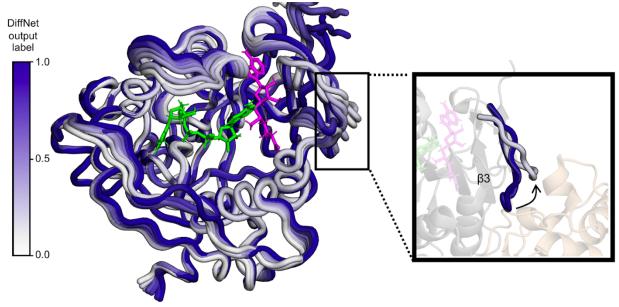


Figure A.4.6 DiffNets predict that $\beta4$ peels away from $\beta3$ in Nsp16 inactive structural states. (Left) Structural states changing from inactive to active (white to purple) as predicted by the DiffNet. (Right) The loop connecting $\beta3$ and $\beta4$ peels away from $\beta3$ into solution in predicted inactive states.

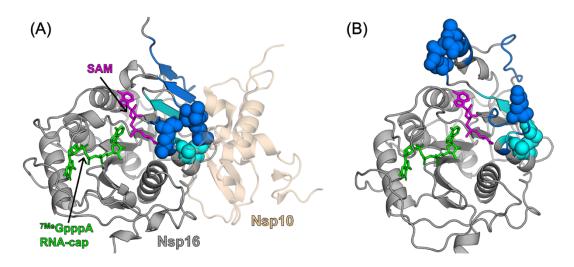


Figure A.4.7 Displacement of Nsp10 binding residues by cryptic pocket opening.

(A) Structure of Nsp16 in cryptic pocket closed state is shown in grey. Cryptic pocket forming residues and the residues undergoing opening motion are shown in cyan and blue, respectively. Cryptic pocket residues that contact Nsp10 are depicted in spheres. (B) Opening motion of the cryptic pocket shows the displacement of Nsp10 binding residues.

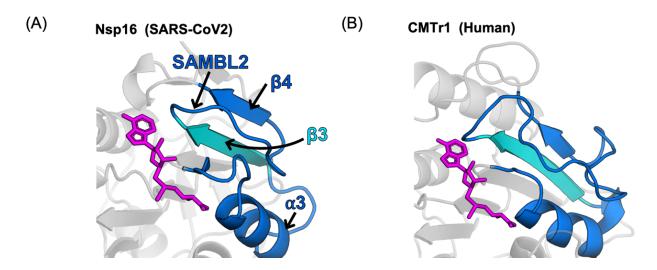


Figure A.4.8 Structural comparison of β 3- β 4 cryptic pocket in SARS-CoV2 Nsp16 and human CMTr1. (A) β 3 and residues lining the cryptic pocket in SARS-CoV2 are shown in cyan and blue, respectively. (B) Regions of human CMTr1, structurally equivalent to β 3 and the pocket lining regions are depicted in cyan and blue, respectively.

Table A.4.1 Timescales for transitioning between the pocket closed and open states in Nsp16 homologs.

8		
Nsp16 homolog	Transition time (microseconds)	Transition time (microseconds)
	'Closed' → 'Open'	'Open' → 'Closed'
SARS-CoV2	77.0	81.4
SARS-CoV1	26.0	13.5
MERS	19.5	9.9

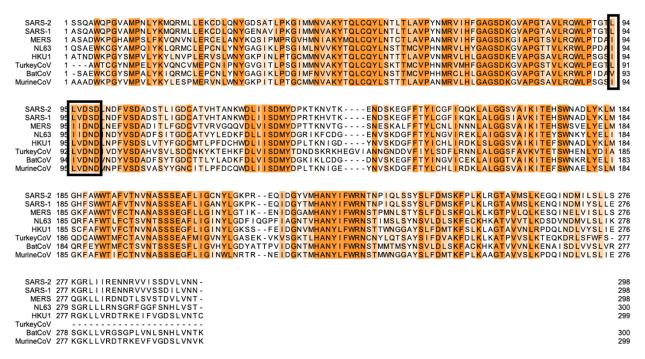


Figure A.4.9 Multiple sequence alignment of Nsp16 homologs from coronaviruses.

The color ranges from white to orange for the sequence conservation score ranging from 0 to 10, where 10 denotes 100% sequence identity. Residues of $\beta 3$ are enclosed in the black box. Uniprot ids of the sequences used for the alignment are given in the Methods section.

Curriculum Vitae

Michael D. Ward

Ph.D. Candidate, MolSSI Graduate Research Software Fellow Washington University School of Medicine Department of Biology and Biomedical Sciences

660 S. Euclid, St. Louis, MO 63110 St. Louis, MO

email: mdward@wustl.edu phone: 860-712-7612

EDUCATION

Washington University in STL, PhD Comp. Systems Biology, GPA: 3.82
Advisor: Prof. Gregory R. Bowman

University of Connecticut, MS Molecular and Cell Biology, GPA: 4.0
Advisor: Prof. Eric R. May

University of Connecticut, BS Biology, Minors: Chem, Bioinf., GPA: 3.65

Sept 2017-Present

RELEVANT SKILLS

- Extensive programming in Python, MATLAB and shell scripting
- Extensive experience with deep learning architectures (CNN/RNN/GNN) using PyTorch, TensorFlow
- Experience with large compute part of a small, core team with largest supercomputer in the world (Folding@Home, >1 exaflop). I was featured in a video with Microsoft CEO, Satya Nadella.
- Extensive experience with scientific computing libraries (Jupyter/Pandas/Scikit/etc.)
- Experience with distributed computing and HPC schedulers (slurm, LSF)
- Experience with multithreaded programming and MPI
- Version control using git/mercurial
- Unit testing (pytest)

CODE EXAMPLES

- Personal GitHub page
 - o https://github.com/Mickdub
- Freely available deep learning package (Main contributor)
 - o https://github.com/bowman-lab/diffnets
- Python package for clustering and building markov models (Contributor)
 - o https://github.com/bowman-lab/enspara
- Python package for predicting metabolism and toxicity of small molecules
 - Code samples available upon request

CAREER EXPERIENCE

Data Science Intern, Recursion Pharmaceuticals, Salt Lake City, UT

Sept 2021–Present

- Developed an algorithm to detect and filter out images with artifacts.
- Adapted SOTA convolutional neural networks (e.g. DenseNet161) to train on images of human cells to learn embeddings that encode biology in a way that is useful for finding new therapeutics.
- Used CNNs to filter out classes with noisy labels, which resulted in a dataset with a richer signal to learn from (i.e. multiclass classification accuracy improvement from ~15% to ~75% top1 accuracy).

Research Assistant, Washington University, St. Louis, MO

Sept 2017 – Present

- Developed a self-supervised learning approach combining deep learning and expectation maximization to detect features in protein simulations that explain how amino acid substitutions alter protein function [manuscript link] [code link]. Awarded MolSSI software fellowship for this work.
- Part of a small team that generated 100TB+ protein simulation datasets using our distributed computing platform Folding@Home. I was featured in a <u>video</u> with several CEOs including Satya Nadella (Microsoft), Jensen Huang (NVIDIA), and Lisa Su (AMD).
- Designed pipelines to automate analysis of 100TB+ protein simulation datasets which included clustering, constructing Markov state models, training with neural nets, automatic feature detection, etc. Author on three manuscripts [link], [link] published in *Nature Chemistry*, *Nature Communications*, and *Biophysical Journal*.
- Collaborated with engineers from Microsoft AI For Health using machine learning models to predict interactions between SARS-CoV-2 protein and human proteins in an effort to accelerate drug development [link]. Now, we are using graph neural networks (GNNs) and 3D-CNNs to aid in efforts to predict new druggable sites from protein crystal structures. I am currently adapting a state-of-the-art GNN [link] for this purpose.
- Combined tree-search algorithms and neural networks to predict how drugs get metabolized [link].

Research Assistant, University of Connecticut, Storrs, CT

Jan 2014-July 2017

- Quantified how nanoparticles flow in blood to learn how to target cancer cells for drug delivery [link].
- Evaluated new computational models of biophysical interactions between lipid membranes and proteins. Identified how to tune particular parameters of the model to make it more robust [link].

PUBLICATIONS

- 1. M. Malik, **M.D. Ward**, et al. Naturally Occurring Genetic Variants in the Oxytocin Receptor Alter Receptor Signaling Profiles. *ACS Pharmacology & Translational Science*. 4, 5, 1543-1555. (2021)
- 2. N.R. Flynn, **M.D. Ward**, et. al. Bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors. *Metabolites*. 11(6), 390. https://doi.org/10.3390/metabol1060390 (2021)
- 3. **M.D. Ward**, M.I. Zimmerman, A. Meller, M. Chung, S. Swamidass, G.R. Bowman. Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets. *Nature Communications* **12**, 3023 https://doi.org/10.1038/s41467-021-23246-1 (2021)
- 4. M.I. Zimmerman, J.R. Porter, **M.D. Ward**, et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chemistry* https://doi.org/10.1038/s41557-021-00707-0 (2021)
- 5. Neha Vithani*, **Michael D. Ward***, Maxwell I. Zimmerman, Borna Novak, Jonathan H. Borowsky, Sukrit Singh, and Gregory R. Bowman. SARS-CoV-2 Nsp16 activation mechanism and a cryptic pocket with pancoronavirus antiviral potential. *Biophysical Journal* https://doi.org/10.1016/j.bpj.2021.03.024 (2021) *Co-first authors
- 6. J. Cubuk, J.J. Alston, J.J. Incicco, S. Singh, M.D. Stuchell-Brereton, **M.D. Ward** et al. The SARS-CoV2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nature Communications* 12, 1936 https://doi.org/10.1038/s41467-021-21953-3 (2021)
- 7. M. Kshirsagar, N. Tasnina, **M. Ward** et al. Protein sequence models for prediction and comparative analysis of the SARS-CoV-2 —human interactome. *Biocomputing 2021*, pp. 154-165. DOI: 10.1142/9789811232701_0015 (2020)
- 8. N. Flynn, L.N. Dang, **M.D. Ward**, S. Swamidass. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation. *J. Chem. Inf. Model.* 60(7):3431-3449. DOI:10.1021/acs.jcim.0c00361 (2020)

- 9. **M.D. Ward**, S. Nangia, E.R. May. Evaluation of the Hybrid Resolution PACE Model for the Study of Folding, Insertion and Pore Formation of Membrane Associated Peptides, *J. Comput. Chem.*, 38(16):1462-1471, (2017)
- 10. Erik J. Carboni, B. Bognet, G. Bouchillon, A. Kadilak, L. Shor, **M.D. Ward**, Anson W.K. Ma. Direct Tracking of Particles and Quantification of Margination in Blood Flow. *Biophys J.* DOI:10.1016/j.bpj.2016.08.026. (2016)

PRESENTATIONS

Talks

- 1. **M.D. Ward***, A. Meller*, et al. "Predicting cryptic pocket opening from protein structures using graph neural networks". *NeurIPS* Machine Learning for Structural Biology workshop. Dec. 13, 2021.
- 2. **M.D. Ward.** "Deep learning the structural determinants of proteins biochemical properties by comparing structural ensembles with DiffNets". Science Friday, Department of Biochemistry and Molecular Biophysics. March 2021.
- 3. **M.D. Ward,** N. Vithani. "Hunting for opportunities to target SARS-CoV-2 via the Nsp16 protein". Folding@Home consortium invited talk. Twitch.tv [Link], February, 2021.
- 4. **M.D. Ward**, M.I. Zimmerman, A. Meller, M. Chung, S. Swamidass, G.R. Bowman. "DiffNets: Self-supervised deep learning to identify the mechanistic basis for biochemical differences between protein variants" May Lab Group Meeting. University of Connecticut, Storrs, CT. August, 2020.
- 5. **M.D. Ward.** "Leveraging deep learning to understand protein dynamics for applications in precision medicine". Thesis Proposal Exam. Feb. 2020.
- 6. **M.D. Ward**. "Elucidating the step-by-step formation of reactive metabolites that cause adverse drug reactions". CSB Qualifying Exam. Feb, 2019.
- 7. **M.D. Ward**, M.I. Zimmerman, G.R. Bowman. "Improving FAST using Google's PageRank Algorithm". BMB Science Friday. Washington University School of Medicine. March, 2018.
- 8. **M.D. Ward.** "Advantages and Limitations of Different Resolution Force Fields to Study Membrane Active Peptides". Master's Thesis Exam. July, 2017.
- 9. **M.D. Ward**, E.R. May, "A Viral Lytic Peptide Interacts with Membranes via Two Chemically Distinct Interfaces" UConn Graduate Seminar, Storrs, CT, Dec 4, 2015.
- 10. **M.D. Ward**, E.R. May, "Probing pH-Dependent Activity of a Viral Lytic Peptide" UConn Undergraduate Student Research Colloquium, Storrs, CT, May 1, 2015.

Posters

- M.D. Ward, M.I. Zimmerman, S.J. Swamidass, G.R. Bowman. "An Interpretable Deep Learning Approach to Detect Biophysical Effects of Protein Mutations". Center for Science and Engineering of Living Systems Fall Poster Session, Washington University in St. Louis, St. Louis, MO. November 11, 2019.
- 2. **M.D. Ward**, N. Flynn, R. Farmer, W. Pomerantz, G. Miller, S.J. Swamidass. "Assessing the liability of isoxazole containing compounds to form reactive metabolites". American Chemical Society National Meeting, San Diego, CA, August 25, 2019.
- 3. **M.D. Ward**, N. Flynn, T.B. Hughes, N.L. Dang, S.J. Swamidass. "Inferring Intermediates of Metabolic Reactions to Uncover Mechanisms of Toxicity" ACS SWRM 2018, Little Rock, AK.
- M.D. Ward, S. Nangia, E.R. May, "Evaluation of the Hybrid Resolution PACE Model for the Study of Folding, Insertion and Pore Formation of Membrane Associated Peptides" Biophysical Society 61st Annual Meeting, New Orleans, LA, Feb 15, 2017.
- 5. **M.D. Ward**, E.R. May, "Evaluation of the Hybrid Resolution PACE Model for the Study of Folding, Insertion and Pore Formation of Membrane Associated Peptides" North Eastern Structure Symposium (NESS), Farmington, CT, Oct 14, 2016.

- 6. **M.D. Ward**, E.R. May, "Probing the Function of Different Regions of a pH-Dependent Viral Lytic Peptide" 251st American Chemical Society National Meeting and Exposition, San Diego, CA, March 16, 2016.
- 7. **M.D. Ward**, E.R. May, "Probing pH-Dependent Activity of a Viral Lytic Peptide Using Molecular Dynamics Simulations" Frontiers in Undergraduate Research Poster Exhibition, Storrs, CT, April 11, 2015
- 8. **M.D. Ward**, E.R. May, "Replica Exchange Simulations of a Viral Peptide Interacting with an Implicit Membrane" UConn Molecular and Cell Biology Retreat, Bolton, CT, Sept 6, 2014.
- 9. **M.D. Ward**, E. Carboni, Anson W.K. Ma, "A Microfluidic Study of Nanoparticles in Simulated Blood Flows: Understanding the Effect of Margination" AIChE Northeast Regional Student Conference, Storrs, CT, April 6, 2014.

Journal Clubs

- 1. **M.D. Ward**. "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders" from Pacific Symposium on Biocomputing. Computational Molecular Biology Journal Club, Washington University School of Medicine, April, 2018.
- 2. **M.D. Ward**. "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer" from *Clinical Cancer Research*. Cancer Informatics Journal Club, Washington University School of Medicine, Nov, 2018.
- 3. **M.D. Ward**. "Deep Residual Learning for Image Recognition" from *CVPR*. Computational Molecular Biophysics Journal Club, Washington University School of Medicine, Jan, 2019.

HONORS, AWARDS, FELLOWSHIPS

2011-2015 Capitol Scholarship

2013-2014 New England Scholar

2015 Cum Laude, Biological Sciences

2017 NSF GRFP Honorable Mention

2019 ACS Chemical Computing Group Research Excellence Student Travel Award

2020 Google Cloud COVID-19 Research Grant

2020 Molecular Sciences Software Institute (MolSSI) Seed Software Fellowship

TEACHING EXPERIENCE

University of Connecticut

Teaching Assistant, Principles of Biology, BIOL 1107 4 credits, Fall 2015, co-taught lab portion with Kunal Dolas

Teaching Assistant, Introduction to Biochemistry, MCB 2000 4 credits, Spring 2016, taught lab portion

Washington University in Saint Louis

Teaching Assistant, Missouri Natural Heritage, BIO 2342 3 credits, Spring 2019