THE UNIVERSITY OF CHICAGO

A DYNAMICAL AND QUANTITATIVE APPROACH TO CHARACTERIZE NON-CANONICAL IMMUNE RECOGNITION EVENTS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
AND
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY
CHRISTOPHER THOMAS BOUGHTER

CHICAGO, ILLINOIS
MARCH 2021

To Elizabeth White, my Nana, a compassionate, caring, consistent role model

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES viii

ACKNOWLEDGEMENTS ix

ABSTRACT xi

LIST OF PUBLICATIONS BASED ON WORK IN THIS THESIS xii

1 INTRODUCTION 1
- 1.1 An Introduction to Adaptive Immunity 1
- 1.2 Canonical Forms of Immune Recognition 3
  - 1.2.1 Antibodies as Highly Specific Neutralizing Binders 4
  - 1.2.2 T Cells: Sentries of the Adaptive Immune System 5
- 1.3 Deviations from the Norm 7
  - 1.3.1 Polyreactivity in Molecules Designed for Specificity 7
  - 1.3.2 An MHC Independent Mechanism for T Cell Activation 9
- 1.4 Biophysical Approaches to Study Immunology 11
- 1.5 Goals and Potential Applications 13

2 BIOCHEMICAL PATTERNS OF ANTIBODY POLYREACTIVITY 15
- 2.1 Introduction: Understanding the Role of Polyreactivity in Natural Immune Responses 15
- 2.2 Generating a Novel Pipeline to Predict Polyreactivity 18
  - 2.2.1 Generation of a Polyreactive Antibody Database 18
  - 2.2.2 A Surface-Level Analysis of Polyreactive Antibody Sequences 19
  - 2.2.3 A Position Sensitive Matrix Representation of Sequences Provides Further Insights into Polyreactivity 23
- 2.3 Systematic Determination of the Key Contributors to Polyreactivity 27
  - 2.3.1 Linear Analysis Methods for the Identification of Polyreactivity . . . 27
  - 2.3.2 An Information Theoretic Approach 32
- 2.4 Molecular Dynamics Simulations Corroborate Informatics 35
- 2.5 Discussion 46

3 UNDERSTANDING THE ROLE OF BTN3A1 IN Vγ9Vδ2 T CELL ACTIVATION 52
- 3.1 Introduction: A Persistent Challenge to Our Understanding of T Cell Activation 52
- 3.2 Extracellular Conformational Change as a Driver of T Cell Activation 54
  - 3.2.1 Using Atomic Force Microscopy to Probe Conformational Change . . 55
  - 3.2.2 Dynamic Tests of BTN3A1's Extracellular Domain Stability 60
- 3.3 An Intracellular Model for pAg-Induced Activation 66
  - 3.3.1 Characterizing pAg Dynamics Within the B30.2 Binding Pocket . . . 67
  - 3.3.2 Investigating B30.2 Structural Changes in Response to pAg Binding . 71

3.3.3 Quantifying the Precise Contribution of pAg to Intracellular Dimerization . . . 75
3.4 Discussion . . . 80

4 COMPUTATIONAL DEVELOPMENTS IN BIOPHYSICS AND MOLECULAR IMMUNOLOGY . . . 86
4.1 Introduction: The Role of Physics and Computation in Immunological Research . . . 86
4.2 AIMS – An Automated Immune Molecule Separator . . . 89
4.2.1 Construction of a Bioinformatic Platform for Repertoire Analysis . . . 90
4.2.2 Linear Discriminant Analysis as a Tool for Repertoire Analysis . . . 91
4.2.3 Using Information Theory to Characterize Diversity and Crosstalk . . . 94
4.2.4 An Application to the Identification of MHC-Like Molecules . . . 96
4.3 A Refined Pipeline for Free Energy Calculations . . . 100
4.3.1 An Introduction to Free Energy Calculations . . . 101
4.3.2 Troubleshooting Toolkits for Implementation . . . 104
4.3.3 A Streamlined Method for Identifying the Most Probable Transition Pathway . . . 110
4.3.4 A Novel Approach for Optimizing Replica Exchange Performance . . . 115
4.4 Discussion . . . 118

5 CONCLUSIONS & PERSPECTIVES . . . 122
5.1 Expanding the Definition of Canonical Immune Recognition . . . 122
5.2 On the Biological and Biophysical Implications of Polyreactivity . . . 123
5.3 Towards Identifying the Mechanism Behind Vγ9Vδ2 T Cell Activation . . . 124
5.4 Perspectives on a Cross Disciplinary Interrogation of Adaptive Immunology 126

APPENDIX - MATERIALS & METHODS . . . 128

REFERENCES . . . 134

LIST OF FIGURES

2.1 A comparative genetic analysis of human-derived polyreactive and non-polyreactive antibody sequences . . . 21
2.2 Amino acid frequency in poly- and non-polyreactive antibody sequences . . . 22
2.3 Position-sensitive quantification of CDR loop properties of antibody sequences . 24
2.4 Quantification of biophysical properties across poly- and non-polyreactive antibody CDR loops . . . 26
2.5 Application of principal component analysis to antibodies tested for polyreactivity 28
2.6 Application of linear discriminant analysis to antibodies tested for polyreactivity 30
2.7 A visualization of the linear weights used for identification of polyreactive antibodies 31
2.8 An information theoretic analysis of antibody sequences . . . 33
2.9 Statistical significance of mutual information calculations . . . 34
2.10 Visualizations of differential charge localization on antibody surfaces . . . 36
2.11 A structural and dynamic analysis of antibody flexibility . . . 38
2.12 Quantification of antibody flexibility using RMSF . . . 40
2.13 Identification of structural motifs in polyreactive antibodies using tICA . . . 42
2.14 Identification of structural motifs in non-polyreactive antibodies using tICA . . 43
2.15 RMSD of polyreactive antibodies simulated for 1 μs . . . 45
2.16 RMSD of non-polyreactive antibodies simulated for 1 μs . . . 45
2.17 Quantification of mutual information of the parsed dataset . . . 49

3.1 A conformational change-based model for BTN-mediated pAg recognition . . . 55
3.2 Visualization of the two conformational states of BTN3A1 . . . 57
3.3 Atomic force microscopy scans of BTN3A1 in supported lipid bilayers . . . 59
3.4 Atomic force microscopy scans of BTN3A1 solubilized in detergent . . . 60
3.5 All-atom molecular dynamics simulations of BTN3A1's extracellular domains . . 62
3.6 Coarse-grained Upside simulations of BTN3A1's extracellular domains . . . 64
3.7 A new clustering-based model for BTN3A1-mediated Vγ9Vδ2 T cell activation . 67
3.8 Dynamics of pAg bound to BTN3A1's B30.2 domain . . . 69
3.9 Triplicate RMSD quantifications of pAg-B30.2 binding simulations . . . 70
3.10 Structural changes induced by pAg binding to BTN3A1's B30.2 domain . . . 72
3.11 Brute-force molecular dynamics simulations of B30.2 homodimers . . . 74
3.12 Visualization of the most probable transition pathway generated by the string method . . . 77
3.13 Converged energy landscapes calculated using the string method . . . 79

4.1 A visual schematic of linear discriminant analysis . . . 93
4.2 Comparisons between natural language and T cell receptor recognition motifs . 96
4.3 A structural comparison of HLA-A and CD1d . . . 98
4.4 Application of AIMS to MHC-like molecules . . . 99
4.5 Visualization of internal coordinate definitions for Euler angle calculations . . . 103
4.6 Instabilities introduced by the quaternion in protein-protein simulations . . . 106
4.7 Visualization of the toy model system . . . 107
4.8 Instabilities introduced by the quaternion in toy model simulations . . . 109

4.9 Single component tests of quaternion instabilities 110
4.10 Comparison between DHAM and averaging approaches for generating the most probable transition pathway 113
4.11 Visualization of novel angular restraints applied to proteins 115
4.12 A novel algorithm for the identification of gaps in replica exchange simulations . 118

LIST OF TABLES

2.1 A quantification of the antibodies used in this study. 19

ACKNOWLEDGEMENTS

Over the course of this journey through graduate school, I have had an overwhelming amount of support from an incredible network of mentors, peers, family, and friends. Starting first with this scientific network, I must thank my greatest supporters in Dr. Erin Adams and Dr. Benoît Roux. The constant mentorship, guidance, and scientific insight they provided throughout my entire graduate career has been invaluable. Erin and Benoît's contrasting mentorship styles proved to be perfectly complementary. Weekly meetings with Erin helped to formulate and fashion new ideas and approaches from the ground up, while monthly discussions with Benoît provided explosions of new avenues of inquiry. These interactions helped to shape me as a scientist and as an individual, and I cannot thank Erin and Benoît enough for all they have done for me.

In addition to these incredible mentors, my scientific endeavors have been greatly supported by a broader support network at the University of Chicago. Starting first with my thesis committee of Dr. Suriyanarayanan Vaikuntanathan, Dr. D. Allan Drummond, and Provost Ka Yee Lee, PhD. The committee has been instrumental in guiding me through the many ups and downs of my research. I thank them for all of the time they have committed to discussing my project with me, which has been critical to my growth as a scientist. While my project never fit quite so neatly into any individuals expertise, the committee nonetheless consistently provided insightful comments, questions, and concerns at every turn.

Next I need to thank the entirety of the Graduate Program in Biophysical Sciences. The administrators and students within the Biophysics Program genuinely make it feel more like a family than a graduate program, and I am incredibly proud to say I have been a part of this scientific family. It all starts from the top, and I would be remiss if I did not specifically call out Dr. Adam Hammond and Dr. Michele Wittels. These two incredible individuals are the glue that holds our program together, and I thank them for their mentorship and friendship over the past few years. Likewise, I need to thank all of the great leaders in BSAB,

and my entering cohort of friends and peers from the 2015 entering class.

The last of my scientific support network I must thank are all of the peers, collaborators, and secondary mentors that have helped to shape my graduate experience. In particular I thank my peers Dr. May Gu, Dr. Caitlin Castro, Dr. Kristof Nolan, Dr. Marta Borowska, Dr. Donghyuk Suh, Dr. Fabian Paul, Ryan Duncombe, Wioletta Nawrocka, and Nabil Faruk for insightful scientific discussions. I thank my collaborators Dr. Thomas Herrmann, Dr. Hugo Mouquet, Dr. Michel Nussenzweig, Dr. Patrick Wilson, and Dr. Albert Bendelac for their help in advancing some of the results discussed within this dissertation. I finally thank my incredible non-dissertation mentors Dr. Abby Stayart, Dr. Briana Konnick, and Dr. Justin Jureller for the profound impact they have all had in the way I view my scientific career.

Lastly, and perhaps most importantly, I must thank all of my friends and family that have supported me over the years. Having an occasional escape from science may be one of the most important ways of staying sane throughout graduate school. I must first thank my amazing parents, Lisa Bizzozero and Thomas Boughter, for their constant love and support. Next I thank the rest of my family who have always been there for me; Ami Boughter, Erika Boughter, Tori Bizzozero, Ashley Bizzozero, Paul Bizzozero, and Beth Rohrick. Finally, I thank all of my friends that have helped make my life incredibly fun and fulfilling; Dr. Jay Pittman, Dr. Will Riedl, Matt Reyer, Bryan Glick, Bryan Gorham, Joe Beaudet, Timmy Cruise, Mikey Mitton, Tomas Kerikas, Tommy George, Andy Vander Wyden, Dr. Zach Rokop, Josh Foster, Calvin Kuo, and Sean Gallup.

Everything I have accomplished up to this point is thanks in part to all of the people listed above. There have been many others who have supported me throughout the years, and I apologize that I did not have the space to thank every single one of them by name.

ABSTRACT

The adaptive immune system is a rich and complex field of study, with incremental improvements in our understanding of the fundamental machinations of some of its core components providing translational developments such as improved vaccines against infectious disease, drugs that can limit the severity of allergic reactions, and immunotherapies for treatment of a wide range of cancers. At the level of basic science, specific subsets of the adaptive immune response each have key questions spanning topics ranging from evolution and cellular communication to non-equilibrium thermodynamics and statistical mechanics. As such, investigations aimed at interrogating adaptive immunity require increasingly interdisciplinary approaches. In this thesis, we outline research at the interface of molecular immunology and computational biophysics, focusing on non-canonical immunological niches that break from the classical descriptions of adaptive immunity.

We first investigate the role of broad reactivity to diverse molecular species in antibodies, molecules that have long been suggested to be highly specific binders to single molecular targets. Through a novel bioinformatic approach, we are able to identify the critical molecular features that confer this broad reactivity, referred to as polyreactivity, in antibodies. Next, we turn our attention to uncovering an elusive activation mechanism of a specific subset of T cells, Vγ9Vδ2 T cells. Unlike the canonical $\alpha\beta$ T cells, these Vγ9Vδ2 T cells are activated independent of antigenic peptides or major histocompatibility complexes. Using a combination of computational approaches, we reassess a prominent model in the field and propose a new, clustering based model for activation. While these interdisciplinary approaches provide fundamental insights into complex biological phenomena, they also represent a powerful analytical framework for broader inquiries into molecular processes in immunology. Collectively, the approaches and computer code outlined herein should serve as a template to improve the pace of scientific discovery in this space.

LIST OF PUBLICATIONS BASED ON WORK IN THIS THESIS†

1. Gu, S., Sachleben, J.R., **Boughter, C.T.**, Nawrocka, W.I., Borowsak, M.T., Tarrasch, J.T., Skiniotis, G., Roux, B. and Adams, E.J. Phosphoantigen-Induced Conformational Change of Butyrophilin 3A1 (BTN3A1) and its Implication on Vγ9Vδ2 T Cell Activation. *Proceedings of the National Academy of Sciences*, **2017**, *114 (35)*, E7311–E7320.

2. Gu, S., Borowska, M.T., **Boughter, C.T.**, and Adams, E.J. Butyrophilins and $\gamma\delta$ T Cell Recognition. *Seminars in Cell and Developmental Biology*, **2018**, *84*, 65-74.

3. Fichtner, A.S., Karunakaran, M.M., Gu, S., **Boughter, C.T.**, Borowska, M.T., Starick, L., Noehren, A., Goebel, T.W., Adams, E.J., and Herrmann, T. Alpaca (Vicugna pacos), the First Non-Primate Species with a Phosphoantigen-Reactive Vγ9Vδ2 T Cell Subset. *Proceedings of the National Academy of Sciences*, **2020**, *117 (12)*, 6697–6707.

4. Castro, C.D, **Boughter, C.T.**, Broughton, A.E., Ramesh, A., and Adams, E.J. Diversity in Recognition and Function of Human $\gamma\delta$ T cells. *Immunological Reviews*, **2020**, *298*, 1-19.

5. **Boughter, C.T.**, Borowska, M.T., Guthmiller, J.J., Bendelac, A., Wilson, P.C., Roux, B., and Adams, E.J. Biochemical Patterns of Antibody Polyreactivity Revealed Through a Bioinformatics-Based Analysis of CDR Loops. *eLife*, **2020**.

†. The following chapters of this dissertation contain sections and figures adopted from the listed publications with modifications. Chapter 1: publications 1, 2, & 4; Chapter 2: publication 5; Chapter 3: publications 1, 2, 3, & 4; Chapter 4: publication 5.

CHAPTER 1

INTRODUCTION

1.1 An Introduction to Adaptive Immunity

When considering the most fundamental characteristics ascribed to living organisms, the maintenance of homeostasis, i.e. a steady, ordered internal state, stands out as a broadly defined trait. With this definition, we require that living organisms have the ability to make a concerted effort to thwart any environmental changes that induce significant deviations from normal day to day functions. Included in these environmental changes is the exposure of the organism to a vast array of pathogens, such as bacteria, viruses, and parasites. Generally, we consider the homeostatic responses to these invaders as a form of an immune response. Throughout evolutionary history, a trend of increasing complexity of these immune responses can be traced, moving from simple pattern recognition receptors to formidable repertoires of highly specific receptors. Prokaryotes utilize intracellular defense mechanisms such as the CRISPR/CAS system responsible for destruction of phage DNA [1], plants and insects each employ proteins with leucine rich repeats to recognize and neutralize pathogen [2, 3], and vertebrates make use of highly variable receptors to identify intruders with exquisite specificity. This last, complex form of immunity is referred to as adaptive immunity, and can be found in various forms across vertebrates from jawless fish to jawed vertebrates such as cartilaginous fish and humans [4, 5].

The adaptive immune system is comprised of an extensive assemblage of receptors designed to respond in a highly specific manner to foreign pathogens. No two adaptive responses to a novel pathogen are alike, with unique receptors generated, selected, and expanded each time the system is challenged. The primary cell subsets within the adaptive immune system, B cells and T cells, play non-degenerate roles in the body's response to infection. B cells' most important function lies in their ability to produce antibodies; large, secreted proteins

that bind directly to pathogens, blocking pathogenic function and identifying these invaders as targets for clearance. T cells have a multitude of complementary roles and responsibilities throughout the course of an immune response; participating in the direct killing of pathogens, the stimulation of B cells, and the activation of a variety of innate immune cells [6].

In order to carry out these functions, B and T cells must utilize cell surface receptors to directly recognize pathogen-derived molecular fragments, collectively called antigens. These antigens vary widely in form, from proteins, fats, and sugars to fragments of genetic material and foreign small molecules. However, at any given moment, the average healthy adult human has on the order of only 10^8 circulating lymphocytes that must be prepared to respond to all possible pathogenic antigens an individual will encounter in their lifetime [6]. Immediately, this raises a critical, and perhaps central, question of adaptive immunity: how can a finite number of circulating receptors recognize the near interminable space spanned by all possible molecular species that signal infection and dysregulation? The solution to this problem, in part, lies in the massive combinatorial diversity resulting from V(D)J recombination, a key step in the generation of any B or T cell receptor [7].

In the case of B cells, V(D)J recombination is merely the first step in the generation of receptor diversity. Upon recognition of target epitopes, i.e. the specific region of the pathogenic antigen, these antibodies undergo multiple rounds of somatic hypermutation and affinity maturation inside a germinal center, whereby the amino acid sequence of the epitope-binding surface is selected for optimal binding to the target [8–10]. The longer this affinity maturation process extends, the higher the affinity of the antibodies towards their target antigen, primarily through mutagenesis of the six complementarity determining region (CDR) loops of the antibody [8]. Using a combination of affinity matured CDR loops, these antibodies bind strongly to the target and aid in invader neutralization or act as signals to other components of the immune system.

T cells, conversely, rely solely on the process of V(D)J recombination to generate diverse

receptors, but generally recognize antigen in a significantly more restricted context. Rather than recognizing free antigen as antibodies do, a specific subset of T cells, $\alpha\beta$ T cells, recognize peptide presented by major histocompatibility complex (MHC) class I and class II molecules. These $\alpha\beta$ T cells are considered the canonical class of T cell, making up the majority of T cells circulating in the periphery. A second, less well studied subset of T cells, $\gamma\delta$ T cells, develop normally in mice independent of the presence of MHC [11, 12]. Instead, $\gamma\delta$ T cells recognize molecules that are structurally unrelated to MHC, and highly conserved across the human population. For either class of T cell, recognition of antigen by the T cell receptor leads direct killing of the target, the secretion of inflammatory molecules, or the recruitment of other cells to the site of the recognition event.

For the purpose of this work, it is helpful to group these forms of immune recognition even more broadly into processes considered either "canonical" or "non-canonical". In our "canonical" classes of adaptive immunity, we focus primarily on the life cycle of a standard antibody, and the behavior and activation mechanism of $\alpha\beta$ T cells. Conversely, we consider the binding of polyreactive antibodies to broad molecular species and the recognition of foreign pathogens by $\gamma\delta$ T cells as our "non-canonical" immune phenomena. Comparing and contrasting each helps to frame the big-picture questions in this work.

1.2 Canonical Forms of Immune Recognition

When considering the key functionalities of the adaptive immune system the average biologist, and potentially even the average immunologist, will think of two processes: the binding of an antibody to a single, specific molecular target, and the recognition of peptides bound to a major histocompatibility complex (pMHC) by a T cell receptor (TCR). While these processes have been comprehensively studied, a brief review of the critical mechanisms of recognition events will help to contextualize the work presently being considered.

1.2.1 Antibodies as Highly Specific Neutralizing Binders

Antibodies are arguably the most important component in a given adaptive immune response, capable of blocking pathogenic functionality or marking their target for clearance by other components of the immune system. Recently, antibodies have become entrenched in the popular lexicon; the presence or absence of anti-SARS-CoV-2 antibodies has become a prominent metric for identifying individuals suffering or recovered from COVID-19, the disease behind the 2020 global coronavirus pandemic [13]. In addition to the role of naturally derived antibodies in disease detection, antibodies designed and tested in laboratories can be used to cure disease. One such therapeutic antibody treatment, developed by Regeneron Pharmaceuticals, has shown success in neutralizing SARS-Cov-2 and reducing viral load [14]. Likewise, anti-cancer antibodies are the key components in many noteworthy immunotherapies, including the Nobel Prize-winning checkpoint blockade therapies [15–17].

In these therapeutic applications, drug developers must generate antibodies specific to their molecular target *de novo*, often through a combination of immunization strategies and rational design. This process of design and testing is incredibly expensive, and as such significant effort is currently being expended to assess the *a priori* "developability" of antibodies as therapeutics. The determinants of antibody developability have been investigated at length through experimental assays, in silico structural prediction-based methods, sequence-based analysis and their correlations with clearance, sequence-based SASA predictions, and sequence-based aggregation propensity predictors [18–22]. In studies focused on the performance of antibody therapies in clinical trials, poor specificity was seen to be a negative indicator of clinical success, in part due to the accelerated systemic clearance of intravenous antibody transfusions, suggesting that therapeutic antibodies should strive towards a drug-like specificity to achieve maximal developability [23–27].

Of note, this poor specificity is not correlated with decreased thermostability of the tested antibodies, suggesting that otherwise well-behaved antibodies that maintain the capability of binding to their primary target may still display reactivity towards unrelated mo-

lecular targets [21]. While this unwanted feature of therapeutic antibodies has plagued drug developers, researchers studying natural immune responses have stumbled upon a similar molecular feature of antibodies which they refer to as "polyreactivity". The precise definition of antibody polyreactivity and the potential benefits it confers to adaptive immunity are highlighted in Section 1.3.1.

1.2.2 T Cells: Sentries of the Adaptive Immune System

Whereas the antibodies secreted by B cells act primarily as a means of targeting pathogenic epitopes to neutralize threats, T cells have the dual responsibility of responding to pathogenic invaders and maintaining proper functionality of host tissues. These T cells utilize a cell surface receptor, called the T cell receptor (TCR), to surveil the relative health of cells within a given microenvironment. Broadly, T cells recognize three distinct classes of antigens; "self" antigens which signal normal cellular function, "non-self" antigens derived from pathogens, and "altered-self" antigens that can be found in cancers or other dysregulations of cellular function [28, 29]. For a vast majority of T cells, these antigens are not recognized alone, but instead are presented by a major histocompatibility complex (MHC). The TCR makes a multitude of strong contacts with both the peptide and the highly conserved platform domain of the MHC, in stark contrast to the highly variable epitope forms recognized by antibodies. This MHC restriction simultaneously simplifies our ability to control T cell recognition, while complicating our capacity to understand it.

The TCR-pMHC interaction displays such exquisite specificity that a simple peptide, 9-11 amino acids in length on average, provides the demarcation between a passive interaction and a fully-fledged immune response. This has massive implications in our ability to fight a wide array of cancers, through so-called "cancer vaccinations" [30,31]. Cancers arise through mutations that alter the function of regulatory proteins, which can lead to uncontrolled growth and spread of cancerous cells. When these proteins reach the end of their life cycle, they are marked for degradation, where some fragments of the protein, including the mutated

fragment, will be processed intracellularly and loaded into MHC molecules for presentation to T cells. This mutated peptide fragment from the cancerous cell is an example of an altered-self antigen, which are referred to collectively as neoantigens [29]. While some neoantigens can be recognized immediately by circulating T cells, others are less immunogenic, and need a boost from external sources, such as cancer vaccines. This process is prohibitively costly at present, due to the high level of personalized care required, but has proven successful in a multitude of cases [32].

Cancer vaccinations and the synthesis of novel peptides are exciting approaches to controlling the T cell response and utilizing it as a powerful therapeutic. However, there exist an abundance of shortcomings with targeting the TCR-pMHC interaction as a therapeutic target. First and foremost, the requirement for the personalization of each treatment is likely inescapable for any given disease. In humans, MHC molecules are highly diverse at the population-level [33]. This diversity means that not all engineered peptides will successfully bind to the MHC molecules present in all individuals. Human genetic diversity also complicates the identity of the peptide itself, independent of its ability to bind to MHC. Synthetic or cancer-derived peptides must be screened for close matches to peptide fragments in healthy tissues; if T cells raised against this novel peptide recognize healthy tissue, the resulting autoimmune responses can be disastrous [34, 35].

If these issues can be surmounted, either through a breakthrough in our ability to synthesize universally immunogenic peptides or a significant decrease in the costs associated with personalized therapeutics, additional complications arise in the fight against cancer. Any targeted treatment of cancers acts as an acute selection pressure for the cancerous cells. Cancers accumulate mutations and evolve on a rapid timescale, and as such centering treatments against a single biomarker may provide an avenue of escape for cancerous cells. Indeed, there have already been reported instances of cancers downregulating surface expression of MHC in response to neoantigen vaccines, rendering the TCR-pMHC based therapy obsolete [36]. To combat this rapid evolution of cancers, we must accumulate a formidable arsenal

of approaches, taking inspiration from less well studied corners of the immune system.

1.3 Deviations from the Norm

The outstanding questions in the study of these canonical forms of immune recognition have exceptional potential for the improvement of human health and our general understanding of the primary defense mechanisms of vertebrates. However, we are increasingly finding that there exist prominent exceptions to the neat rules of adaptive immune recognition outlined by previous decades of research. These less traditional mechanisms of adaptive immune recognition represent an untapped avenue for new therapeutic advances. Non-canonical forms of immune response represent the frontier of immunology, with researchers in these spaces redefining what is "known" about antibodies and T cells every few years. My own research has been primarily concerned with the characterization of the previously mentioned polyreactive antibodies, and the elucidation of a new mechanism for T cell activation found specifically in Vγ9Vδ2 T cells.

1.3.1 Polyreactivity in Molecules Designed for Specificity

While the process of affinity maturation and somatic hypermutation of antibodies results in high-affinity binders to a particular epitope, some antibodies have been shown to display signs of reactivity towards diverse off-target epitopes. This broad but low-affinity binding has been termed "polyreactivity". Importantly, polyreactive antibodies frequently display the same strong binding to their selecting antigen but have an additional ability to bind biophysically unrelated molecular species. This raises interesting questions: how can molecules that are explicitly selected for high affinity binding to a single target also bind completely unrelated ligands, and does this polyreactivity have any function in an immune response? Furthermore, could polyreactive antibodies that have not been explicitly selected for binding to a target ligand play a key role in the immune system?

Antibody polyreactivity has been hypothesized to be beneficial in the early stages of

antibody maturation, acting as a pool of diverse binders ready to recognize novel antigens and initiate the more stringent selection process [37]. To this end, a majority of B cell receptors and antibodies which have not undergone somatic hypermutation, including those on immature B cells and early "natural" antibodies, have been found to be polyreactive to some extent and are suggested to have an innate-like response to pathogens [38, 39]. While these mostly unmutated polyreactive antibodies remain at low frequency in antigen-experienced individuals, a distinct population of polyreactive antibodies that have undergone selection are still expressed by mature B cells that circulate in blood [40]. In fact, some studies have found the polyreactivity status of an antibody is mostly independent of the number of somatic hypermutations in the antibody sequence [41,42]. In line with this finding, only 5-10% of the repertoire of naive B cells circulating in the periphery are polyreactive, but this increases to 20-30% in the memory B cell compartment, showing a distinct capability of polyreactivity to survive selection [40, 43]. These results suggest that polyreactivity can persist, or perhaps even be selected for during the selection process within the germinal center. Further potential roles for polyreactivity will be discussed in Chapter 2.

Despite this potential positive role of polyreactivity in antibody development and maturation, the mechanism behind such polyreactivity remains a mystery. Generally, biophysicists often think of binding specificity being conferred through one of two distinct mechanisms, with interactions described by models of either conformational selection or induced fit. Conformational selection, similar to the "lock and key" conceptualization of protein-protein interactions, supposes that throughout the dynamic motion of a given protein one particular conformation will be a near perfect match to the target interface of the ligand [44,45]. In this model the two interfaces have neatly aligned charge-charge pairings, hydrophobic patches in similar locations, and complementary hydrogen bond donors and acceptors. Induced fit, on the other hand, assumes a protein interface which is flexible around the ligand interface, adopting a conformation not normally sampled in an unbound state in solution [44, 45]. The entropic costs of this conformational shift are then compensated by the significant increase in positive enthalpic contributions to the binding. These models are not mutually exclusive,

and protein-protein interactions likely often reside somewhere on a spectrum between these two extremes, meeting in the middle at a "conformational melding" model [46].

The "true" model that describes antibody binding has important implications into the mechanisms of antibody polyreactivity. If antibody binding is primarily described by conformational selection, then the diverse polyreactive ligands must overcome the delicate specificity of neatly arranged lateral organizations of charge, hydrophobicity, and polarity of a binding surface explicitly generated for a single target. Conversely, an antibody binding strongly to its target ligand using an induced fit mechanism would need to be highly flexible to accommodate the disparate structural features of polyreactive ligands. Understanding how polyreactive antibodies are able to recognize both their primary and polyreactive ligands has broad implications in understanding the role polyreactive antibodies play in an immune response, and in improving the design process of antibody therapeutics. My results outlined in Chapter 2 of this thesis provide a strong foundation upon which to further understand this biophysical process.

1.3.2 An MHC Independent Mechanism for T Cell Activation

While the binding of polyreactive antibodies to broad ligands represents a significant departure from our previous understanding of antibodies as specific binders, as outlined in Section 1.2.1, the function of these polyreactive antibodies remains similar. Initial forays into the identification of the function of antibody polyreactivity still center around their ability to neutralize pathogenic targets or identify molecules for clearance [47]. Conversely, $\gamma\delta$ T cells represent a significant departure from the classical picture of an $\alpha\beta$ T cell response discussed in Section 1.2.2. $\gamma\delta$ T cells, the second major lineage of T cells in the human adaptive immune system, are significantly less well studied than their $\alpha\beta$ TCR bearing counterparts.

In healthy individuals, approximately 5% of T cells circulating in the periphery bear $\gamma\delta$ TCRs [48, 49]. While this low abundance in the periphery may cause some to question their importance, these cells nonetheless play a key role in the adaptive immune response, making

up a significant percentage of intraepithelial T cells, reaching as high as 30% in sites such as the colon [50]. Additionally, high levels of tumor infiltrating $\gamma\delta$ T cells have been found to improve the prognoses of patients with colorectal cancer [51]. The differential prevalence and localization of $\gamma\delta$ T cells suggests a non-degenerate role in the adaptive immune response. While a multitude of putative ligands for $\gamma\delta$ T cells have been identified [52], the precise role of these cells remains unclear. In our attempts to elucidate this role, one specific subset of $\gamma\delta$ T cells, Vγ9Vδ2 T cells, have been the subject of intense scrutiny over the past decade. These Vγ9Vδ2 T cells are known to rely on a class of molecules called butyrophilins to be activated, but the precise mechanism of this activation remains a mystery.

Butyrophilin molecules are a broad class of single-pass transmembrane proteins bearing Ig-folded extracellular domains and a range of functional intracellular domains. Butyrophilins were first identified in the fat globules within milk isolated from bovine epithelial cells, providing the etymology of this protein, from the Greek butyros "butter fat" and philos "having affinity for" [53, 54]. Almost exactly 30 years later, a butyrophilin sub-family, BTN3A, was identified by Harly et al. and found to be a key mediator in the activation of Vγ9Vδ2 T cells [55]. Further work by the Adams lab showed that the intracellular domain of BTN3A1, one of the BTN3A family members, was able to coordinate binding to phosphoantigen (pAg), the known antigen of Vγ9Vδ2 T cells [56]. However, the downstream effects of this antigen binding event are still unclear. On the approach of the 10-year anniversary of this discovery, massive gaps in our understanding of the precise role of butyrophilins in the activation of T cells remain despite significant efforts by our lab and others.

While progress has been made in the structural elucidation of BTN3A1 and the identification of the key players in activation [56, 57], what role these molecular-level features play at the cellular scale is still unclear. BTN proteins share some structural similarity with co-stimulatory molecules, and in fact are members of the B7-superfamily of co-stimulatory molecules, yet butyrophilin-mediated T cell activation is TCR-dependent, i.e. the sequence of the specific Vγ9Vδ2 clone confers different levels of sensitivity to BTN [58]. What we

do know is that BTN3A is expressed ubiquitously across all tissues in humans, and that there are no known BTN3A polymorphisms [59]. This makes butyrophilins an appealing target for the development of co-treatments administered alongside other immunotherapies, providing a potential MHC-independent back door for T cell activation. Engineering such precise control over these processes requires a detailed understanding of the mechanisms of activation. My work aimed at progressing this understanding and challenging existing models is discussed at length in Chapter 3 of this work.

1.4 Biophysical Approaches to Study Immunology

While these non-canonical forms of immune recognition represent a significant challenge to our understanding of the immune system, overall we as a field have made substantial progress towards characterizing some of the key cellular and molecular players in other aspects of immune recognition and regulation. These recent results have laid a strong foundation for researchers from other fields to make substantial contributions to the study of the immune system. Increasingly over the past two decades, huge contributions to the field are being made by biochemists and biophysicists. Some of the most exciting outstanding questions in the field are inherently biophysical in nature, focused specifically on the single molecule level of immune interactions.

How do antibodies and T cell receptors carefully parse through diverse molecular species and identify immunogenic targets with extreme specificity? Once these molecules recognize these targets, how does the cell determine the appropriate response from the integration of binding strengths and molecular partners? How are MHC-like molecules loaded with antigen, and what determines the activating potential of peptide-MHC pairs? Each of these processes are determined entirely by their fundamental physical properties, as described by the fundamental theories of thermodynamics and statistical mechanics. By leveraging the application of these fields of study to biophysical processes in immunology, we can begin to make inroads into understanding the fundamental physical rules governing these recognition

events.

The most widely adopted tool to study the biophysical basis of molecular immunology has been the elucidation of the atomic structure of proteins using X-ray crystallography. The first protein crystal was solved in 1958, culminating in the award of the Nobel prize in 1962 to John Kendrew and Max Ferdinand Perutz for their discovery of the molecular structures of hemoglobin and myoglobin [60, 61]. Over the ensuing 60-plus years, a total of over 150,000 new structures have been solved and deposited in the Protein DataBase (PDB), making critical contributions to a wide range of biological fields [62]. While a powerful technique, these structural characterizations have a major shortcoming in their static depiction of highly dynamic processes. Proteins themselves are inherently dynamic, and these dynamics are key for their function [63]. To generate insights into the dynamic nature of crystallized proteins, we utilize a tool called molecular dynamics.

Many fields of physics rely on numerical simulation to generate insights that are difficult or impossible to probe experimentally. To probe the dynamical motion of proteins in an attempt to recreate the molecular movements of these molecules as they occur in vivo, biophysicists utilize molecular dynamics simulations, bringing static crystal structures to life. The first molecular simulation, an investigation into the side chain movement of bovine pancreatic trypsin inhibitor (BPTI) in 1977, only covered a mere 9.2 picoseconds of "real" time [64]. However, the foundations provided by this first BPTI simulation in vacuum paved the way for rapid advances in the study of molecular simulation, eventually leading to the establishment of molecular dynamics as a field unto itself, earning Martin Karplus a share of the 2013 Nobel Prize in Chemistry. Yet despite massive improvements in computational power over the past few decades, there remain limitations on the size, level of detail, and timescales accessible to molecular dynamics simulations.

These restrictions of computational power constrain the dynamical processes we are able to sufficiently sample. Immunology operates on timescales spanning nearly 17 orders

of magnitude, from the picosecond-scale motion of amino acid residues to the days-long (10^5 seconds) evolution of an adaptive immune response. When concerning ourselves only with the behavior of single molecules, we still must consider dynamic processes ranging from the picosecond- to seconds-scale. Yet given the constraints at present, we can sample only into the microsecond regime, or the millisecond regime if significant resources are available. In order to expand the scope of insights provided by molecular dynamics simulations, new techniques of enhanced sampling and advanced calculations are constantly adapted from fundamental physical theories and adapted for use in standardized simulation toolkits. In Chapter 4, I will discuss my contributions to improving these approaches and formalizing them into accessible protocols for all to use.

1.5 Goals and Potential Applications

B and T cells were discovered a mere 60 years ago [65, 66], and yet our understanding of each is already sufficient to engineer these cells for our own benefit. In just the past decade, we have created entirely new ways to cure disease using approaches such as adoptive cell therapy (ACT) and antibody therapeutics. However, our work on this front is far from over. Despite this significant progress, there remain a wealth of novel, poorly characterized pathways in immunology that question what we know about the rules of immune recognition.

Antibody polyreactivity and butyrophilin-mediated activation of Vγ9Vδ2 T cells contrast with the canonical forms of immune recognition that have been thought to be nearly fundamental tenets of the field. Antibodies have long been considered to function solely as highly specific binders to their molecular targets. Likewise, the activation of T cell receptors was long thought to be completely restricted to recognition of antigen in the context of MHC-like molecules. The discovery of polyreactive antibodies suggests there may be some new, added function of antibodies that has not been considered before. Vγ9Vδ2 T cells represent a similarly new model for expanding our understanding of immune recognition.

Throughout my thesis research, I have been focused on using a combination of experimental and computational approaches to dissect the molecular details of these non-canonical forms of immune recognition. This interdisciplinary approach has generated novel scientific insights into these immunological processes and new tools and approaches for further study of much broader forays into biochemistry and molecular biophysics. This dissertation will outline the major results from these studies and discuss their implications to their specific sub-fields and the broader fields of immunology and computational biophysics as a whole.

CHAPTER 2

BIOCHEMICAL PATTERNS OF ANTIBODY POLYREACTIVITY

2.1 Introduction: Understanding the Role of Polyreactivity in Natural Immune Responses

The mere mention of polyreactive antibodies can immediately introduce confusion into standard scientific discourse surrounding the nature of antibody-antigen interactions. Frequently, some posit that polyreactivity is merely the sign of an immature antibody that has not undergone rigorous selection for binding to an antigen. Others suggest that the poor-specificity of the antibody is the sign of an unstable secondary structure. As mentioned in Chapter 1, the antibody polyreactivity considered in this work is not explained by either of these hypotheses. The polyreactivity status of an antibody is independent of the number of somatic hypermutations in the antibody sequence [41, 42], and the thermostability of a given antibody correlates poorly with the polyreactivity [21]. Instead, it appears that polyreactivity has a key role in the adaptive immune response.

In addition to the benefits polyreactivity confers to antibody development discussed in Chapter 1, polyreactivity may in fact augment the efficacy of a given immune response by mature antibodies. Polyreactive IgA antibodies have been shown to have an inherent reactivity to microbiota in the mouse gut, with a predicted role in host homeostasis [67]. These previously identified antibodies so far have no known primary ligands yet play a key role in facilitating the gut immune response to the plethora of exogenous antigens encountered in the dynamic dietary and microbial environment of the gut. This implies the existence of antibodies whose primary function is to act as polyreactive sentries in the gut, yet the downstream effects of polyreactive antibodies coating commensal bacteria is so far unclear. Similar polyreactive IgA and IgG mucosal antibodies were found in the gut of human im-

munodeficiency virus (HIV) infected patients, but these antibodies either had low affinity to the virus or lacked neutralization capabilities [68]. The benefit of singular antibody sequences with the ability to sample large portions of the commensal population may represent an improvement in efficiency of the homeostatic machinery of the gut.

While the precise role of these primarily polyreactive gut antibodies is still a topic of debate, polyreactivity has been suggested to augment the immune response in other immunological niches. Broadly neutralizing antibodies (bnAbs), which bind robustly to conserved epitopes on the surface glycoproteins of influenza viruses or HIV are more likely to be polyreactive [69–71]. In one study of HIV binding antibodies, over half of all tested bnAbs were found to be polyreactive [72]. These bnAbs have been the subject of intense study for their potential as the central components of an HIV treatment or as the byproduct of an immune response to a universal Influenza vaccine [71, 73–75]. One hypothesized mechanism for the capability of polyreactive antibodies to confer this broad neutralization in the face of a changing viral epitope is heteroligation, the ability of a single antibody to bind the primary target with one binding domain and use the other binding domain to bind in a polyreactive manner [41]. This heteroligation allows the antibody to take advantage of the significant avidity increase afforded by bivalent binding, despite the low envelope protein density of HIV or a geometry which does not readily lend itself to bivalent binding on the surface of influenza viruses [76].

In accordance with this role of polyreactivity in responses to deadly diseases such as HIV and Influenza, as well as the therapeutic applications discussed in Chapter 1, many researchers have worked to identify the biophysical underpinnings of polyreactivity in natural immune responses. The most popular hypotheses for the primary biophysical predictors of polyreactivity have included CDR3 length [42], CDR3 flexibility [72], net hydrophobicity [77] and net charge [78]. More observational studies have found an increased prevalence of arginine and tyrosine in polyreactive antibodies [26, 79]. While these previous studies represent substantial advances in the study of polyreactivity, they have often been limited in scope,

focusing on a singular antibody source and primarily focused on CDR3H. Comparing across these individual antibody sources highlights discrepancies between the proposed predictors of polyreactivity. The aforementioned properties determined to be key to polyreactivity in previous studies were found to be statistically insignificant in studies of HIV-binding and mouse gut polyreactive antibodies [41, 67].

Clearly, a computational framework that would enable us to predict the polyreactivity of a given antibody *a priori*, whether evaluating the efficacy of a natural immune response or the potential fate of a therapeutic antibody, would be tremendously useful. Such a framework, for example, could be used to assist in the isolation of broadly neutralizing anti-viral antibodies, or speed up the process of therapeutic antibody screening. To achieve this goal, a thorough understanding of the molecular features behind polyreactive binding interactions is critical. Experimental approaches utilizing next-generation sequencing and ELISA allow for the identification of hundreds of polyreactive antibody sequences. However, the systematic characterization of these antibodies is difficult.

Issues immediately arise when defining the conditions by which we determine an antibody to be polyreactive. While polyreactivity may exist on some continuous spectrum, we are inclined to frame the problem as binary. This binary discretization is useful for the identification of meaningful differences yet must be recognized as an imperfect assumption. In addition to this more philosophical challenge, experimental efforts must also overcome significant hurdles. Detailed biochemical studies of polyreactive antibodies via protein crystallography, quantitative binding experiments, and mutagenesis provide exceptional insight but are inherently low throughput. Structural modeling of these polyreactive antibodies represent a high throughput approach, but models of flexible loops are relatively unreliable, and are unlikely to capture nuances in side-chain placement [80]. A bioinformatics-based approach, centered around high throughput analysis that minimizes structural assumptions while maintaining positional context of amino acid sequences would provide a thorough, unbiased analysis of existing data and create a powerful pipeline for future studies.

In this study, we show that, using just the amino acid sequences of antibodies from a database of over 1,000 sequences tested for polyreactivity, unifying biophysical properties that distinguish polyreactive antibodies from non-polyreactive antibodies can be identified. We find that, while charge and hydrophobicity are in fact important determinants of polyreactivity, the characteristic feature of polyreactive antibodies appears to be a shift towards neutrality of the binding interface. In addition, loop crosstalk is more prevalent in the heavy chain of polyreactive antibodies than non-polyreactive antibodies. From these properties, a machine learning-based classification software is developed with the capability to determine the polyreactivity status of a given sequence. This software is generalizable and can be re-trained on any binary classification problem, as will be discussed in-depth in Chapter 4.

2.2 Generating a Novel Pipeline to Predict Polyreactivity

2.2.1 Generation of a Polyreactive Antibody Database

Our aggregate database of over 1,000 antibody sequences is compiled from our own previously published and new data, in addition to data from published studies by the Mouquet and Nussenzweig labs (Table 2.1) [41, 67, 68, 70, 72]. Using an ELISA-based assay, the reactivity of each antibody is tested against a panel of 4-7 biochemically diverse target antigens: DNA, insulin, lipopolysaccharide (LPS), flagellin, albumin, cardiolipin, and keyhole limpet hemocyanin (KLH). This panel has become increasingly prevalent in the literature for experimental measures of polyreactivity in antibodies [21,41,42,67,68,70–72,81,82]. The ligands represent a diverse sampling of biophysical and biochemical properties; for example, enrichment in negative charge (DNA, insulin, LPS, albumin), amphipathic in nature (LPS, cardiolipin), exceptionally polar (KLH), or large in size (KLH, flagellin). From this panel, a general rating of "polyreactive" or "non-polyreactive" is given to 529 and 524 antibodies, respectively. For the purposes of this study, antibodies are determined to be polyreactive if the authors of the original studies determined a particular clone binds to two or more ligands in the panel. Those that bind to one or none of the ligands in the panel are deemed

non-polyreactive.

Dataset	Polyreactive	Non-Polyreactive	Total
Mouse IgA	205	240	445
HIV Reactive	172	124	296
Influenza Reactive	152	160	312
	529	524	1053

Table 2.1: A quantification of the antibodies used in this study.

A limitation of this full polyreactivity dataset is that there exists an intermediate between the two classes. As discussed in the introduction, it is not immediately obvious where the line for polyreactivity should be drawn. An antibody that binds to 2-3 ligands may not necessarily achieve broad reactivity through the same mechanism as an antibody that binds 4 or more ligands from a panel of 6 or 7. To remove these ambiguities, a so called "parsed" dataset is generated whereby antibodies that bind 4-7 ligands are labelled polyreactive, antibodies that bind 0 panel ligands are labelled non-polyreactive, and those that bind 1-3 are removed from the analysis. The results presented below utilize the full dataset of 1053 antibody sequences, unless otherwise noted.

2.2.2 A Surface-Level Analysis of Polyreactive Antibody Sequences

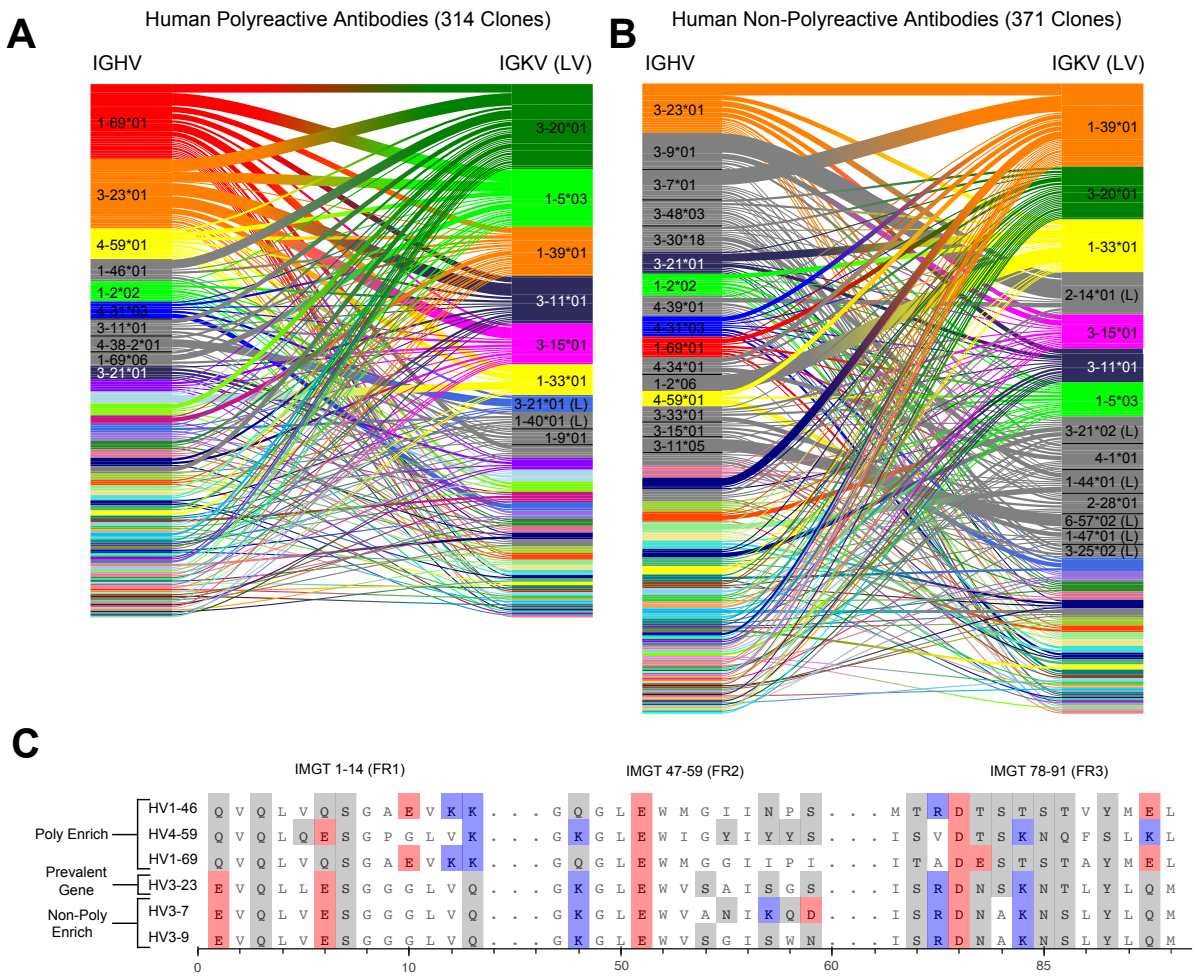
As a first pass at the given dataset, we focus on the most simplistic of the possible explanations for differences between polyreactive and non-polyreactive antibodies, specifically the J- and V-gene usage of each group. Figure 2.1A and 2.1B, rendered with code adapted from the Dash et al. derived program TCRdist [83], represents each antibody V-gene as a line connecting a single heavy and light chain gene for the full human-derived antibody dataset (685 sequences). Direct comparisons between mouse and human derived antibodies are difficult at the gene usage level, so the mouse data are analyzed separately (data not shown).

Genes are identified from nucleotide sequences using NCBI's IgBLAST command line tool [84]. Heavy and light chain genes that are shared between polyreactive and non-

polyreactive sequences are colored for the top labelled instances. Genes which are labelled but not found above a 2% threshold in the opposite population are colored grey, while those that do not have a visible name are colored randomly to highlight variation in gene usage. From this comparison, it is clear that the variable gene usage is skewed between polyreactive and non-polyreactive sequences, with an enrichment of V_H1-69, V_H1-46, and V_H4-59 in the polyreactive population, a trend that persists in the parsed dataset (data not shown). In contrast, no qualitative differences in the J-gene usage are readily discernible between these two groups (data not shown).

While the full alignment of these most used heavy chain variable genes shows a high degree of sequence similarity, Figure 2.1C highlights the regions of highest dissimilarity between the biophysical properties of amino acids in prevalent genes within each population. V_H3-23, the most prevalent gene in the non-polyreactive human dataset and the second most prevalent gene in the polyreactive human dataset, can be used as a reference for comparisons between genes enriched in each individual population. This reference gene shares a high degree of sequence similarity with the second and third most frequently occurring genes in the non-polyreactive dataset, V_H3-7 and V_H3-9, save for a lysine and aspartic acid pair in framework 2 of V_H3-7. The genes enriched in the polyreactive dataset, however, are quite different from this reference. All three of the polyreactive enriched genes have charged residues where the non-polyreactive enriched genes have hydrophilic residues (or vice versa) at IMGT positions 1, 13, and 90. These initial results hint at some systematic differences between the polyreactive and non-polyreactive antibody populations.

Figure 2.1D quantifies the extent of the difference in gene usage in each population by comparing these most prominent genes from our accumulated dataset of HIV- and influenza virus-reactive antibodies. While the two most common genes in the polyreactive dataset account for 27% of the human polyreactive antibodies in this study, the top three most common genes in the non-polyreactive dataset account for just over 17% of the total population.



D

Genes	Count Polyreactive	Percent Polyreactive	Count Non-Polyreactive	Percent Non-Polyreactive
HV3-23*01	41	13.06 %	29	7.82 %
HV3-9*01	3	0.96 %	21	5.66 %
HV3-7*01	5	1.59 %	18	4.85 %
HV1-69*01	44	14.01 %	12	3.23 %
HV4-59*01	18	5.73 %	9	2.43 %
HV1-46*01	13	4.14 %	5	1.35 %

Figure 2.1: **A comparative genetic analysis of human-derived polyreactive and non-polyreactive antibody sequences uncovers population level differences.** Gene usage diagrams comparing (A) human polyreactive and (B) non-polyreactive sequences show a qualitative difference in the VH gene usage. Shared colors indicate identical genes, grey indicates genes that are not seen in the other population at a level over 2%. Unlabeled genes are colored randomly to highlight genetic variation in the populations. (C) Sequence alignment of the most prevalent genes in the polyreactive and non-polyreactive populations compared to a reference gene common to each population. Hydrophobic amino acids are colored white, hydrophilic amino acids are colored grey, and positively or negatively charged amino acids are colored blue or red, respectively. (D) Percentage and raw count of observed gene usage for the polyreactive and non-polyreactive sequences.

In addition to being the most prevalent gene in the polyreactive dataset, V_H1-69*01 has also been found historically to be more prevalent in broadly neutralizing antibodies against influenza viruses, in line with the previously mentioned overlap between bnAbs and polyreactivity [71, 82].

Overall, there is a noticeable difference between the gene usage frequency of polyreactive and non-polyreactive antibodies, but the overlap in the usage of the two populations suggests that gene usage alone is not sufficient to distinguish the two groups. While there exist qualitative differences between framework sequences enriched in the polyreactive dataset compared to the non-polyreactive population, a look at the amino acid usage of the CDR loops of each group shows no significant differences (Figure 2.2). This implies that the positional context of a given amino acid is critical to tease out differences in antibody binding properties.

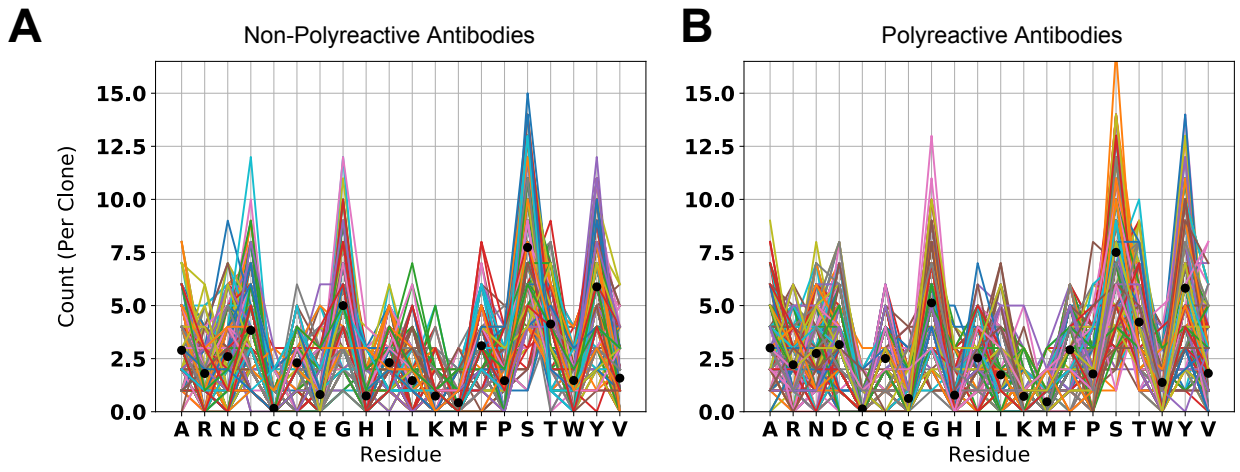


Figure 2.2: **The raw count of amino acids found in polyreactive and non-polyreactive antibody sequences shows no notable differences.** Amino acid usage plot highlighting the occurrence of each amino acid in non-polyreactive (A) and polyreactive (B) CDR loops. Each line represents an individual clone, and each point along the line represents the count of each amino in that given clone. Black dots represent the average counts per clone.

2.2.3 A Position Sensitive Matrix Representation of Sequences Provides Further Insights into Polyreactivity

To identify deeper trends in the biophysical properties of polyreactive antibodies, we utilize a new methodology to analyze and represent a range of different properties inherent to these sequences. While the framework regions of antibodies are highly conserved, the CDR loops vary significantly in length and show very low conservation between populations. This makes alignment of CDR loops difficult without creating subgroups for loops of identical length. To overcome this, the sequence data is re-organized into a matrix representation (Figure 2.3A). Each sequence is aligned by the center of each CDR loop, with spaces between the loops set to zero and each amino acid encoded as a number from 1 to 21.

While this alignment method excludes the framework regions of the antibodies and slightly averages out some of the properties at the edge of the CDR loops, we reason that most of these differences are evident in the gene usage analysis of the previous section. From this simple alignment, no obvious patterns emerge separating polyreactive and non-polyreactive antibodies, however we can clearly see that mouse gut-derived IgA antibodies have generally shorter CDR3H loops, and more conserved CDR3L sequences when compared to the human-derived antibody sequences. All subsequent analysis is derived from this matrix representation of the sequences.

With this new positionally sensitive and quantitative alignment method, we are able to further dissect the differences in amino acid sequences presented in Figure 2.1. Figure 2.3B uses this positional sequence encoding to determine the amino acid frequency difference between polyreactive and non-polyreactive sequences. For example, phenylalanine is found at position 93 in roughly 40% of polyreactive sequences and nearly 60% of non-polyreactive sequences. Therefore position 93, amino acid F has an intensity of -0.2 in Figure 2.3B.

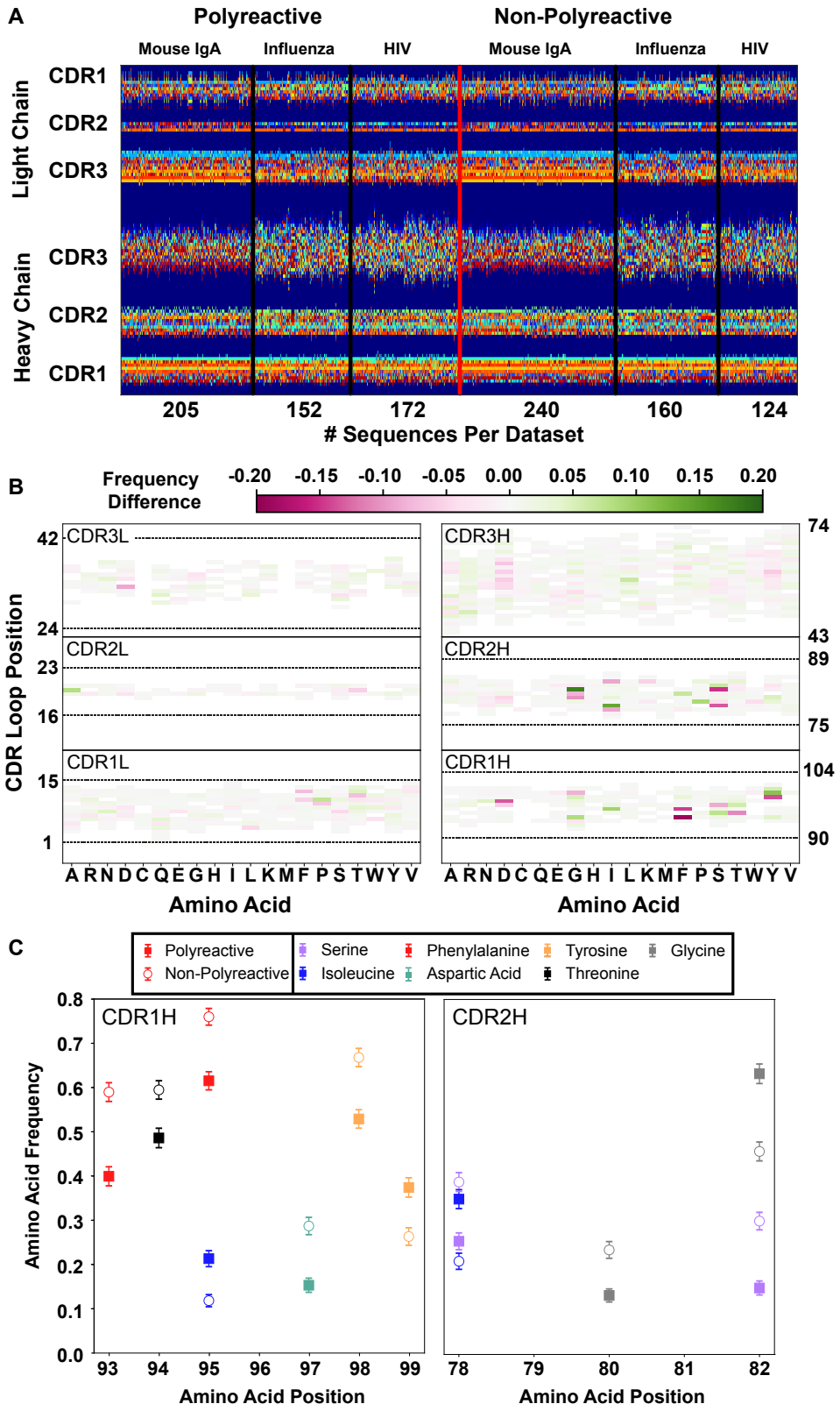


Figure 2.3: **A new representation of CDR loop sequences improves the position-sensitivity of quantitative antibody analysis.** (Caption continued on next page)

Figure 2.3 (Continued): (A) Matrix representation of the amino acid sequences used in this study provides a framework for further analysis. Each amino acid is encoded as a number from 1 to 21, represented by a distinct color in the matrix. A 0-value is used as a buffer between loops and is represented by the dark blue regions. The red line separates polyreactive and non-polyreactive sequences. (B) Amino acid frequency difference between polyreactive and non-polyreactive sequences for all six CDR loops. Residues more common in polyreactive sequences are shown in green, while those more common in non-polyreactive sequences are shown in pink. Loop positions correspond to the numerical position within the matrix of panel A. (C) An in-depth representation highlighting the amino acid frequencies used to create panel B. Only frequency changes greater than 10% are shown for clarity.

From Figure 2.3B, it is evident that most of the major differences are in the germline encoded regions CDR1H and CDR2H, in line with the observations from Figure 2.1 that suggest polyreactive antibodies have a distinct gene usage when compared to non-polyreactive antibodies. Figure 2.3C further expands on these differences, showing the largest changes in amino acid frequencies between the two populations. We can see that there is a slight decrease of phenylalanine frequency in CDR1H of polyreactive antibodies, in favor of isoleucine. Additionally, there is a general shift towards hydrophobicity in CDR2H, as the hydrophilic residue serine at matrix positions 78 and 82 is less prevalent in polyreactive antibodies, instead replaced by the more hydrophobic residues isoleucine and glycine. In the parsed dataset, these differences become larger in magnitude, particularly in CDR1L, where phenylalanine is again found less frequently in polyreactive sequences (data not shown).

This increased prevalence in loop hydrophobicity of polyreactive antibodies has been suggested before in the literature [72] along with a net increase in positive charge [78], so we next aimed to analyze this matrix systematically using biophysical properties inherent to the loops. A simple analysis of the full human and mouse-derived dataset investigating classical parameters explored previously by other groups (CDR loop length, net charge, net hydrophobicity, and gene usage) and some new properties (side chain flexibility, side chain bulk, and Kidera Factors [85]) show some significant differences between polyreactive and non-polyreactive antibodies (Figure 2.4A, B). The versatility of the positionally sensitive amino acid matrix allows for the application of multiple "property masks" to tease out the specific regions of each CDR loop that contributes most to these significant differences. Given

a property, amino acid charge for example, we can replace each simple 1-21 representation with a distinct representation based upon amino acid properties.

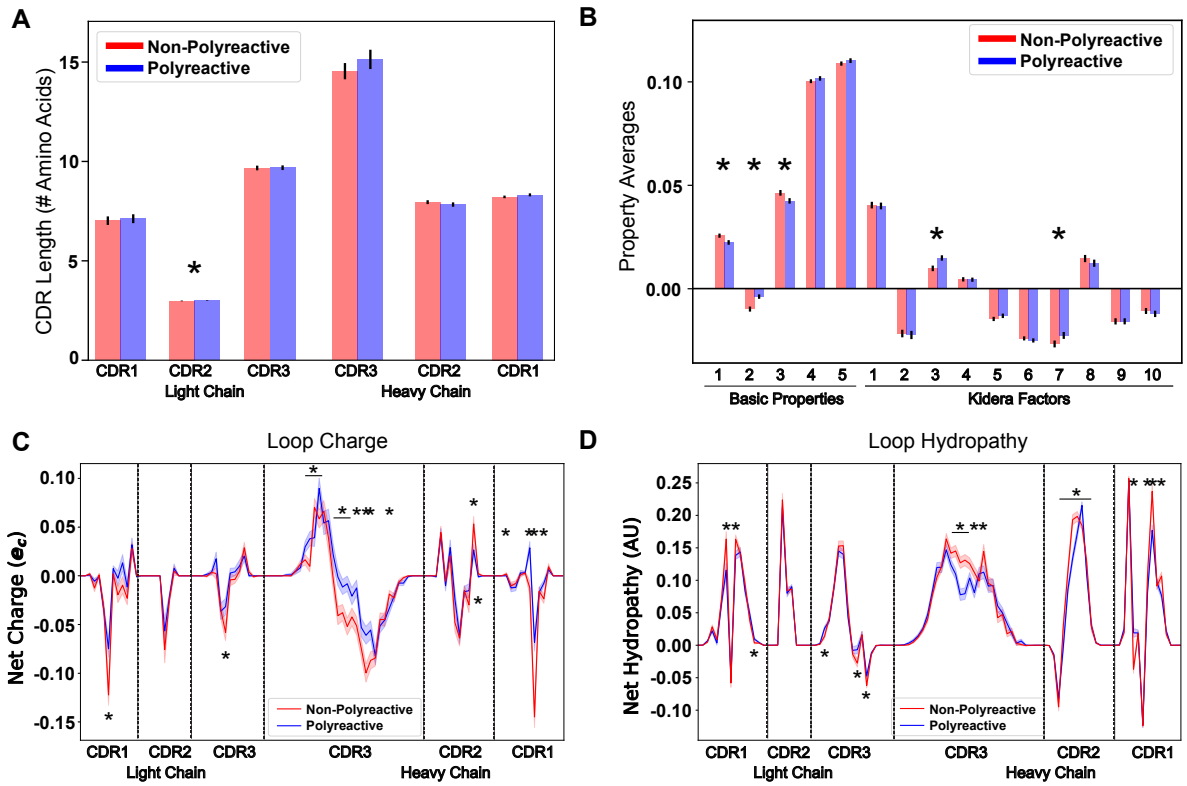


Figure 2.4: **Position-sensitive quantification of CDR loop properties of mouse and human antibody sequences highlights differences between polyreactive and non-polyreactive populations.** Plotting the average CDR loop lengths (A) and net antibody biophysical properties (B) show small but significant differences when analyzed in bulk. Basic properties 1-5 are hydropathy 1, charge, hydropathy 2, side chain flexibility, and side chain bulk. Plotting the average net charge (C) and hydropathy (D) as a function of position of polyreactive and non-polyreactive sequences highlights significant differences in CDR3H. Light shadow around lines represent bootstrap standard errors. All uncertainties obtained via bootstrapping. Stars indicate p-value ≤ 0.05 calculated via nonparametric Studentized bootstrap test. Bars with a single star above represent contiguous regions of significance. p-values in panels (A) and (B) corrected for multiple tests using the Bonferroni correction.

In the matrix of Figure 2.3A leucine, histidine, and arginine are represented by the integers 3, 16, and 17. As an example, when the charge property mask is applied, the matrix representations of these three amino acids in all sequences is changed to 0.00, 0.091, and 1.00, respectively. We apply 62 such masks to this matrix, including simple metrics like charge, hydropathy, side chain flexibility, and side chain bulkiness to go along with more

carefully curated metrics from the works of Kidera et al. and Liu et al [85, 86]. A complete description of these properties can be found within the extended data provided in the GitHub distribution.

The application of these masks gives an entirely new matrix describing the localization of amino acids with a given property. By averaging across all sequences in the polyreactive or non-polyreactive dataset when these masks are applied, we can readily see differences in charge patterning and hydropathy when comparing polyreactive and non-polyreactive sequences (Figure 2.4C, D). Including errors obtained via bootstrapping, we see that these differences are most pronounced in the center of CDR3H, with some differences also apparent in the remaining five loops.

This analysis shows an overall bias towards neutrality in these regions; i.e. neither positively nor negatively charged, neither strongly hydrophilic nor hydrophobic. These results also contextualize the findings of Figure 2.3C. The trend towards hydrophobic residues in CDR2H of polyreactive antibodies importantly does not make these regions net hydrophobic, but instead make these regions slightly less hydrophilic on average. This effect is yet again more pronounced in the parsed dataset (Figure 2.4D,E), with a strong trend towards interface neutrality. Conversely, when comparing bootstrap samples drawn from the null distribution, i.e. the "polyreactive" or "non-polyreactive" labels are given to antibody sequences at random, we see no difference between the biophysical properties of the two populations (data not shown).

2.3 Systematic Determination of the Key Contributors to Polyreactivity

2.3.1 Linear Analysis Methods for the Identification of Polyreactivity

Along with simple property averaging, these masks also give a high dimensional space from which we can determine, in an unbiased way, the primary factors that discriminate

polyreactive and non-polyreactive antibodies. As a first pass, we apply a principal component analysis (PCA) to the matrix of all antibody sequences in an attempt to separate the polyreactive or non-polyreactive populations along the axes of highest variation in the dataset. Unfortunately, the principal components of these data do not effectively distinguish between the two populations (Figure 2.5).

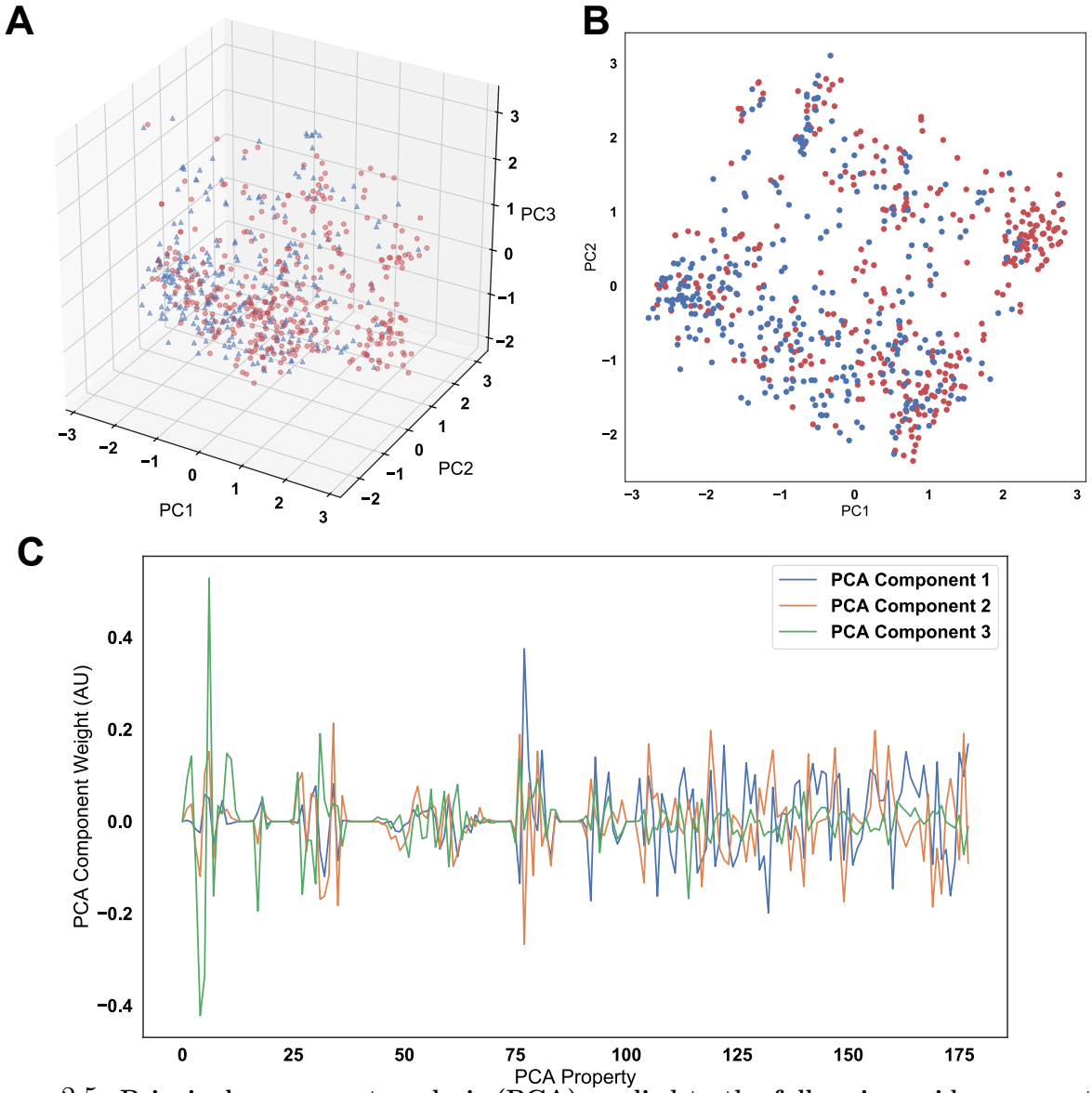


Figure 2.5: **Principal component analysis (PCA) applied to the full amino acid usage matrix and the top 75 discriminating vectors used for linear discriminant analysis.** The analysis shows an inability to distinguish the two populations when showing the first three (A) and first two (B) principal components. (C) Examination of the weights of these first three components shows there is no one property disproportionately contributing to the variance in the dataset. The vector normal of each set of weights is equivalent to 1.

To further investigate the physical and sequence-based properties of polyreactivity in antibodies in a more targeted manner, we employ linear discriminant analysis (LDA), a common algorithm often applied in classification problems [87–89]. Further discussion of linear discriminant analysis can be found in Chapter 4. Figure 2.6A shows the results of LDA when applied to the parsed dataset comprised of 311 polyreactive antibodies and 362 non-polyreactive antibodies. As discussed in the introduction, the framing of polyreactivity as a binary problem is not a perfect assumption. The inclusion of intermediate levels of polyreactivity further confounds this issue. Indeed, the application of LDA to the full dataset shows a reduced ability to split polyreactive and non-polyreactive antibodies (data not shown), likely due to this spectrum of polyreactivity. By considering only the parsed dataset for these classification analyses, we can improve confidence that the differences identified are those that separate strongly polyreactive and strongly non-polyreactive antibodies.

LDA can be operated in two distinct modes; one which is capable of identifying key discriminating properties, and one that functions as a canonical classification algorithm. The distinctions between these two modes are discussed in Chapter 4. In the first mode, we find that the data can be split more effectively when the parsed dataset is broken up into the distinct "reactivity" groups, i.e. those antibodies specific for influenza viruses, HIV, or found in the mouse gut (Figure 2.6A). This suggests there may be some bias due to antigen specificity, or lack thereof, whereby influenza virus-specific antibodies take a slightly different path towards polyreactivity compared to HIV reactive or mouse gut IgA antibodies. However, when using the classification mode, the classification accuracy is roughly equivalent across all tested datasets (Figure 2.6B). Testing this classifier with a scrambled dataset, where the labels are randomly assigned, shows the expected decrease in classification accuracy for each individual dataset for all ranges of input features.

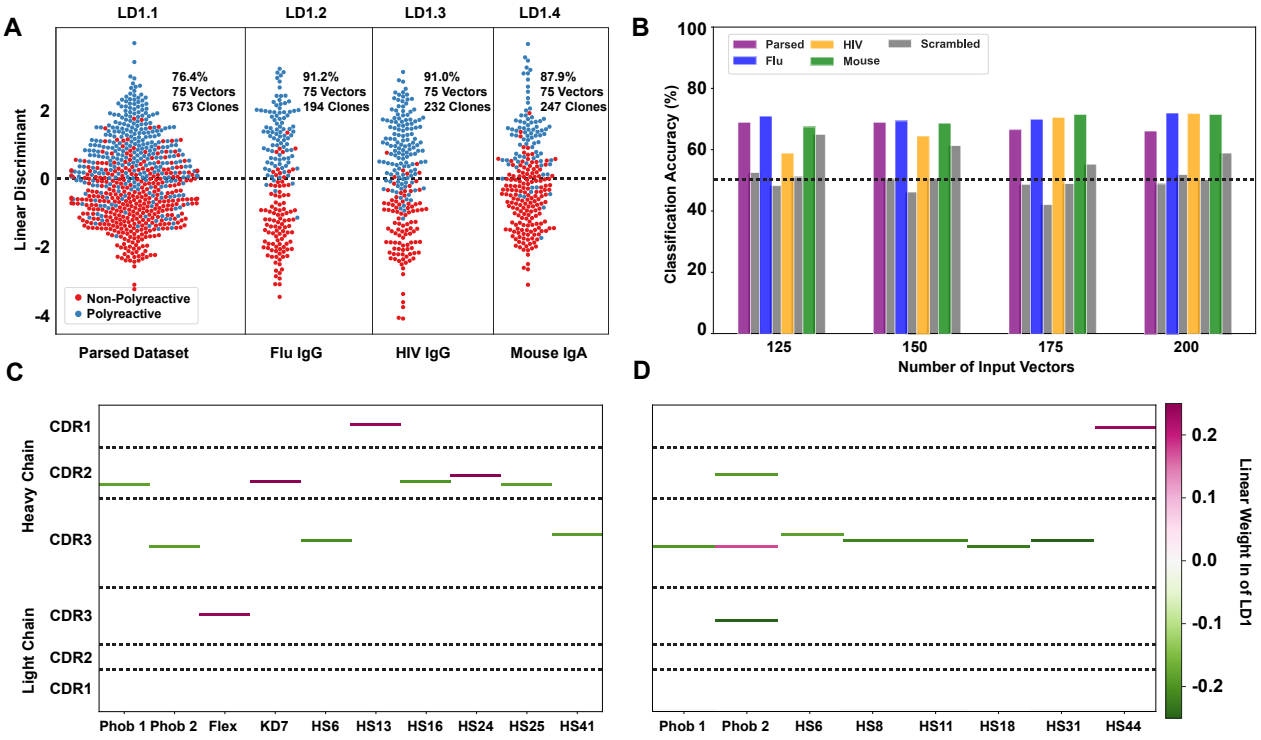


Figure 2.6: **Linear discriminant analysis (LDA) can meaningfully separate the two populations and these meaningful differences can be used to generate a polyreactivity classifier.** LDA applied individually to the complete parsed, Influenza, HIV, and mouse datasets. Percentages indicate the accuracy of the linear discriminant in labelling polyreactive and non-polyreactive antibodies. For these data, the plotted linear discriminants are comprised of different linear weights. (B) Accuracies of a polyreactivity classifier with a separate test and training dataset. Groupings in this figure are the same as those in panel A. A support vector machine is generated for each individual population, and the reported values are accuracies calculated through leave one out cross validation. Shown are test data and a scrambled dataset where the labels of "polyreactive" or "non-polyreactive" are applied randomly (grey bars). The dotted line indicates 50% accuracy threshold. (C) Property matrices highlighting the top 10 weights of the linear discriminants in panel A for the parsed dataset with 75 vectors (C) and the HIV dataset with 75 vectors (D). Color bar represents the normalized weight of each property, where pink rectangles represent properties correlated with increased polyreactivity, and green rectangles represent properties correlated with decreased polyreactivity.

When applying LDA in the first mode (Figure 2.6A), we can directly pull the linear weights of each component comprising linear discriminant 1 and reveal which biophysical properties at each CDR position best distinguish between the two populations. The differences in the linear weights from the heavy chain CDR loops comprising each discriminant show clear differences when comparing the complete parsed dataset (Figure 2.6C) to the HIV only dataset (Figure 2.6D). In the parsed dataset, the discriminating weights are heavily concentrated in CDR2H. Whereas in the HIV dataset, these weights are centered around

the CDR3H loop. Only the top ten linear weights are shown in Figure 2.6C, D. The full matrix of linear weights can be found in Figure 2.7.

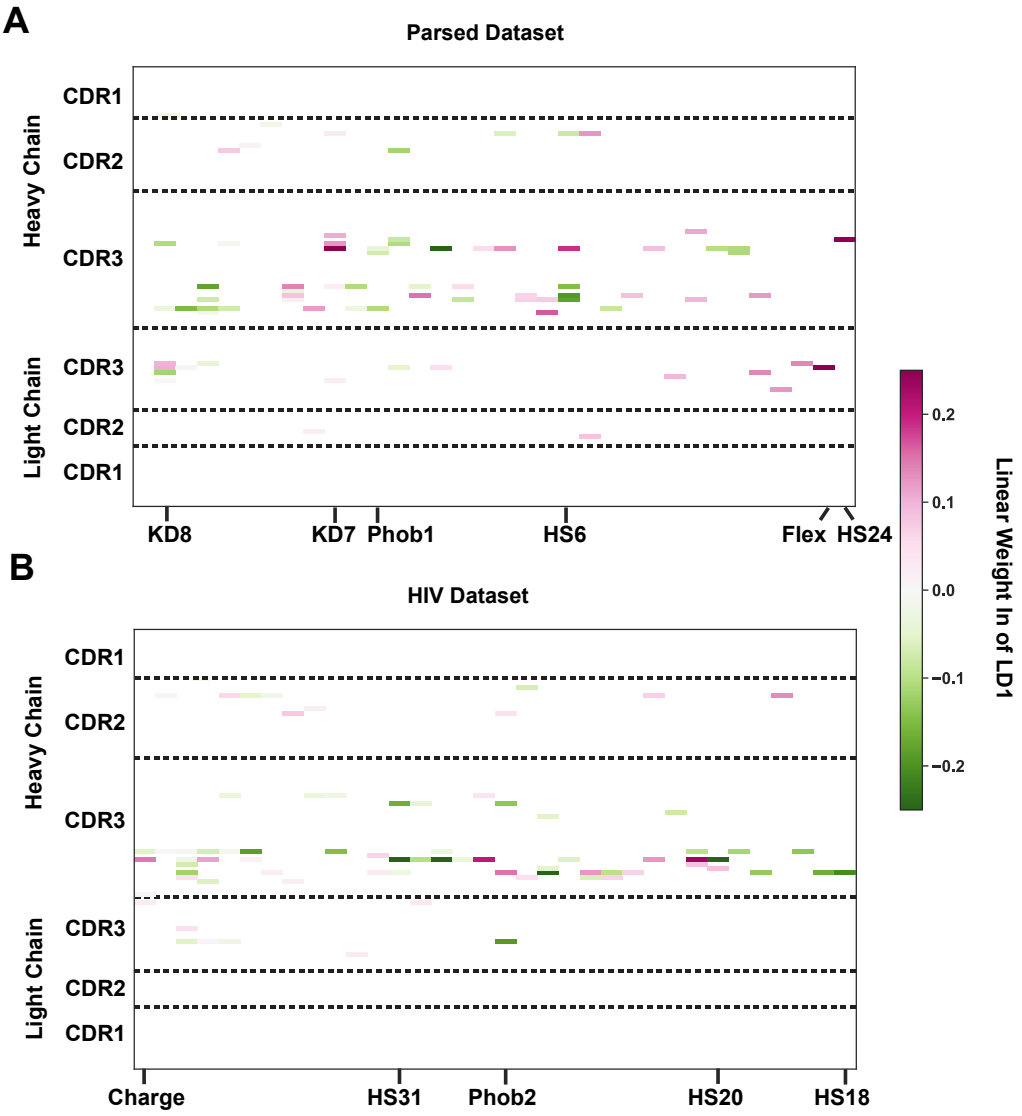


Figure 2.7: **The complete representation of the 75 linear weights that most effectively separate polyreactive and non-polyreactive sequences in the parsed complete dataset (A) and the parsed HIV dataset (B).**The x-axes each represent a single biophysical property selected after parsing down the full feature list using a maximal difference algorithm and a correlation analysis.

The predominant discriminating factors between datasets might be due to the significant difference in CDR3H length between the mouse (IgA) and the human datasets, which confounds the analysis in this region. However, when examining each individual subset of the complete dataset we do find that there are common properties that seem to be the

primary discriminators (i.e. largest linear weights). These are hydropathy 1, hydropathy 2, and hotspot variable 6 (a structural parameter related to α-helix propensity).

2.3.2 An Information Theoretic Approach

While analysis of the biophysical property differences between polyreactive and non-polyreactive sequences provides some insight into the molecular basis for the polyreactivity phenomenon, a broad unifying pattern which could discern the biophysical mechanism behind polyreactivity is not readily evident across all types of antibodies. To probe these polyreactive sequences in a quantitative yet more coarse manner, we applied the formalism of information theory to our dataset of antibody sequences. In brief, we use the concepts of Shannon Entropy, a proxy for diversity, and mutual information, a metric capable of quantifying crosstalk, for the analysis of our antibody sequences. A thorough explanation of our use of these concepts of information theory can be found in Chapter 4.

Figure 2.8A shows the Shannon entropy distribution for the full dataset of polyreactive and non-polyreactive antibodies. Given there are only 20 amino acids used in naturally derived antibodies, we can calculate a theoretical maximum entropy of 4.2 Bits, which assumes that every amino acid occurs at a given position with equal probability. Although the observed entropy of the CDR3H loop approaches this theoretical maximum, it hovers below it (3.5 Bits) due to the relative absence of the amino acids cysteine and proline in the center of this loop. The difference in the entropy distributions in CDR1H are consistent with the bias in amino acid usage in this region, shown previously in Figure 2.3.

When calculating the mutual information, we are interested in the location of the CDR loops in relation to one another, to help explain why we see patterns of increased crosstalk. To orient ourselves in physical space, Figure 2.8B gives an example crystal structure (PDB: 5UGY) [90] highlighting the lateral arrangements of the CDR loops.

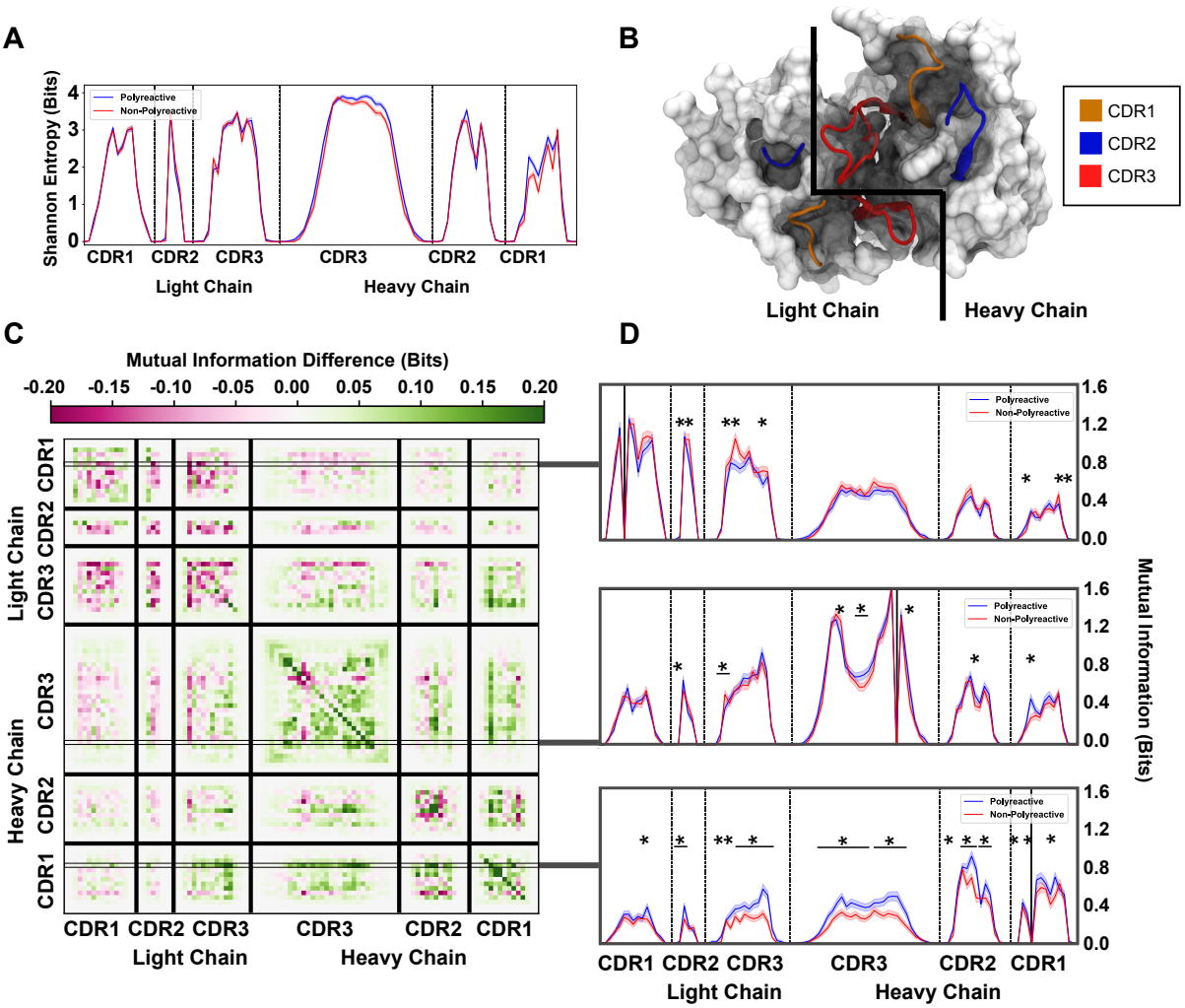


Figure 2.8: **An information theoretic analysis of antibody sequences shows an increase in polyreactive antibody loop crosstalk.** (A) The sequence diversity of the polyreactive and non-polyreactive datasets, quantified using Shannon Entropy, highlight similar diversities between the two groups. (B) A crystal structure (PDB: 5UGY) provides a visual representation of the lateral organization of the CDR loops on the antibody binding surface. (C) The difference in mutual information between polyreactive and non-polyreactive sequences shows that CDR loops of the heavy chain have more crosstalk in polyreactive antibodies. Each individual row represents the given condition, whereas each column gives the location the mutual information is calculated. (D) Singular slices of the mutual information show the data in (C), projected from the matrix onto a line, highlighting the significance of the differences at these particular locations. The positions of the "given" amino acid, i.e. the particular Y in $H(X|Y)$, are highlighted by grey boxes in panel C. Solid black lines indicate where on the X-axis this "given" amino acid is located. Stars indicate statistical significance ($p \leq 0.05$) calculated through a nonparametric permutation test. Bars with a single star above represent contiguous regions of significance.

The matrix in Figure 2.8C shows that the mutual information between CDR loops on this binding surface is increased in the heavy chains of polyreactive antibodies over non-polyreactive ones, suggesting an increase in loop crosstalk in antibodies that exhibit poly-

reactivity. Interestingly, it appears that there is a corresponding decrease of loop crosstalk in the light chains of polyreactive antibodies. This observed crosstalk persists across all polyreactive antibodies within all subsets of our tested dataset and is evident both in intra-loop and inter-loop interactions. Figure 2.8D highlights some examples of the interesting significant differences of this crosstalk at distinct given positions within CDR1L, CDR1H, and CDR3H. A complete plot of the statistically significant differences ($p \leq 0.05$) of Figure 2.8C shows that a large portion of these differences are in fact significant (Figure 2.9).

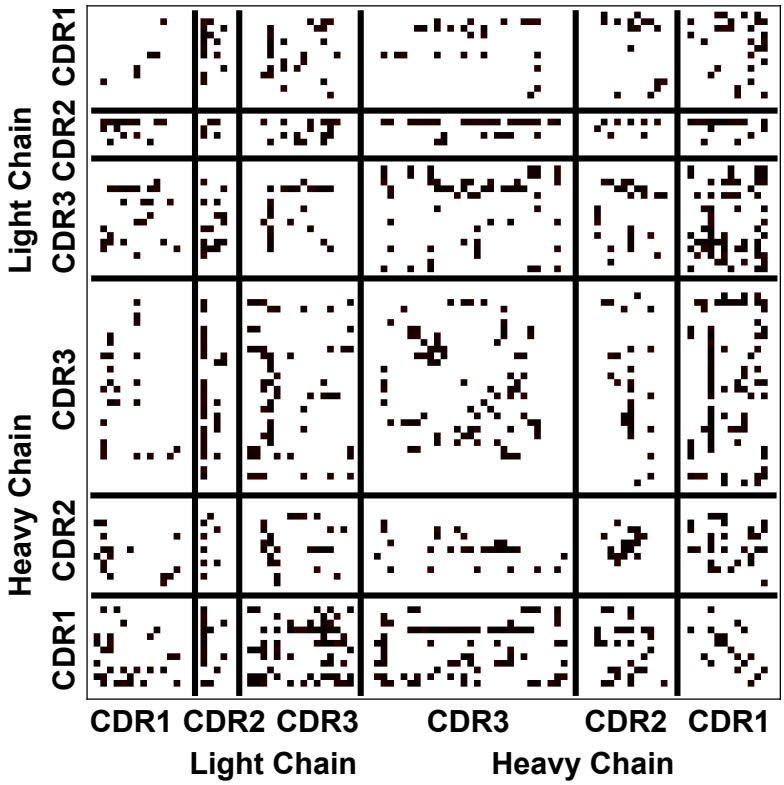


Figure 2.9: **The statistical significance of the values reported in Figure 2.8C.** Each black dot represents statistical significance (p 0.05) at that given location. Significance is calculated using a non-parametric permutation test.

The ordering of these entropy and information plots is chosen to reflect the spatial arrangement of the loops on the antibody surface; as such they show also that mutual information between loops drops off with physical distance between these loops. In other words, loops (and residues) that are located close to each other will have more of an effect on their direct neighbors as opposed to those that are more physically distant. This increased

mutual information suggests that in the heavy chains of polyreactive antibodies, there is enhanced cooperativity or co-evolution of the amino acids of intra- and inter-CDR loop pairs.

2.4 Molecular Dynamics Simulations Corroborate Informatics

While these calculations of mutual information provide strong evidence of an increase in intra- and inter-CDR loop crosstalk, what this loop crosstalk entails physically is not immediately clear from these measurements. The mutual information increase could come from gene usage being somehow coupled, amino acid composition correlating with the cognate ligand, or the amino acids directly interacting physically with each other. In some way, this crosstalk appears to be selected for in the polyreactive population. To address this concern, we set out to directly ascertain the source of this crosstalk using a structural and dynamic approach. Dr. Marta Borowska, a former graduate student in the University of Chicago's Department of Biochemistry and Molecular Biology, made a substantial contribution to this effort through the crystallization of five antibodies derived from the bioinformatic dataset analyzed above [91]. These antibodies are all tested for polyreactivity, with three identified as highly polyreactive (5-7 ligands recognized) and two that are completely non-polyreactive (0 tested ligands recognized).

Immediately, we see patterns emerge when comparing and contrasting the crystallized structures of polyreactive and non-polyreactive antibodies (Figure 2.10). The surface of the polyreactive antibodies 2G02, 43G10, and 338E6 appear to display a flatter binding interface, compared to the protrusions and crevices of non-polyreactive antibodies 4C05 and 3B03. Importantly, while we see a multitude of hydrophobic residues on the surface of polyreactive clone 2G02, we likewise see a large patch of hydrophobicity on non-polyreactive clone 3B03, countering the proposed role of these non-polar groups as key to polyreactivity [23, 72, 92]. Instead of highly hydrophobic binding surfaces bearing the brunt of responsibility, these structures suggest that charged residues are the key to polyreactivity. Figure 2.10 highlights

"naked" charges, i.e. unpaired charges on the binding surface, in yellow, whereas counterbalanced charge-charge interactions within the surface of a single antibody are highlighted in black. In polyreactive antibodies, we see at most one such naked residue within the binding interface, whereas both of the non-polyreactive antibodies have three or more.

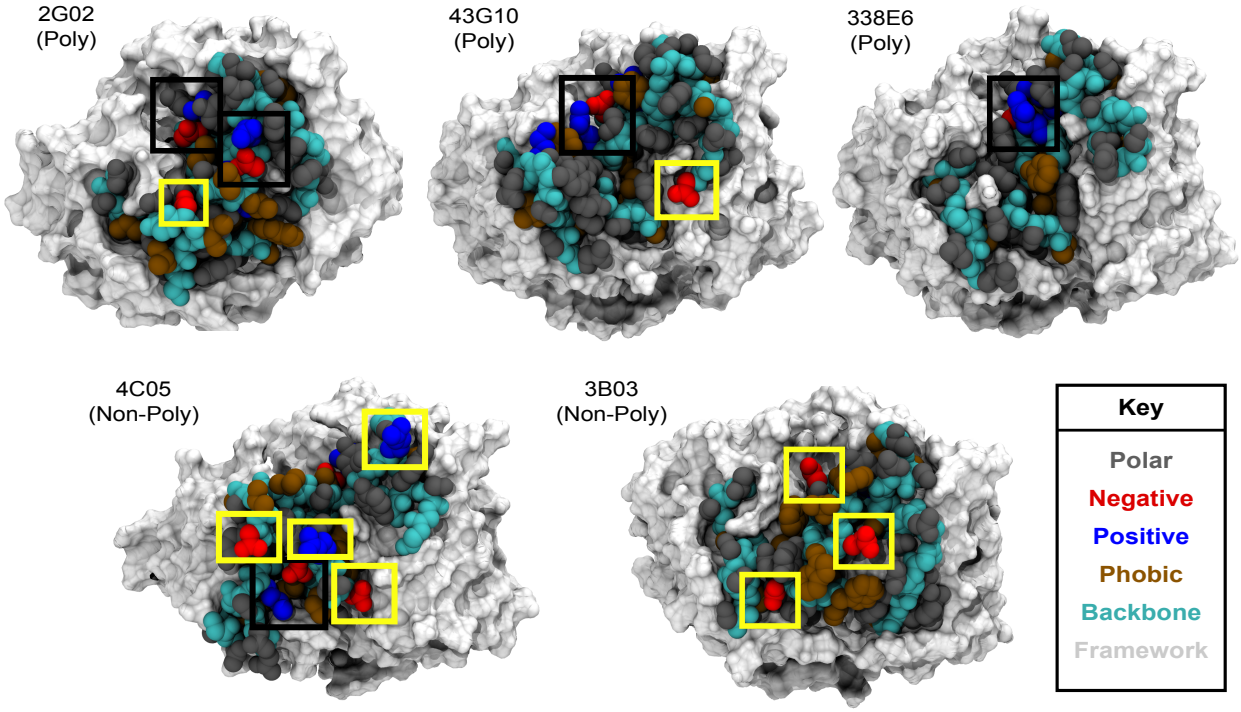


Figure 2.10: **Crystal structures of polyreactive and non-polyreactive antibodies display differing charge localization patterns.** Black boxes highlight instances of charge sequestration on the surface of antibodies via inter- and intra-CDR loop interactions. Yellow boxes highlight unpaired charges on the surfaces of antibodies.

Critically, these polyreactive antibodies do not have fewer charged residues, but instead these charged residues are more frequently sequestered. This sequestration of charge on the surface of polyreactive antibodies agrees well with the corresponding increase in mutual information, suggesting the crosstalk manifests as direct side chain interactions across the binding surface of polyreactive antibodies.

These direct side chain interactions across and within the CDR loops of polyreactive antibodies further calls into question the role of flexibility in antibody polyreactivity. Previous research has suggested that polyreactive antibodies are more flexible than non-polyreactive

antibodies [47,72], and more broadly there is a belief that specificity in antibodies is conferred via rigidification of the CDR loops [93–96]. One can imagine that a highly interconnected protein interface should remain relatively rigid, rather than adopt sufficiently flexible poses to conform to a variety of unrelated polyreactive ligands. We can test the relative rigidity of each of these structures using brute force molecular dynamics simulations, looking not for any specific motion of individual residues but instead to ascertain the overall flexibility of each region.

To carefully dissect this nuanced question of flexibility using molecular dynamics simulations, we include an additional antibody structure to be simulated to bring our totals to three polyreactive simulations and three non-polyreactive simulations. This sixth antibody structure, derived from the work of Guthmiller et al., includes two prominent Arg→Lys mutations alongside five additional germline reversion mutations of the polyreactive 2G02 antibody that abrogate said polyreactivity [47]. This newly non-polyreactive clone, referred to as the 2G02 mutant hereafter, provides an excellent comparison group to observe the dynamical nature of polyreactivity in antibodies. To probe these dynamics, all six structures are fully hydrated in a periodic water box, solvated in 0.15M NaCl, and equilibrated for 500 picoseconds and run for 1 microsecond of simulated time. Further simulation details and descriptions of the analysis can be found in the Appendix. These 1 μs simulations were run in duplicate, with the second replica run for 500ns, as there is little evidence of significant changes in dynamics over the latter half of the full microsecond scale simulations.

Beginning with comparisons between the polyreactive and non-polyreactive versions of antibody 2G02, we see distinct structural differences between the two simulated systems exacerbated by temporal evolution. Figure 2.11 provides representative frames from the 1 μs simulations of the polyreactive 2G02 antibody (Figure 2.11A) and the non-polyreactive 2G02 mutant (Figure 2.11B). While the mutation of arginine to lysine results in no net changes in charge on the antibody surface, arginine is a significantly more flexible side chain [97], capable of contortions which prove favorable for intra- and inter-loop interactions.

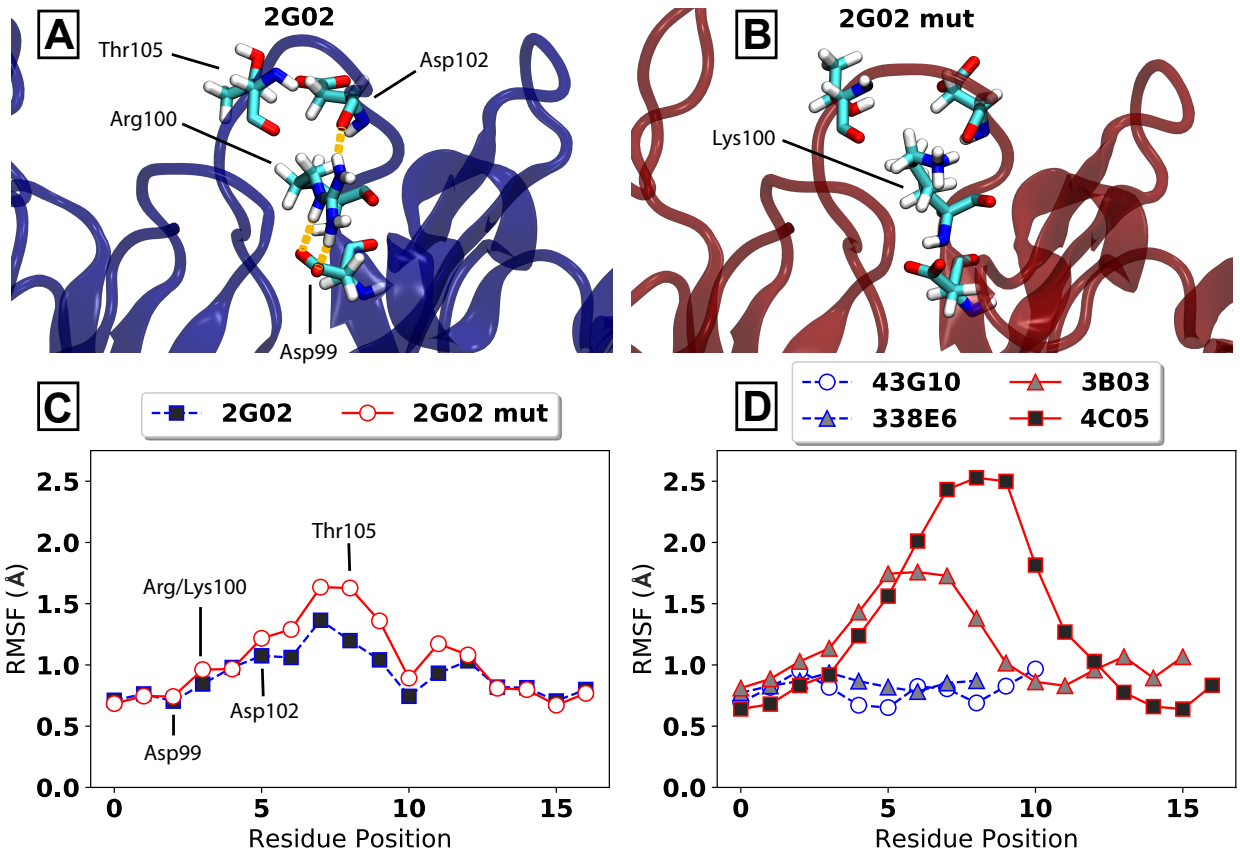


Figure 2.11: **The binding surfaces of polyreactive antibodies are more rigid than their non-polyreactive counterparts.** (A) A representative snapshot of molecular dynamics simulations of the polyreactive 2G02 antibody shows an intricate intra-loop hydrogen bonding network. Key residues are labeled, and hydrogen bonds are shown in yellow. (B) A similar snapshot of the mutated, non-polyreactive 2G02 antibody highlights the disruption of the hydrogen bonding network by the R100K mutation. (C) Quantification of the relative flexibilities of the molecular dynamics simulations highlighted in panels (A) and (B) via root mean square fluctuation (RMSF) of CDR3H. Key residues are labeled. (D) RMSF measurements of the CDR3H loops of remaining polyreactive (blue) and non-polyreactive (red) MD simulations show similar trends.

In 2G02, Arg100 forms a complex hydrogen bonding network, whereby a salt bridge between Asp99 and Arg100 stabilizes the loop sufficiently, creating a hydrogen bonding partner for the backbone oxygen of Asp102 and, less frequently, the backbone oxygen of Thr105. When this arginine is mutated to lysine, the hydrogen bonding network is completely disrupted, with the lysine making contact with neither the side chain of Asp99 nor the backbones at the distal portion of the loop. This disruption of the hydrogen bonding network has direct consequences on the flexibility of these loops, as quantified by root mean square fluctuation (RMSF). RMSF is a useful metric for quantifying the average fluctuation of single

residues over the course of entire simulated trajectories. RMSF is given by equation 2.1:

$$RMSF = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \|r_t - r_{ref}\|^2} \tag{2.1}$$

Here we calculate the distance between r_t, the position of the residue of interest at time t, and some reference position r_{ref}, then average this distance over time. We repeat this measurement for multiple residues across the CDR loops to quantify the flexibility over the course of each simulation. Figure 2.11C shows the impact of the disruption of the hydrogen bond network of 2G02 by the R100K mutation on the flexibility of CDR3H. Due to the proximity of D99 and R/K100 to the core of the protein, these mutations have little impact on this tightly packed region of the protein. However, between the more exposed residues of D102 and G106 we see a significant increase in the flexibility of the mutated 2G02 antibody compared to the wild-type. This mutation directly disrupts the crosstalk discussed in section 2.3.2, suggesting this increase in flexibility may be responsible for the loss of polyreactivity in mutant 2G02.

We can further compare RMSF across the remaining four antibody simulations. Figure 2.11D highlights exceptionally high flexibility in the CDR3H loops of non-polyreactive antibodies 4C05 and 3B03, compared to the robust rigidity of the loops of polyreactive antibodies 43G10 and 338E6. We see that if we compare these fluctuation quantifications of CDR3H across all simulated antibodies that 2G02 acts as something of a boundary case, with antibodies more flexible than wild-type 2G02 being non-polyreactive, and those less flexible being polyreactive (Figure 2.12A). Interestingly, we see a reversed pattern emerge when looking at the flexibility of CDR1L. RMSF measurements in CDR1L compared across all simulated systems appear to display an increased flexibility in polyreactive antibodies, save for polyreactive antibody 338E6 (Figure 2.12B). Again we see that the flexibility of CDR1L of the non-polyreactive, mutated form of 2G02 trends towards that of the other non-polyreactive antibodies.

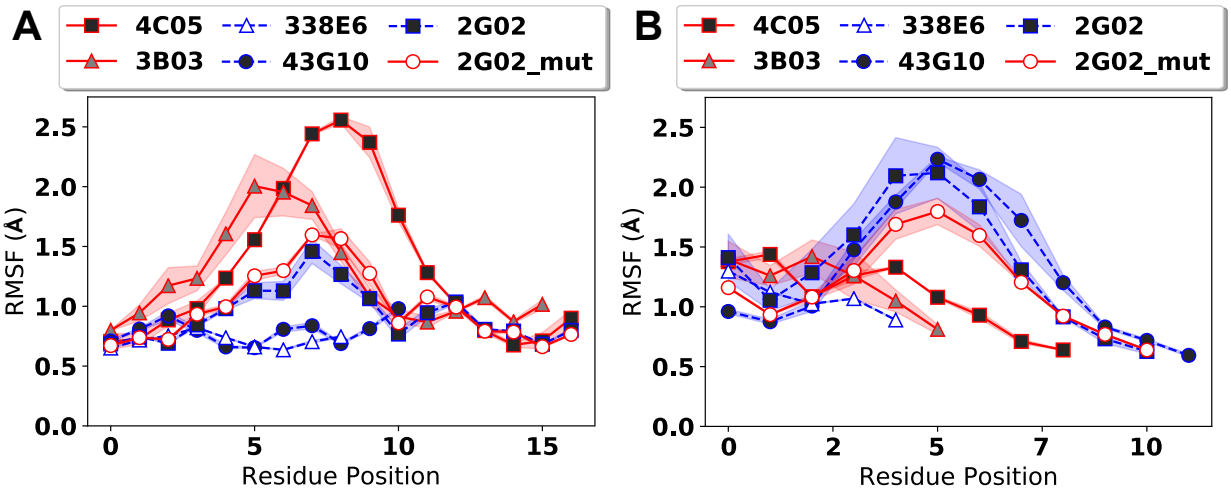


Figure 2.12: **Rigidification of polyreactive antibodies is not conserved across all CDR loops.** (A) The data of Figure 2.11C,D concatenated into a single plot highlights the relative flexibility of CDR3H of non-polyreactive antibodies. (B) Root mean square fluctuation (RMSF) measurements of CDR1L displays a reversed trend, whereby polyreactive antibodies appear to have more flexible loops. Data are averaged over two replicates, with standard deviations given by shadows surrounding each line.

While root mean square fluctuation is a powerful metric for quantifying flexibility, we can obtain a more complete story through visual inspection of trajectories. However, due to the significant length of the trajectories, we require a systematic way to identify distinct states across the course of the simulation. A common approach for identifying distinct, metastable states in molecular dynamics simulations is through time-lagged independent component analysis (tICA), a process originally developed in the signal processing literature [98]. tICA has been shown to identify slow degrees of freedom for protein motion [99] and is useful as a means of pre-processing data for further downstream analysis through significant dimensionality reduction [100]. Whereas PCA generates an orthogonal basis set across the dimensions of highest variance in the data, tICA incorporates temporal information to generate an orthogonal basis set through the slowest changing structural features.

Using PyEMMA, a Python library for the generation of Markov models [101], we isolate the cα backbones of each CDR loop and quantify the backbone dihedrals across the entire trajectory for each simulated system. We select a time lag of 1ns, narrowing our analysis to motions slower than this cutoff, and generate the first four independent components (ICs)

from the tICA processing. We then project the isolated backbone dihedral data on the first two independent components, i.e. the two tICA eigenvectors with the largest associated eigenvalues, for the polyreactive (Figure 2.13) and non-polyreactive (Figure 2.14) data. Within these projections, each isolated density in tICA space represents similar protein conformations. However, the relative similarity of these conformations within or across each region varies from projection to projection. As such, we utilize k-centers clustering to mathematically identify distinct groups in the tICA projections. We color each cluster and add these to the tICA projections, and then render a representative structure from each cluster, with the color of the rendered image matching the color of the k-center cluster marker.

It is important to note that these clusters are calculated using all four calculated independent components, helping to explain why some of the clusters on the two-dimensional projections appear to be close together and not necessarily spanning all visible densities in each plot. Looking first at the polyreactive antibodies, we find that despite what appears to be substantial breadth in the tICA space, the identified structures from each cluster show limited conformational change. The clusters of antibody 338E6 (Figure 2.13A) span across all four highly sampled regions, yet the renders of these structures identify each distinct conformation as simple translations through space, with each loop maintaining identical conformations. In the tICA projections of data from antibodies 43G10 (Figure 2.13B) and 2G02 (Figure 2.13C), the clusters instead appear to identify distinct conformational states of CDR1L, showing good agreement with the previously discussed RMSF data. Critically, we see that for these polyreactive antibodies, the CDR3 loops align nearly perfectly, independent of which cluster they are associated with.

Conversely, the non-polyreactive antibodies show significant differences across each cluster identified within the tICA projections. In antibody 4C05 (Figure 2.14A) these differences are primarily localized in CDR3H and CDR3L, with little motion across the other loops. Antibody 2B03 (Figure 2.14B) has a highly flexible heavy chain, particularly CDR3H, while the light chain adopts identical conformations across all identified clusters.

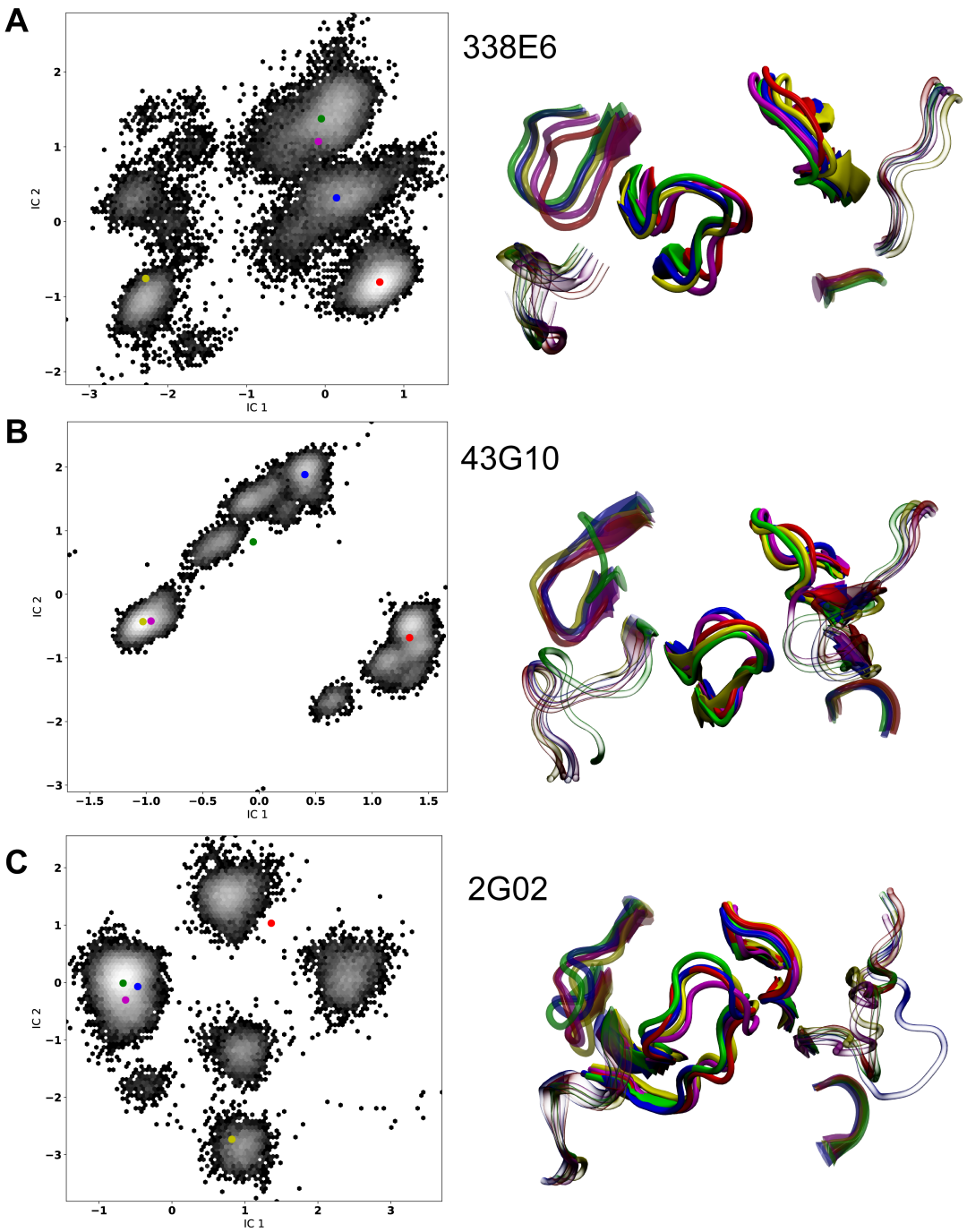


Figure 2.13: **tICA plots and representative structures of polyreactive antibodies highlight exceptional rigidity of their CDR loops.** Left: representative tICA plots from 1 microsecond simulations of polyreactive antibodies. Individual clusters within tICA space identified via k-centers clustering are represented by colorful circles. Right: representative structures of the CDR loops from each identified cluster from tICA plots. Colors of the structures match those found in the tICA plots.

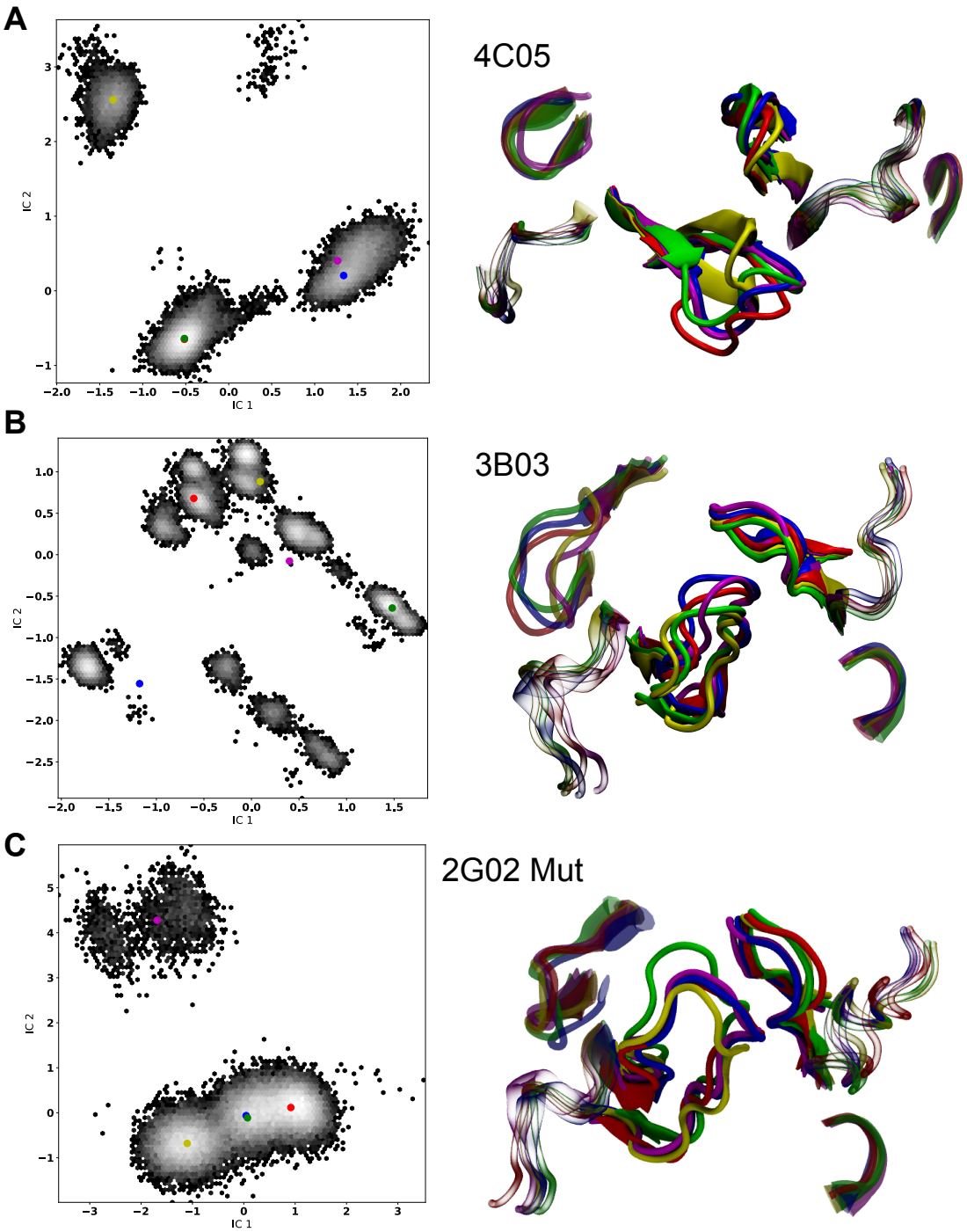


Figure 2.14: **tICA plots and representative structures of non-polyreactive antibodies highlight exceptional flexibility of their CDR loops.** Left: representative tICA plots from 1 microsecond simulations of polyreactive antibodies. Individual clusters within tICA space identified via k-centers clustering are represented by colorful circles. Right: representative structures of the CDR loops from each identified cluster from tICA plots. Colors of the structures match those found in the tICA plots.

Despite the presence of only one large basin in the tICA projection, the rendered clusters of the mutant form of antibody 2G02 show changes in the conformation of CDR3H, differing strongly from the observed tICA projections of the wild-type form.

Each tICA plot and subsequent render paints a vivid picture of the metastable states adopted by each antibody throughout the course of the simulation. However, the use of tICA to analyze MD data has important caveats. While the representative structures rendered above give a strong picture of the conformational changes that occur throughout each trajectory, the paths of the backbone motions between these states are lost. Likewise, due to the time averaging in the RMSF calculations, we lose all dynamic information. We can instead turn to root mean square deviation (RMSD):

$$RMSD(t) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \|r_n(t) - r_{n_0}\|^2} \tag{2.2}$$

RMSD is nearly identical to RMSF, but instead of averaging the displacement of each residue over time, we average the displacement over all residues at each time point. In this way, we can look at the gross structural changes of each loop through time. Figure 2.15 shows these RMSD traces for each polyreactive antibody, while Figure 2.16 shows these data for the non-polyreactive antibodies. Antibodies 338E6, 43G10, and 2G02 all display rigid CDR3H loops, while 4C05, 3B03, and the 2G02 mutant all show increased flexibility in this same loop. Interestingly, we see that CDR1H of polyreactive antibodies 338E6 and 43G10 are significantly more dynamic than those in non-polyreactive antibodies. However, this trend is not seen when comparing either wild-type or mutant 2G02, as differences between these two are only seen in CDR3H and CDR1L, as was seen in RMSF calculations. Overall, we find good agreement with the RMSF and tICA data outlined above, maintaining consistent trends across all measured metrics.

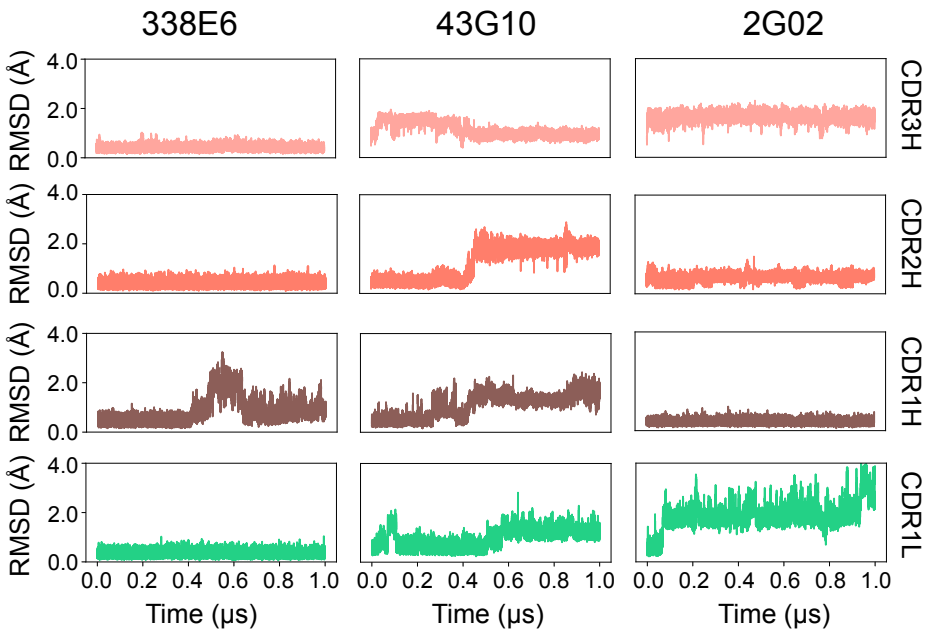


Figure 2.15: **Root mean square deviation of the CDR loops of polyreactive antibodies throughout 1 μs simulations show limited dynamics in CDR3H and CDR2H.** RMSD traces over all simulated time across four of the six CDR loops. Key differences in dynamics across poly- and non-polyreactive antibodies can be found within the heavy chain and CDR1L.

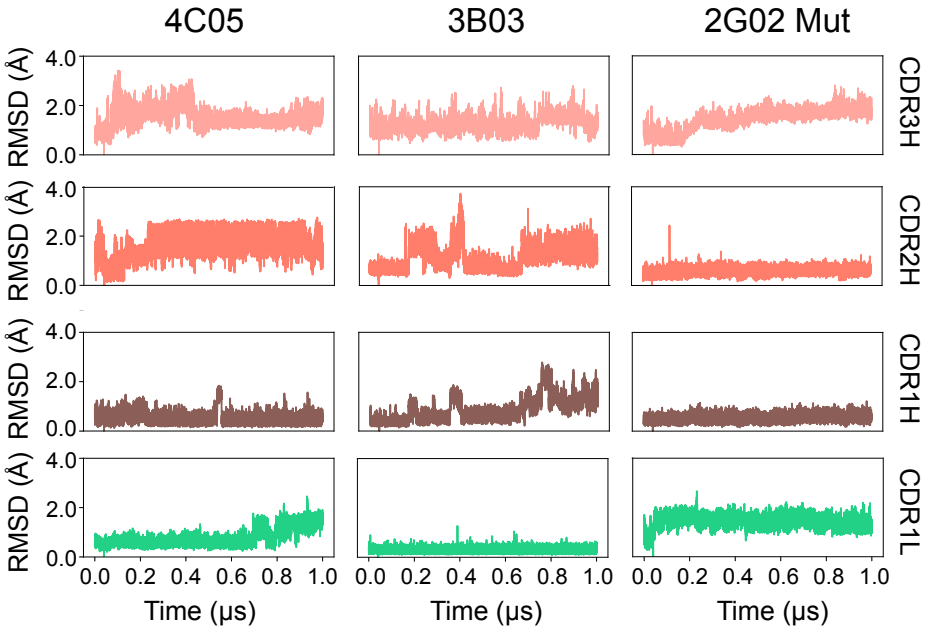


Figure 2.16: **Root mean square deviation of the CDR loops of non-polyreactive antibodies throughout 1 μs simulations show high flexibility in CDR3H and CDR2H.** RMSD traces over all simulated time across four of the six CDR loops. Key differences in dynamics across poly- and non-polyreactive antibodies can be found within the heavy chain and CDR1L.

2.5 Discussion

Previous research has highlighted the importance of hydrophobicity, charge, and CDR loop flexibility on antibody specificity. In this work, we expand upon these previous results with a new bioinformatic and biophysical characterization of polyreactive antibodies. The software generated for this study provides a powerful computational tool which can be utilized by researchers interested in discerning differences between populations of adaptive immune molecules in broad contexts. Building off of the efforts of our own work and that of experimental collaborators, we were able to aggregate to date one of the largest publicly available datasets of antibodies tested for polyreactivity. Differences in the germline gene frequency and amino acid frequencies show there exists some underlying differences between polyreactive and non-polyreactive antibodies. A surface level analysis of this dataset is able to discriminate certain features of polyreactive and non-polyreactive antibodies, namely that on average, polyreactive antibodies are less strongly negatively charged, less hydrophilic, and have a higher prevalence of antibodies with longer CDR loops of the heavy chain. Importantly, however, these binding surfaces do not have a net positive charge nor are they net hydrophobic.

Our results highlight an increase in V_H1-69 gene usage in polyreactive antibodies, an interesting finding given the substantial literature outlining its importance in diverse immune environments. In addition to the aforementioned role of V_H1-69 in broadly neutralizing anti-influenza and anti-HIV antibodies [69–72], autoreactive chronic lymphocytic leukemic B cells commonly express receptors bearing V_H1-69 [102,103], and anti-HIV antibodies which target the membrane-proximal external region of HIV-1 envelope glycoproteins frequently utilize V_H1-69 [104]. While previous reports suggest that the key feature permitting these auto-reactive or polyreactive interactions of V_H1-69 is an exceptionally hydrophobic CDR2H loop [92] our results suggest this does not explain the over-representation of this antibody in the polyreactive dataset, as on average the CDR2H of polyreactive antibodies is strongly hydrophilic. Instead, certain structural or dynamic features of the antibody may contribute

to its out-sized role in critical biological contexts.

To dig deeper into the biophysical differences between polyreactive and non-polyreactive antibodies, we created an adaptable software for the automated analysis of large antibody datasets and the application of a new analysis pipeline for the study of polyreactive antibodies. Overall, the improvements of this software to the current state of antibody sequence analysis are sufficient to highlight key differences in the two populations with improved spatial resolution. The position sensitive sequence alignment is able to further parse through the genetic differences and show that in general, polyreactive antibodies have a tendency to have more hydrophobic residues in CDR2H, and a decreased preference for phenylalanine in CDR1H. While these observational differences provided some initial insight, a more rigorous biophysical treatment was necessary. With the addition of 62 biophysical properties analyzed using the position sensitive alignment, significant differences between the CDR3H loops in polyreactive and non-polyreactive antibodies become immediately evident, providing a more detailed depiction of the antigen binding surface of polyreactive antibodies.

These data suggest a movement towards neutrality or "inoffensive" residues in the CDR loops of polyreactive antibodies: amino acids that are neither exceptionally hydrophobic nor hydrophilic and with a net charge close to 0. Previous studies have suggested that polyreactive antibodies tend to have more hydrophobic CDR loops, such that low affinity Van der Waals interactions might be the primary means of polyreactive interactions [23, 72]. However, these studies counted the number of hydrophobic residues per sequence or averaged the hydrophobicity of all six CDR loops. While our results partially agree with these previous findings, our analysis extends much further into defining the biophysical basis of this phenomenon. For example, while our position sensitive representation of the sequences shows that CDR3H does become more hydrophobic in polyreactive sequences, it is still net hydrophilic on average. A highly hydrophobic binding surface would provide an avenue for non-specific interactions with other hydrophobic proteins, but it would occlude binding to highly hydrophilic ligands like DNA. A slightly hydrophilic, neutral-charged binding surface

would permit weak interactions with a wide range of ligands.

Using these and other biophysical properties as input feature vectors, we were able to generate a generalizable protocol for binary comparisons between two distinct populations of Ig-domain sequences. This framework is able to successfully split all tested polyreactive and non-polyreactive antibody datasets. Care was taken to not overfit these data and a preliminary classifier built from this algorithm was able to identify the proper number of input vectors for each LDA application. While there are general features which best split the polyreactive and non-polyreactive antibodies in these datasets, including charge, hydropathy, and α-helix propensity, these features alone are not sufficient to discriminate between the two populations. Instead, 75 vectors taken from the position-sensitive biophysical property matrix are necessary to properly split the groups, including both simple properties like charge, hydropathy, flexibility, and bulkiness and more carefully curated properties like the often-used Kidera factors and the hotspot detecting variables of Liu et al [85, 86, 105]. The inability to arrive at a core few biophysical properties that could effectively distinguish polyreactive and non-polyreactive antibodies necessitated the application of further approaches, namely information theory.

The tools provided by information theory proved to be effective in the present study. The classic approach to information theory considers some input, communication of this input across a noisy channel, and then reception of a meaningful message from the resultant output. We can think of the analogous case for these antibodies, whereby the sequence and structure of the antibodies can be seen as our input, the thermal noise inherent to biological systems can complicate biochemical interactions, and the necessary output is antigen recognition, i.e. binding between the antibody and the ligand. Focusing just on the antibody side of this communication channel, we determined the underlying loop diversity through the Shannon entropy of the polyreactive and non-polyreactive datasets. This diversity was found to be nearly equivalent while the mutual information, a metric of "crosstalk" across populations, between and within CDR loops was found to be increased in the heavy chain and decreased

in the light chain of polyreactive antibodies. Importantly, this crosstalk is increased across and within all loops when analyzing the parsed dataset (Figure 2.17).

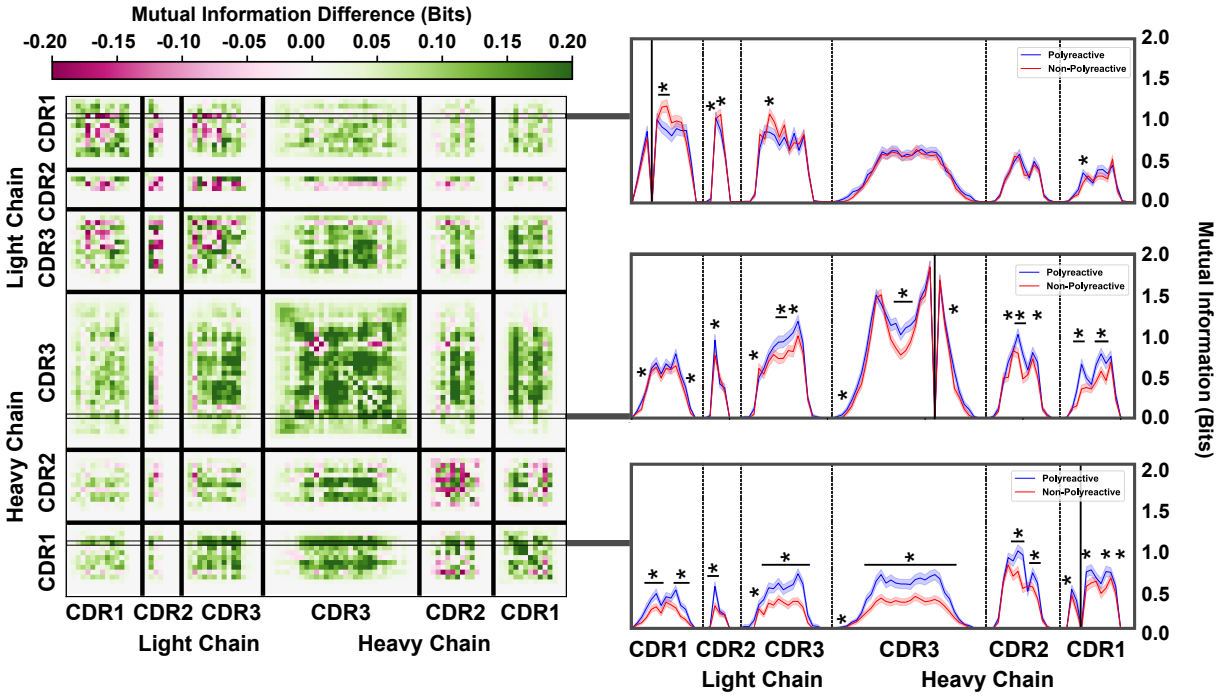


Figure 2.17: **An information theoretic analysis of the parsed antibody sequences shows an increase in polyreactive antibody loop crosstalk that is more pronounced when compared to the full dataset.** The difference in mutual information between polyreactive and non-polyreactive sequences shows that CDR loops have more crosstalk in polyreactive antibodies. Each individual row represents the given condition, whereas each column gives the location the mutual information is calculated (left). Singular slices of the mutual information show the data projected from the matrix onto a line, highlighting the significance of the differences at these particular locations. The positions of the 'given' amino acid, that is the particular Y in H(X|Y), are highlighted by gray boxes (right). Solid black lines indicate where on the X-axis this 'given' amino acid is located. Stars indicate statistical significance (p 0.05) calculated through a nonparametric permutation test. Bars with a single star above represent contiguous regions of significance.

As seen in the crystal structures solved by Dr. Borowska [91] mutual information manifests itself in these polyreactive antibodies as an increase in charge-charge interactions on the binding surface. This helps to explain the minimal change in net charge of antibodies between the two groups, yet the significant move towards neutrality in the CDR loops of polyreactive antibodies. The pairing of two charged groups helps move the binding surface of polyreactive antibodies towards a more "inoffensive" binding surface. A binding surface that is neither exceptionally hydrophobic nor hydrophilic, and lacks a significant positive or

negative charge, represents a relatively appealing binding interface for a low-affinity interaction with a large array of diverse ligands. A patchwork of hydrophobic and hydrophilic non-charged residues exposed to potential ligands generally represents an ideal candidate polyreactive surface. The corresponding decrease in the mutual information between the light chain CDR loops of polyreactive antibodies could be caused by a de-emphasis in the involvement of these loops due to differential binding configurations of polyreactive ligands, as has been previously hypothesized [37, 106].

Finally, the dynamic nature of these structural interactions were probed using all atom molecular dynamics simulations. Six antibody structures were simulated for a total of 12 μs of accumulated simulated time. Using a variety of analytical techniques, we found that specifically the CDR3 loop of the heavy chain of polyreactive antibodies tends to be more rigid than that of non-polyreactive antibodies. Conversely, the CDR1 loop of the light chain of these same polyreactive antibodies appears to be more flexible in polyreactive antibodies. These data directly contradict recent results that propose flexibility as a key feature of polyreactive antibodies [47]. While Guthmiller et al. utilized computationally predicted models of polyreactive and non-polyreactive antibodies, our simulations are based on the first ever crystallized structures of antibodies confirmed to be polyreactive. Although structure prediction has come a long way in the past decade, they still perform relatively poorly in the placement of side chains [80]. As seen in our crystallographic analysis, the relative positioning of side chains has critical implications on the resulting dynamics. Additionally, the sole metric for flexibility in Guthmiller et al. is based upon the spread of data across projections of the dynamic data on to principal components. Our results show that tICA projections, a similar metric for collapsing the dimensionality of dynamic data, do not necessarily reflect flexibility of the antibodies. Overall, we find that polyreactive antibodies are in fact more rigid, but further dynamic studies of these polyreactive antibodies will be needed to confirm these results.

The results obtained using linear discriminant analysis further contextualizes these dy-

namic results, helping to complete the circle of observations between bioinformatics and dynamics. In addition to standard side chain properties, many of the most important features for splitting polyreactive and non-polyreactive antibodies were structural in nature. Specifically, hotspot variables 6, 24, 25, and 41 all correspond to the structural tendencies of a given amino acid. Coupled with the increase in side chain interactions that may be implied by the increased mutual information across the loops of polyreactive antibodies, this potential for increased loop structure in polyreactive antibodies suggests more rigid CDR loops in polyreactive antibodies, as was seen in the RMSF and RMSD measurements taken from molecular dynamics simulations.

Further experimental assays will be necessary to more comprehensively identify the underlying mechanisms of polyreactivity, including additional sequencing and biochemical analyses of polyreactive and non-polyreactive antibodies. Antibodies specific to other pathogens or those from other organisms tested for polyreactivity will help form a more complete picture and improve the generality of the results. As with any machine learning based approach, the classification algorithm is only as good as the data it is trained on. Adding further data in the training set, including more mutations and germline reversions that turn a polyreactive antibody non-polyreactive or vice-versa, will be critical for a comprehensive analysis of polyreactivity. Additionally, a more robust assay for determining polyreactivity such as a chip-based screen to test for binding to many diverse targets would greatly expand our perspective and help understand just how broadly reactive these polyreactive antibodies are. Lastly, a more complete understanding of the germinal center and the selection processes inherent to the affinity maturation process will assist in the determination of whether polyreactivity is a byproduct or a purposeful feature of the affinity maturation process.

CHAPTER 3

UNDERSTANDING THE ROLE OF BTN3A1 IN Vγ9Vδ2 T CELL ACTIVATION

3.1 Introduction: A Persistent Challenge to Our Understanding of T Cell Activation

Shortly after the initial identification of $\gamma\delta$ T cells by Brenner et al. in 1986 [107], this same group at Brigham and Women's Hospital in Boston, Massachusetts determined a role for these cells in humans. Modlin et al. discovered that $\gamma\delta$ T cells are prevalent in lesions derived from leprosy patients, and that these lesion-resident T cells proliferate in vitro in response to mycobacterial antigens [108]. In the same year, a separate group at the National Institutes of Health determined that this same line of T cells is activated in response to antigens derived from *Mycobacterium tuberculosis*, in addition to the previously discovered *Mycobacterium leprae* derived ligands [109]. This line of T cells was later further characterized and found to bear Vγ9+ and Vδ2+ T cell receptors, occasionally referred to as Vγ2Vδ2 TCRs due to conflicts in nomenclature preferences but only referred to as Vγ9Vδ2 TCRs in this work. Despite this explosion of scientific discovery, critical questions surrounding the identity of the ligand itself remained. These mycobacterial antigens were known to require some type of presentation by host cells [108], but any further details were lacking. Were these antigens peptide-based like the majority of previously identified T cell antigens? Were they similarly presented by MHC class I or class II molecules to the TCR?

In 1994, Jean-Jacques Fornié and Marc Bonneville's groups quickly answered this first question, discovering that the ligand in question was not a peptide, breaking sharply from the known paradigm of T cell activation, and instead was a phosphate-containing molecular metabolite [110]. Shortly thereafter, Michael Brenner and Barry Bloom's labs completed the work started in France, nailing down the identity of the first known activating ligand,

isopentenyl pyrophosphate (IPP) [111]. As the field progressed, further ligands were identified, most notably (E)-4-Hydroxy-3-methyl-but-2-enyl pyrophosphate (HDMAPP), the most potent microbially-derived pathogen. These ligands were later classified as a broad cluster of antigens collectively called phosphoantigens (pAgs). These pAgs were known to activate Vγ9Vδ2 T cells, but the key mediating molecule remained a mystery. Further research ruled out the classical antigen presenting MHC class I [11, 112] and class II [12, 110] molecules as the key mediators, creating more questions than were answered. Other activating molecules were proposed, but the true molecule responsible for transmitting the information of pAg accumulation was not identified for nearly twenty years after the first ligand-characterizing studies.

Finally, in 2011, Harly et al. identified butyrophilin 3A1 (BTN3A1) as a necessary protein for the pAg-dependent activation of Vγ9Vδ2 T cells [55]. These researchers discovered that when incubating cells with anti-BTN3A1 antibodies, they subsequently became able to activate Vγ9Vδ2 T cells. Further follow-up experiments showed that knocking down BTN3A1 using small-hairpin RNA completely abrogated the ability of target cells to activate T cells upon the addition of pAg [55]. Proceeding this study, the Adams lab at the University of Chicago determined the structural basis for this ability to sense pAg, identifying the intracellular B30.2 domain of BTN3A1 as a pAg-binding domain [56]. This had profound implications for the mechanism of TCR recognition of antigen, suggesting that intracellular pAg accumulation was somehow mediating an extracellular change, rather than the direct antigenic presentation that has been consistently demonstrated in $\alpha\beta$ T cells. While this insight was a massive breakthrough in our understanding of the Vγ9Vδ2 T cell activation mechanism, it remains to be seen what the downstream consequences of this pAg binding are.

3.2 Extracellular Conformational Change as a Driver of T Cell Activation

There have been a host of theories for the "true" mechanism linking pAg and BTN3A1 to T cell activation. Early on, before the work of Sandstrom et al., there was some suggestion that pAg was actually presented extracellularly with the Vγ9Vδ2 TCR making direct contact with a BTN3A1-pAg complex [113]. This hypothesis has been disproven, repeatedly and resolutely, by numerous groups across the research community, solidifying the role of intracellular pAg binding as the mechanism of BTN3A1-mediated T cell activation [114,115]. Instead, the field has converged upon a so-called "inside-out" model whereby intracellular binding of phosphoantigen leads to an extracellular conformational change of BTN3A1 that is subsequently recognized by the TCR. This model was largely based upon further structural data from the Adams lab in work by Palakodeti et al., who discovered two distinct conformational states of BTN3A1 homodimers [57]. The first state, a proposed "inactive" conformation that adopts a head-to-tail arrangement of the extracellular domains is suggested to then convert into a v-shaped, activating conformation (Figure 3.1). In this model, pAg acts on the intracellular B30.2 domain, and in some way transmits the information necessary to induce a conformational change extracellularly, supposedly through the single-pass transmembrane helix. My work in this space has primarily focused on testing this model, attempting to identify the predominant conformation and the likelihood of this conformational change.

Previous research focused on attempting to validate this model of conformational change have mostly fallen short of their target. In the very paper the model was suggested, Palakodeti et al. found that in solution, only the v-shaped conformation could be observed using a FRET-based approach [57]. Further negative-stain electron microscopy results failed to identify the hypothesized head-to-tail conformation using full-length protein expressed in lipid nanodiscs [116]. Despite the inability to detect this proposed inactive conformation, the shortcomings of each of these methods made explicitly ruling out the head-to-tail model

difficult. Further experimental validation is necessary to confirm or deny the existence of this conformation.

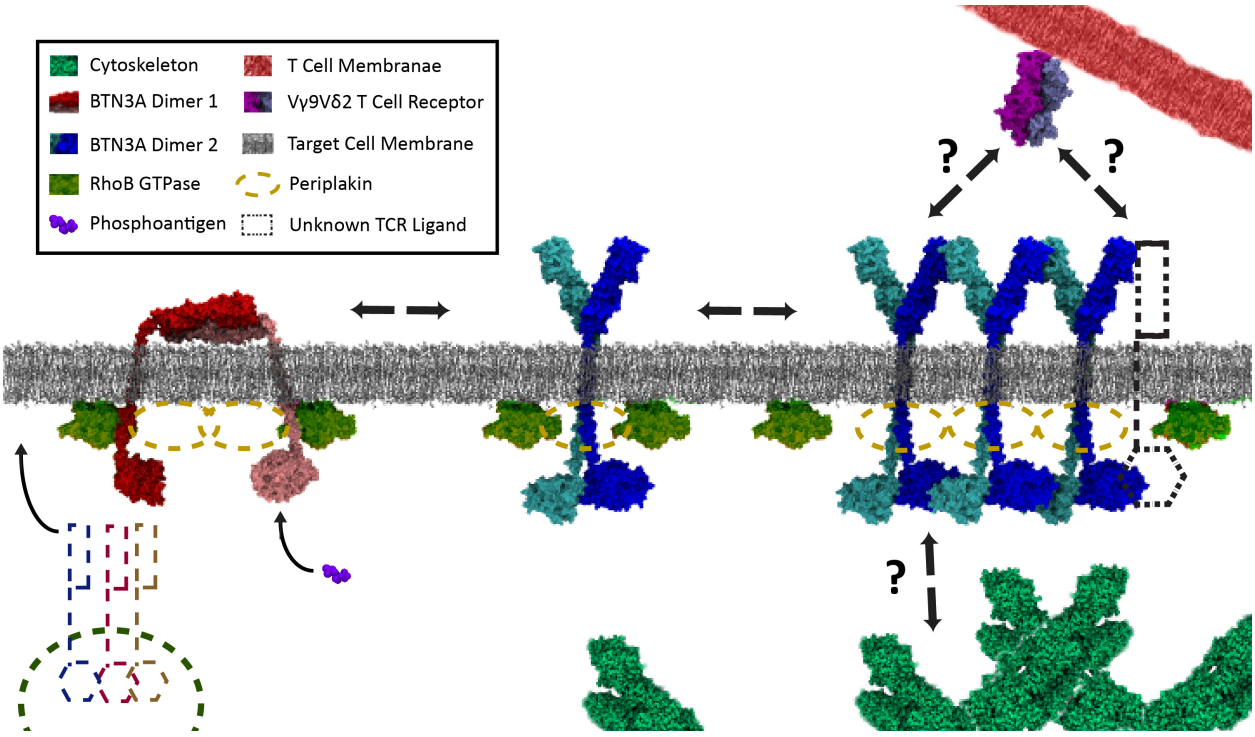


Figure 3.1: **A comprehensive outline of the conformational change model of BTN3A-mediated pAg recognition by Vγ9Vδ2 T cells.** In this model, pAg binds to inactive BTN3A1 homodimers (red and pink), leading to a conformational change producing active BTN3A1 homodimers (blue and cyan). These activated homodimers then cluster at the immune synapse, either through cytoskeletal interactions or some other, unknown mechanism. Accessory proteins rhoB and periplakin are included highlighting their proposed interaction with BTN3A1 intracellular domains, yet their function in this accessory role remains unclear.

3.2.1 Using Atomic Force Microscopy to Probe Conformational Change

Clearly, a method more sensitive than either FRET or negative-stain microscopy is necessary to fully rule out or confirm the existence of the head-to-tail conformation. In an attempt to answer this question, we utilize atomic force microscopy (AFM), an approach capable of identifying conformational changes on the nanometer scale. AFM is a conceptually simple technique, whereby a small composite cantilever is raster-scanned across a nanoscale topographical surface. AFM leverages the simple geometry of long lever arms to amplify

small perturbative deflections incident upon the cantilever probe into much larger motions of a laser whose motion is sensed by a photodetector [117]. Modern atomic force microscopes have sub-nanometer precision in the z-axis, and x-y precision on the order of a few nanometers [118]. This resolution should be sufficient for discriminating between the two proposed conformations of BTN3A1's extracellular domain, with structures given in Figure 3.2A,B. The data generated by AFM resemble topographical maps, from which 2D-slices can be taken to give detailed local height traces. A hypothetical height trace can be seen in Figure 3.2C, providing insight into what we can expect AFM data of these protein samples to look like.

Despite this exceptional three-dimensional resolution, AFM is often a difficult technique to apply in practice due to the requisite surface-restriction of the measurement. Any measured sample must be stably bound to the imaging surface, either through simple electrostatic effects or through more involved functionalization techniques [119, 120]. Two distinct routes are taken to overcome this technical issue. The first approach is centered around creating a closer approximation of the natural environment of BTN3A1 by reconstituting the purified protein into lipid vesicles. These lipid vesicles are then deposited onto the flat mica AFM imaging surface, whereby they spontaneously transition from a spheroid vesicle to a planar bilayer. The second approach is more direct, removing some of the complicated steps involved with reconstituting the protein into a lipid bilayer. Rather than relying on lipid vesicles depositing onto the mica surface, we solubilize the protein in detergent and directly probe the conformation of the protein on the imaging plane. The protocols for purification, reconstitution, and measurement of the protein using each of these approaches can be found in the Appendix.

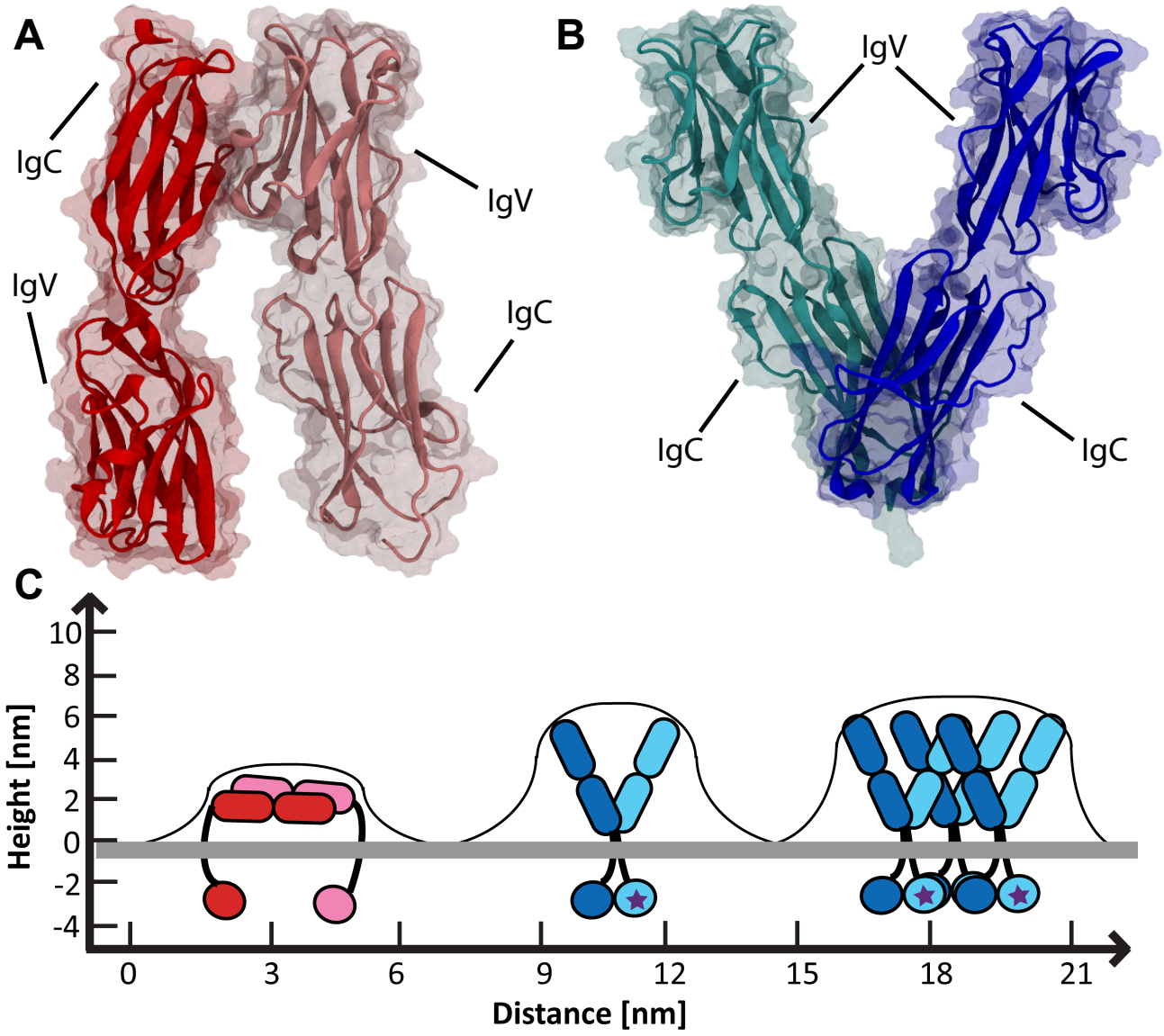


Figure 3.2: **The distinct conformational states of BTN3A1's extracellular domains readily lend themselves to interrogation by atomic force microscopy.** (A) Crystal structure of BTN3A1's extracellular head-to-tail conformation. (B) Crystal structure of BTN3A1's extracellular v-shaped conformation. (C) Hypothetical atomic force microscopy height trace of the conformations highlighted in (A) and (B) in the context of a supported lipid bilayer.

The strength of using lipid vesicles to identify the conformations of BTN3A1 lies in our ability to focus our measurements in the incredibly precise z-dimension. As suggested in Figure 3.2, we expect that the v-shaped conformation should protrude nearly 8nm off the surface of the lipid bilayer, while the head-to-tail conformation juts a mere 4nm above the lipid headgroups. This difference in height is well within the range of detection of a modern atomic force microscope. Consistent with previous experimental results, we are able to identify features consistent with these v-shaped conformations, with a representative image shown in Figure 3.3A. Here dark colors represent low features and bright objects represent features of higher prominence. Specifically, the mica surface can be seen in dark brown, the lipid bilayer in gold, and the v-shaped BTN3A1 in white. Figure 3.3B zooms in on this data, while panels C and D provide three-dimensional renders to give further perspective. The height trace of profile 1, the prominent feature in the data, provided in Figure 3.3E shows good agreement with the predicted height of the v-shaped conformation (8nm). Distinctly lacking from this data is any evidence of the head-to-tail conformation. The faint features around 1nm in height outlined in Figure 3.3F are likely nanoscale salt aggregates, too small to be the head-to-tail conformation.

Unable to observe the head-to-tail conformation using this vesicle-based approach, we further attempted to probe the conformation of BTN3A1 using a more direct measurement. Full-length BTN3A1 was expressed and solubilized in DDM detergent and deposited directly onto mica. Figure 3.4A shows the results of this direct measurement, with each distinct bright spot corresponding to each domain of BTN3A1. The line traces taken from this topographic map (Figure 3.4B) matches up almost perfectly with what we would expect to see for a homodimer of BTN3A1, with a larger, longer extracellular domain and a smaller intracellular B30.2 domain. However, given the poor x-y resolution of atomic force microscopes resulting from tip convolution [121,122], we cannot confidently determine which conformation is observed in this profile.

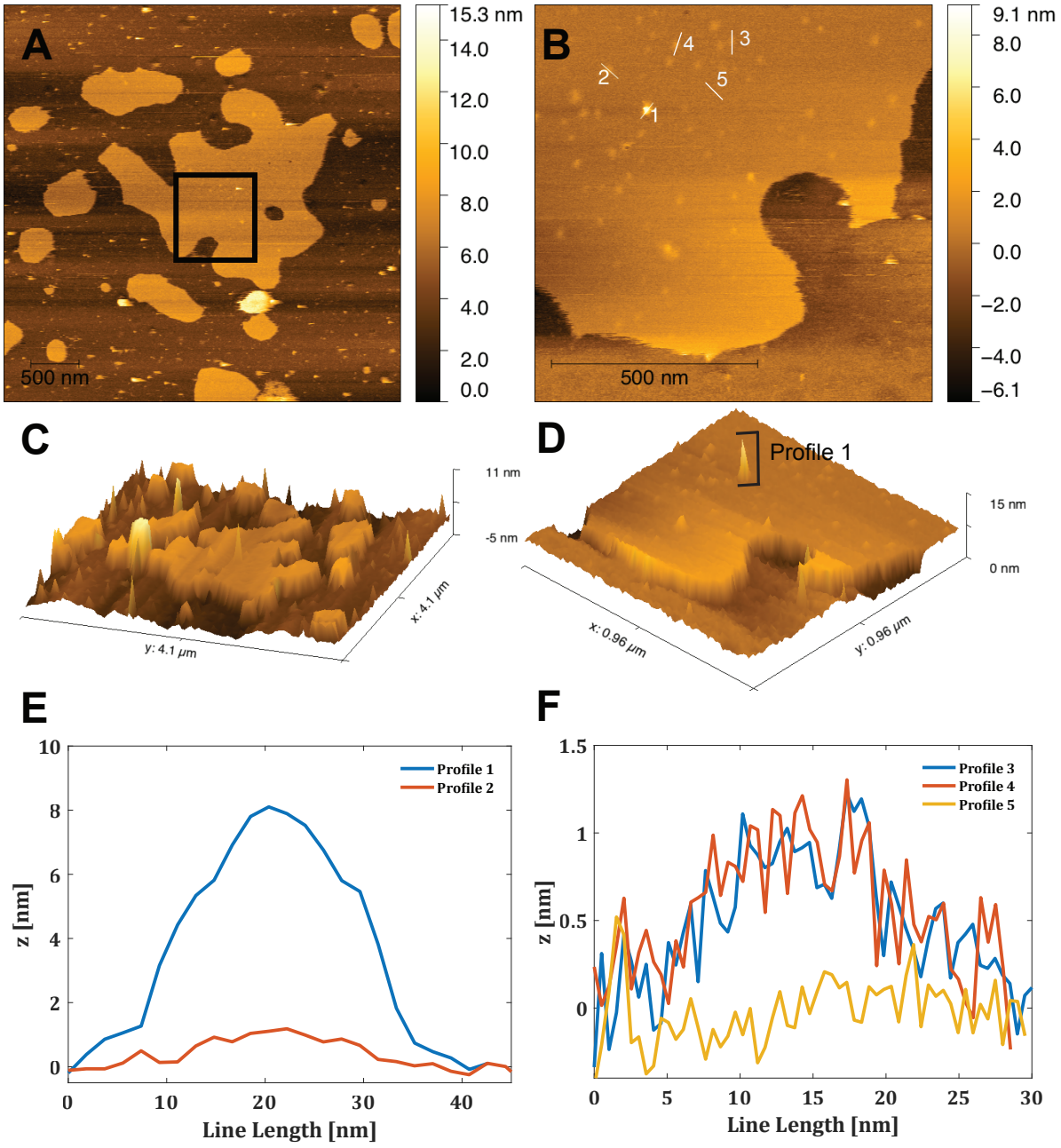


Figure 3.3: **Atomic force microscopy of BTN3A1 in supported lipid bilayers may be capable of detecting the v-shaped conformation, but not the head-to-tail conformation.** (A) 2D scan of BTN3A1 reconstituted into vesicles deposited on mica as supported lipid bilayers. Scale bar bottom left gives XY scale, color scale right gives Z scale. (B) Zoomed scan of the boxed region in (A). Scale bar bottom left gives XY scale, color scale right gives Z scale. (C) 3D representation of the data in (A). (D) 3D representation of the data in (B). (E) Line scan of profiles 1 (putative BTN3A1 v-shaped conformation) and 2 (lipid) from (B). (F) Line scan of profiles 3 (salt), 4 (salt), and 5 (lipid).

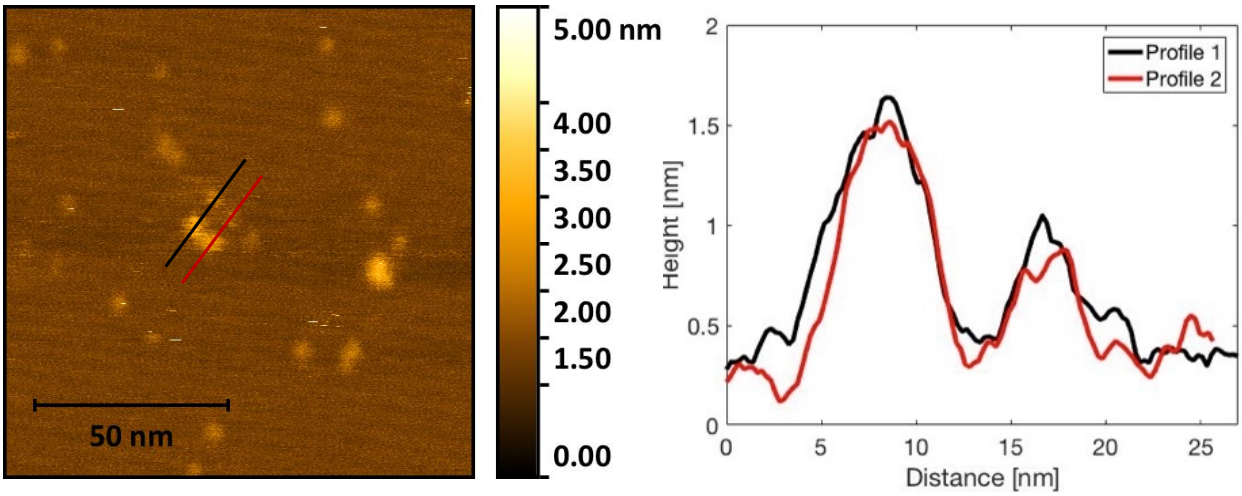


Figure 3.4: **Atomic force microscopy of BTN3A1 solubilized in detergent can resolve full-length BTN3A1, but not the conformation of the extracellular domains.** (A) 2D scan of BTN3A1 deposited directly on mica. Scale bar bottom left gives XY scale, color scale right gives Z scale. (B) Height profiles of the two traces in (A).

While these data further point towards the v-shaped conformation being the primary state, and perhaps the only state, adopted by full-length BTN3A1, they still lack the strength to conclusively rule out the head-to-tail conformation as one that is truly realized in solution. These results help to contextualize BTN3A1 and show that the conformation of the protein appears to not be significantly affected by the lipid bilayer, but further detail is needed to determine the relevance of the two molecular conformations.

3.2.2 Dynamic Tests of BTN3A1's Extracellular Domain Stability

While experimental approaches have thus far fallen short in determining the physiological relevance of the head-to-tail BTN3A1 conformation identified in crystal structures, a computational approach may provide atomistic insights capable of breaking down this problem. Specifically, computation can be used to ascertain the relative stability of each of these conformations. In this work, we first utilize all-atom molecular dynamics to survey the temporal evolution of each of the dimer interfaces of the two crystal structure conformations of BTN3A1. However, as mentioned in Chapter 1, all-atom molecular dynamics is significantly hampered in its ability to sample long timescales. We can overcome this shortcoming via

coarse-graining, a technique whereby waters and side chains are replaced by implicit functional groups to reduce computational cost. Using these two computational techniques, we are able probe the differences in conformational dynamics between the two proposed states.

Each protein structure was downloaded directly from the protein database and set up for simulation using the CHARMM-GUI [123, 124]. Further details of the simulation set up can be found in the Appendix. The two simulated systems are run in duplicate for 500 nanoseconds in periodic boxes fully hydrated with explicit waters and buffered with 0.15M NaCl. We run these simulations in duplicate to mitigate the stochastic initialization effects inherent to all-atom simulations [125, 126]. In each of these systems, we can look at three metrics as a proxy for the relative stability of the two protein conformations. Using the two crystal structures as reference states (Figure 3.2A,B), we can measure the RMSD, the RMSF, and the relative angle between the two domains, all as a function of time across both the interfacial domains and the entire protein structures.

Figure 3.5 shows that the v-shaped conformation is more stable than the head-to-tail conformation across duplicate simulations. We examine both the stabilities of each dimeric interface as well as the entire protein structures, denoted in the figures as "interface" and "full", respectively. Root mean square deviation (RMSD) traces (Figure 3.5A,B) highlight the exceptional stability of the v-shaped conformation, with deviations under 1.5 Å for the interface and under 2.0 Å across the entire structure. Conversely, the head-to-tail simulations rapidly diverge from their crystallographic interface, displaying an interface deviation of over 2.0 Å within the first 200ns of simulated time and a dynamic structural evolution for the remainder of the simulation. We can zoom in and determine where this stability, or lack thereof, comes from in each structure.

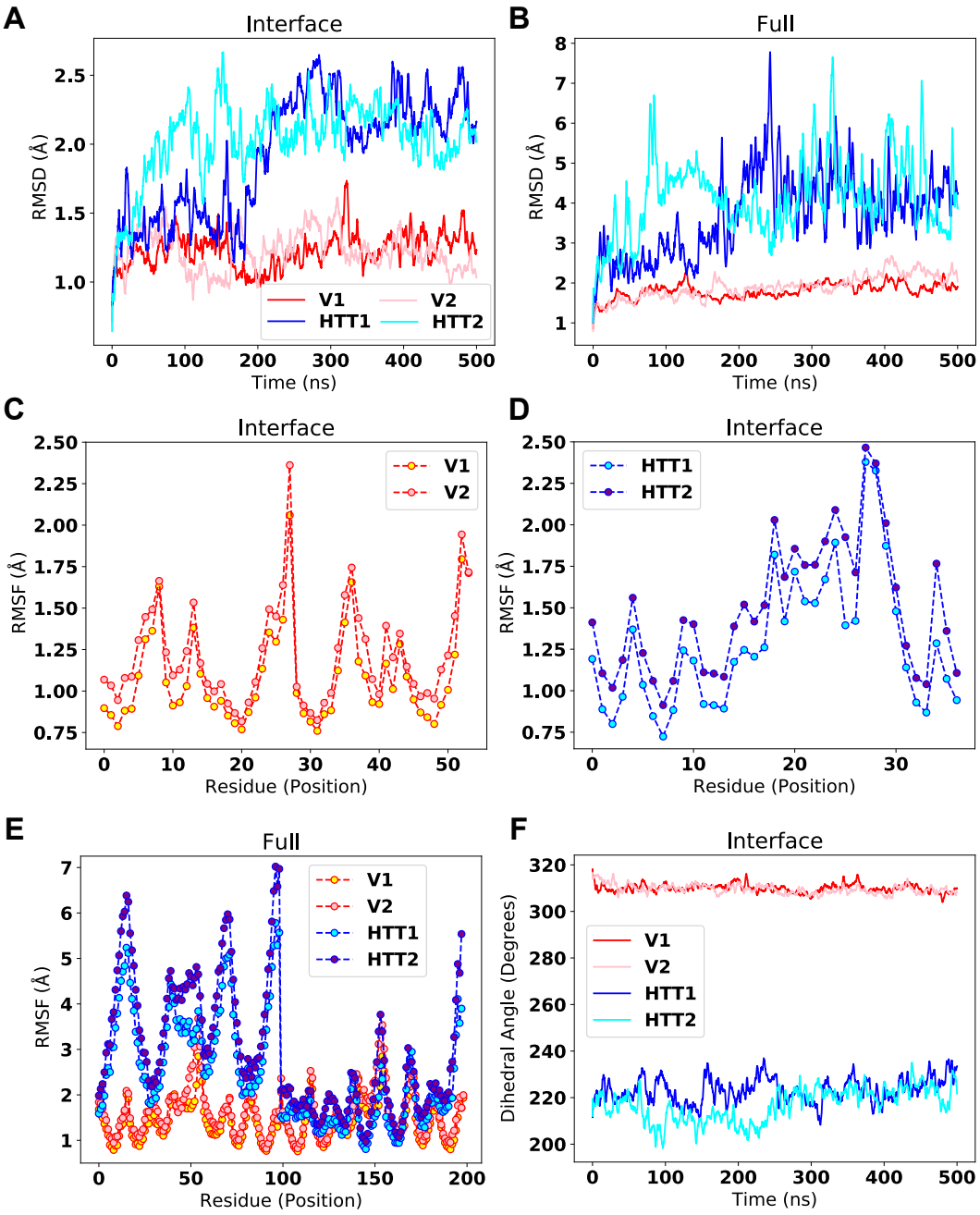


Figure 3.5: **All-atom molecular dynamics simulations of BTN3A1's extracellular domains highlight the exceptional stability of the v-shaped conformation and the relatively transient nature of the head-to-tail conformation.** (A) Root mean square deviation (RMSD) measurements of duplicate 500ns simulations of the v-shaped (V1,V2) and head-to-tail (HTT1, HTT2) dimeric interfaces. Running averages over 20 picoseconds are shown for clarity. (B) RMSD measurements of the full dimers of the v-shaped and head-to-tail conformations. Running averages over 20 picoseconds are shown for clarity. Interface root mean square fluctuations (RMSF) for the v-shaped (C) and head-to-tail (D) conformations. (E) Full protein RMSF measurements. (F) Measurement of the dihedral angles between interfacial domains in all four simulated systems through time.

As discussed in Chapter 2, root mean square fluctuation (RMSF) is a useful metric for tracking the displacement of individual resides averaged over entire trajectories. The interfacial residues are defined as any residues on one monomer within 5.0 Å of the corresponding dimeric partner. Due to this imprecise definition, we see that there exist regions of 5-10 amino acids in the v-shaped RMSF trace (Figure 3.5C) that are incredibly stable through time with an RMSF under 1.0 Å. These regions are comprised of the backbone interactions of the v-shaped dimeric interface, yet again attesting to the exceptional stability of these interactions. No such regions of local stability exist within the head-to-tail interface (Figure 3.5D), and RMSF measurements across the entire dimeric complexes (Figure 3.5E) further underline the dramatic stability differences between each conformation. Lastly, we see that over the course of these simulations, the head-to-tail interface is not maintained, as dihedral angle measurements between the two interacting domains shift through time, whereas v-shaped dimers maintain a dihedral angle between monomers of 315 degrees (Figure 3.5F).

Despite these striking differences, all-atom simulations on the order of hundreds of nanoseconds are often sufficient for sampling significant conformational changes in the protein structure but are unable to capture folding and unfolding events or protein binding events [127]. Further coarse-grained simulations help to fill this gap in sampling, serving to account for possibilities that short term dynamics do not necessarily predict the longer-timescale behavior. In this work we utilize Upside, a coarse-graining software developed in the Sosnick lab at the University of Chicago [128]. The Upside simulation package is unique in its powerful machine-learning based approach to force-field building and the relative simplicity of the model used [129]. We can further enhance sampling using replica exchange molecular dynamics, a powerful technique to sample a protein's conformational landscape across a wide range of temperatures. Replica exchange MD will be discussed at length in Chapter 4.

Simulations were run by Nabil Faruk, a graduate student in the Biophysical Sciences program at the University of Chicago, using replica exchange molecular dynamics simula-

tions on Upside. Throughout the course of these simulations we are able to sample each simulated system from their crystal structure to a nearly unfolded state. As a first attempt at characterizing this folding pathway, and to specifically look at which parts of the proteins unfold, we can compare the radius of gyration (R_g) to the fraction of protein replicas that remain in the bound state at each tested temperature. By plotting these two metrics (Figures 3.6A,B) we can see via the R_g that the head-to-tail conformation begins to unfold much sooner than the v-shaped conformation, but once the protein begins to unfold, they each evolve in a similar manner as a function of temperature.

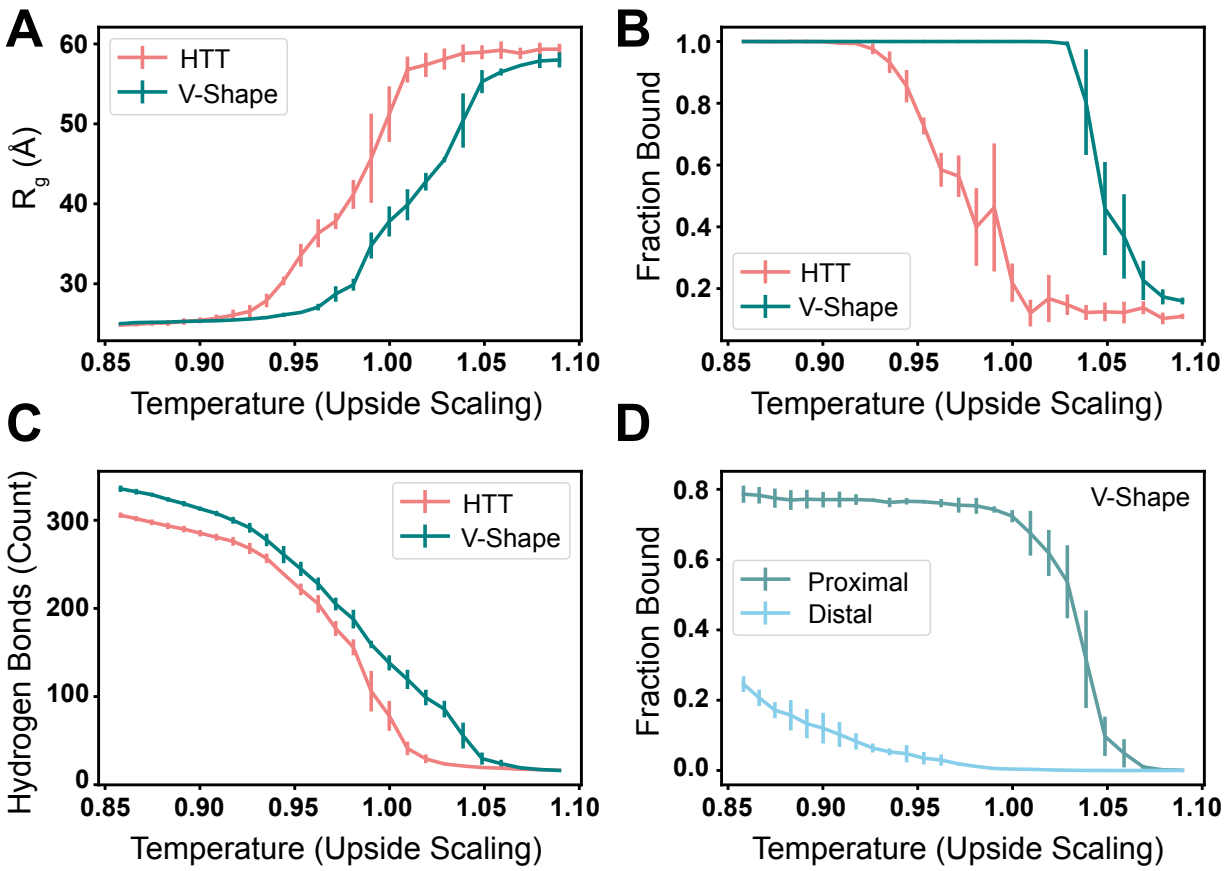


Figure 3.6: **Coarse grained Upside simulations show the v-shaped conformation is considerably more stable than the head-to-tail conformation over long timescales.** (A) Radius of gyration across all tested temperatures of coarse grained umbrella sampling simulations. (B) Fraction bound, i.e. probability of maintaining dimeric interaction, over all tested temperatures. (C) Average hydrogen bond count over the course of Upside simulations. (D) Fraction bound of both the distal (IgV) and proximal (IgC) domains of the v-shaped conformation. See figure 3.2 for reference of structure.

Conversely, we see from the bound fraction quantification that the v-shaped conformation remains a dimer much longer than the head-to-tail conformation, confirming the improved stability of the v-shaped conformation. Indeed, it appears from these data that the monomers adopting the v-shaped dimeric state begin to unfold before the dimeric interface is dissolved. From these data, we can see that the proteins are undergoing significant conformational changes as the temperature is increased, to the point that nearly no hydrogen bonds remain at the highest tested temperature (Figure 3.6C). This suggests a completely unfolded state of both proteins at the highest temperature. At lower temperatures, we find that interactions between the distal domains are slightly increased, persisting just enough to suggest weak interactions (Figure 3.6D).

These data suggest an incredibly stable v-shaped interface, and one that is potentially irreversible over the lifetime of a single dimer complex residing on the surface of a cell. This calls into question the prevailing model for activation. How could the information of a small molecule binding to the intracellular B30.2 domain of BTN3A1 be transmitted through a single-pass transmembrane helix, and how could this single helix provide a strong enough allosteric effect to alter such a stable interface? Similarly, if the v-shaped extracellular conformation exists independent of the presence of pAg, as suggested by the work of Gu et al. [116] and the results of Figure 3.3, then any model of activation would operate on an assumption of some change in the thermodynamic equilibrium of the two conformational states. pAg would then function as a catalyst, shifting this equilibrium away from the head-to-tail conformation and towards the v-shaped conformation. Yet if the v-shaped conformation is nearly irreversible and the head-to-tail conformation is relatively unstable, as suggested by the above simulations, this equilibrium is already significantly unbalanced. Instead, it appears more likely that BTN3A1 is constitutively expressed as a v-shaped dimer, and that some other mechanism plays a key role in the activation of T cells.

3.3 An Intracellular Model for pAg-Induced Activation

In light of these results casting doubt on the activation model conditional on extracellular conformational change, a new model for activation is required. We can begin to build such a model by recontextualizing some of the results in the field, stripping away their implications towards a model based upon extracellular change and focusing instead on the basic results. We can start from the very beginning of the study of BTN3A1, in the work of Harly et al. As mentioned in the introduction of this chapter, BTN3A1 was first identified through an anti-BTN antibody that is capable of activating Vγ9Vδ2 T cells [55]. This activating antibody, referred to as monoclonal antibody 20.1 (mAb20.1), was found to immobilize BTN3A1 on the surface of live cells using fluorescence recovery after photobleaching [55]. Further results from Palakodeti et al. clarifies these findings, providing a possible mechanism for the activating potential of mAb20.1. Co-crystal structures of a single chain variant (scFv) of the 20.1 antibody bound to the BTN3A1 extracellular domain suggest that the full mAb may crosslink butyrophilin molecules [57]. Figure 3.7 depicts a new model based upon these experimental data and incorporating some of the hypotheses investigated in this work, whereby clustering and immobilization of BTN3A1 is driven by intracellular changes induced by phosphoantigen binding.

This new, clustering based activation mechanism is consistent with the results so far in the field, which have been primarily focused on the intracellular consequences of pAg binding [56, 116, 130, 131], fundamental differences in antibody- or pAg-based activation mechanisms [58], or the (recently successful) search for a TCR-contacting molecule [132]. Recently however, the work of Yang et al. has revitalized this conformational change model. Through new crystallographic results, Yang et al. suggest that a single rotomeric shift of histidine 351 (H351) within the B30.2 domain of BTN3A1 may have implications for activation [133]. The authors go on to suggest that this shift leads to a larger allosteric effect that then results in a significant extracellular conformational change.

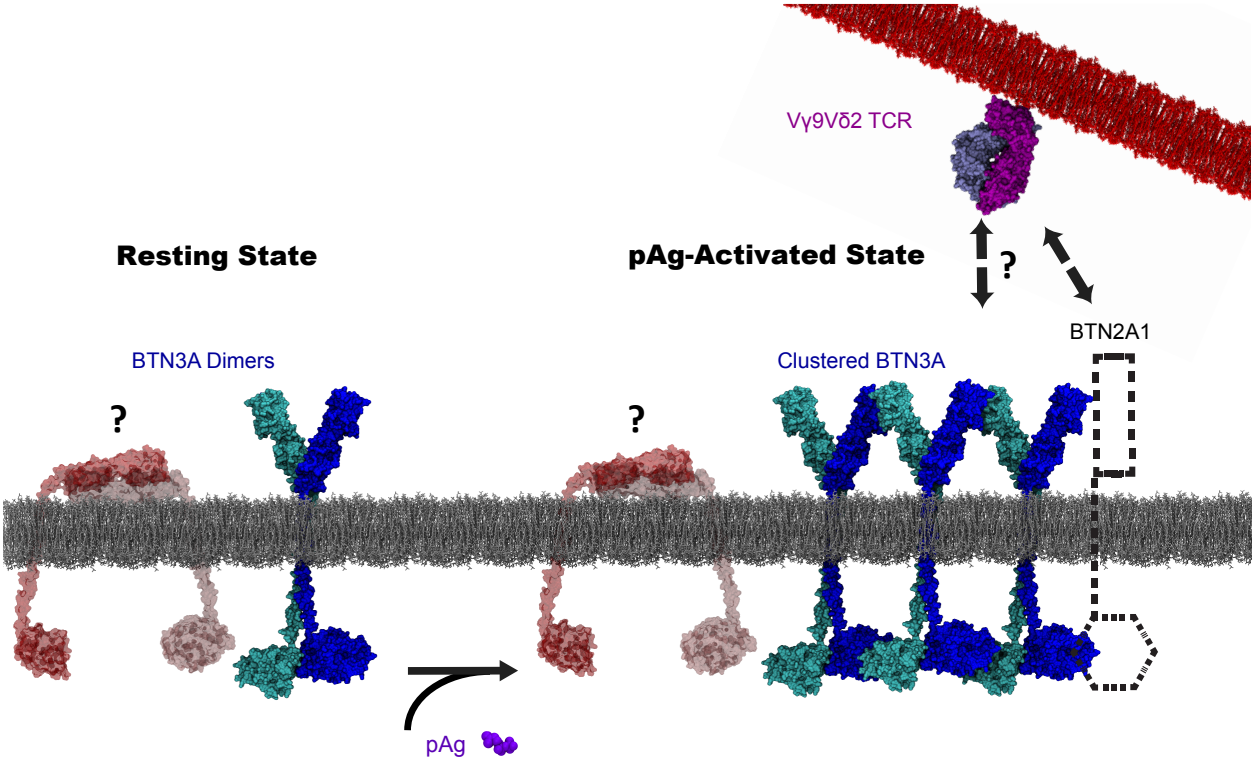


Figure 3.7: **A new clustering-based model for BTN3A1-mediated Vγ9Vδ2 T cell activation.** In this model, we assume the majority of BTN3A1 homodimers on the cellular surface, both active and inactive, exist in a v-shaped conformation (blue and cyan molecules). While we cannot rule out the existence of the head-to-tail conformation of the BTN3A1 homodimer (red and pink), we expect that it does not play a role in pAg-dependent activation of T cells. Instead, pAg binds the intracellular B30.2 domain of the v-shaped homodimer, inducing clustering and potential recruitment of BTN2A1. The role of BTN2A1 in T cell activation is discussed further in the discussion.

Yet as suggested by the results outlined in in Section 3.2, these extracellular changes appear unlikely. Is a single histidine rotomeric shift truly responsible for initiating T cell activation? If not, what are the consequences of pAg binding to the B30.2 domain, and how might this information be transmitted to the extracellular surface? Again, these questions readily lend themselves to interrogation using computational approaches.

3.3.1 Characterizing pAg Dynamics Within the B30.2 Binding Pocket

Working from the starting crystal structure of Yang et al., we can quickly test their proposed hypothesis of H351's key role in the transmission of the signal of pAg binding

to the extracellular surface. Starting from the crystal structure of HDMAPP bound to the B30.2 domain (PDB ID: 5ZXK), we see that the hydroxyl oxygen of the ligand neatly contacts the π-nitrogen of H351 [133]. We would expect that if H351 acts as a "gatekeeping" residue, this mechanism should persist across all activating phosphoantigens. As such we model cHDMAPP, a synthetic form of HDMAPP [134], into an identical binding pose within the pocket of the B30.2 domain. These two all-atom simulation systems are fully solvated in explicit water with 0.15M NaCl using CHARMM-GUI [123, 124] and run in triplicate for 500ns of total simulated time. Further detail can be found in the Appendix.

Figure 3.8A shows a time trace of a representative replica of the distance between cHDMAPP's hydroxyl and H351. Nearly immediately, within the first few nanoseconds of simulated time, we find that the cHDMAPP-H351 interaction is unstable. Shortly thereafter, the hydroxyl seems to move about in an unstructured manner, before adopting what appears to be a semi-stable state roughly 15 Å away from H351. However, this interaction is also relatively short lived, giving way to yet more chaotic motion through the remainder of the trajectory. The cHDMAPP hydroxyl does appear to interact with H351 at multiple time points throughout the full 500ns trajectory, but only does so briefly. While the chaotic motion throughout the majority of the trajectory suggests the cHDMAPP hydroxyl is freely diffusing, the semi-stable state adopted in the short time frame between 20 and 80ns likely occurs via a distinct interaction elsewhere within the B30.2 domain's binding pocket. However, this interaction is also relatively short lived, giving way to yet more chaotic motion through the remainder of the trajectory.

We can identify this distinct state more easily by looking at a representative distance trace taken from the HDMAPP simulations. Figure 3.8B shows not only the 07-H351 distance, but also 07-E416 distance. This measured distance between the glutamic acid and the HDMAPP-hydroxyl is complementary to the 07-H351 distance, identifying E416 as a second significant interaction partner of pAg.

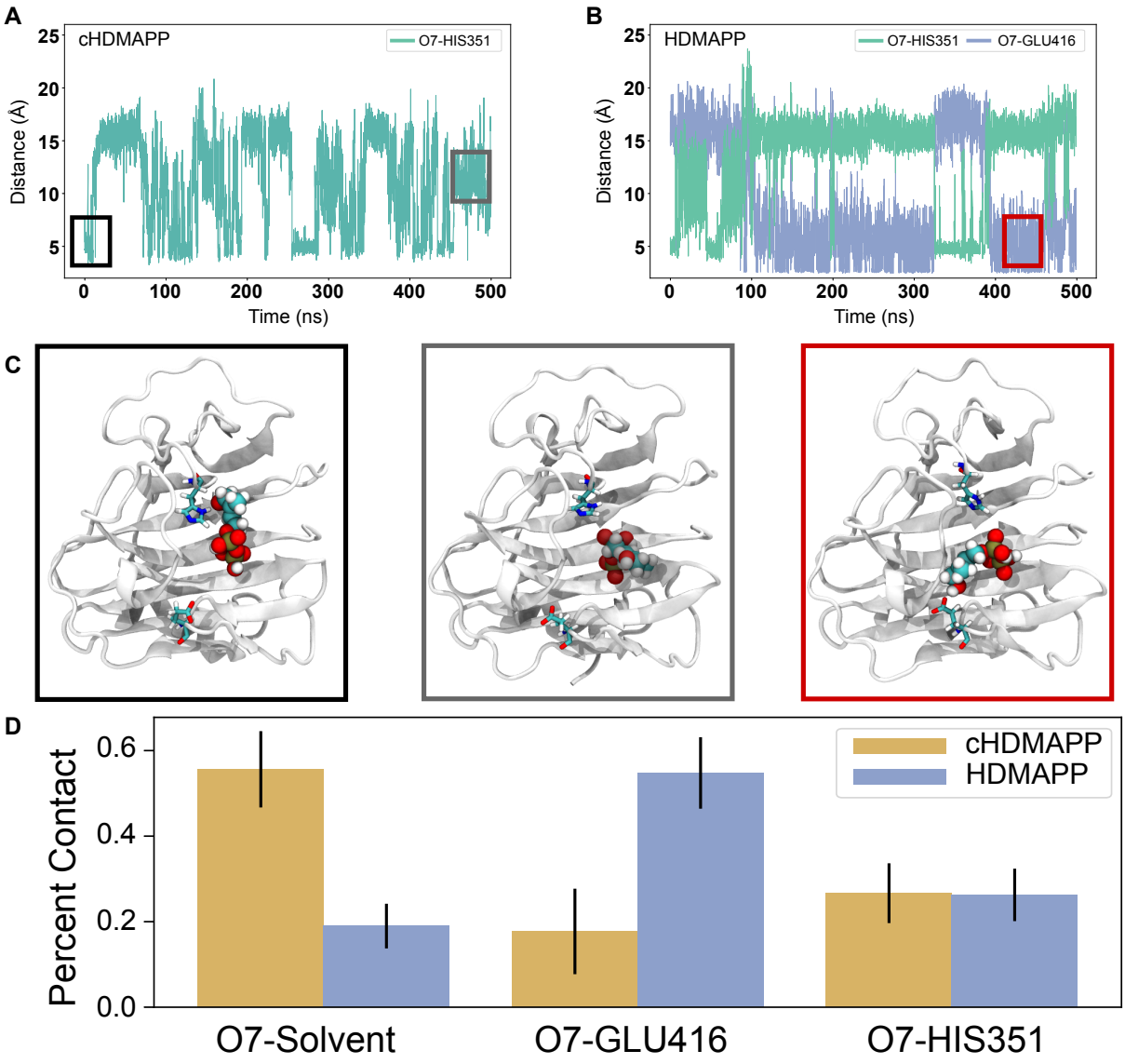


Figure 3.8: **Molecular dynamics simulations of pAg bound to the B30.2 domain show that the hydroxyl tail of HDMAPP and cHDMAPP dynamically sample the binding pocket.** (A) Distance trace through time between the cHDMAPP-hydroxyl (07) and His351 . (B) Distance trace through time between the HDMAPP hydroxyl (07) and both His351 and Glu416. (C) Representative snapshots of pAg-B30.2 simulations. Border colors match the highlighted regions of the distance traces in (A) and (B). (D) Percent contact, i.e. each interacting partner at a distance of under 5.0 Å, of cHDMAPP and HDMAPP with solvent, Glu416, or His351.

When looking at these structures in closer detail (Figure 3.8C) we see that the terminal hydroxyl is the key mediator of each of these interactions. E416, and the histidine-distal region of the pocket as a whole, have been shown previously to play a significant role in

the activation capabilities of the B30.2 domain [114]. Again, we see that these results are consistent across triplicates. We can quantify how each of these triplicate simulations behave and see that on average cHDMAPP spends more time sampling space outside of the pocket than either of the two distinct regions of the binding pocket (Figure 3.8D).

It thus appears that H351 is not the sole mediator for the activation of T cells by pAg. While H351 may play a role, pAg is much more dynamic within the B30.2 binding pocket than has been previously proposed. Instead, pAg's intermittent contact with E416 appears to be equally important, and ligands that rarely contact either region (cHDMAPP) are still capable of activating T cells. Looking across all three triplicates for cHDMAPP and HDMAPP (Figure 3.9), the data show that HDMAPP contacts H351 and E416 nearly equally, while cHDMAPP contacts neither region for a majority of the simulated time.

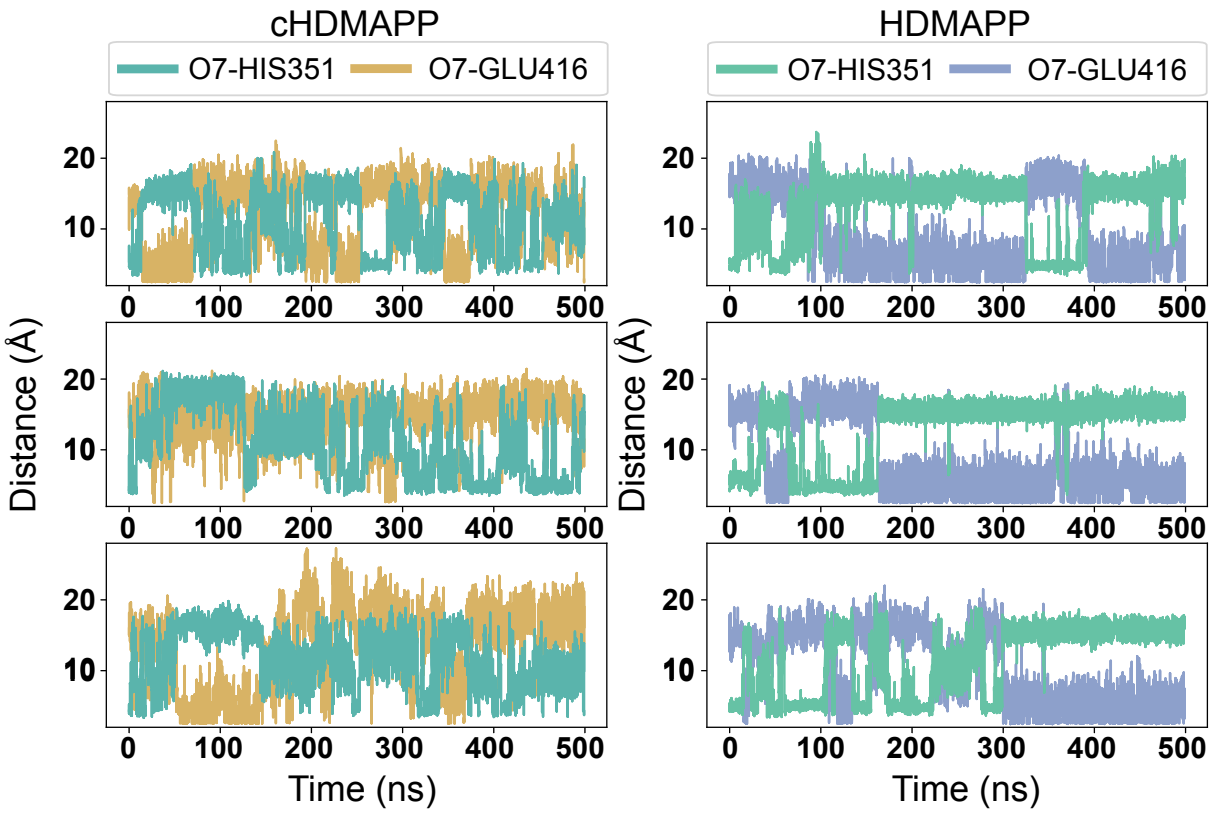


Figure 3.9: **Distance traces between pAg and either His351 or Glu416 through time show consistent behavior across molecular dynamics triplicates.** Triplicate 1 of cHDMAPP and HDMAPP (top) are highlighted in Figure 3.8.

While both of these ligands are able to activate T cells, HDMAPP is the more potent activator [134], suggesting contact with H351 and E416 may augment the activation potential of antigenic diphosphates. Additionally, the key mediating chemical group for this interaction, hydroxyl, is not present in isopentenyl pyrophosphate (IPP) an endogenous phosphoantigen which binds less tightly to BTN3A1's B30.2 domain and is a much less potent stimulant than HDMAPP or cHDMAPP [56, 130, 134]. This suggests pAg, and quite likely the diphosphate group within each pAg, may initiate some other molecular change upon binding to the B30.2 domain.

3.3.2 Investigating B30.2 Structural Changes in Response to pAg Binding

Using cHDMAPP as a test ligand, we can more clearly define the subsequent molecular consequences of pAg binding independent of the contacts made with H351 or E416, providing a broader understanding of how ligands such as cHDMAPP and IPP activate T cells. Initializing these simulations with the cHDMAPP hydroxyl positioned outside of the pocket, we ran 500ns of all-atom simulations in periodic simulation boxes fully hydrated in explicit water with 0.15M NaCl. Two distinct systems, a B30.2 domain with cHDMAPP bound and one without, are run in triplicate. Given the proposed models in the field, we hypothesize that cHDMAPP should induce some local and potentially global conformational change, and that these changes should become more evident when comparing to an apo-structure.

Similar to the simulations focused on antigen dynamics, we rapidly see differences in the observed protein dynamics between the apo- and pAg-bound simulations. Visualization of representative frames from these trajectories show that pAg keeps the binding pocket of the intracellular domain in a more contracted state, while the apo systems are free to adopt a more open, flexible conformation (Fig. 3.10A). Root mean square deviation (RMSD) backbone analysis of these simulations, which reveal structural shifts in the protein backbone over time, show that overall the crystal structure is stable in solution.

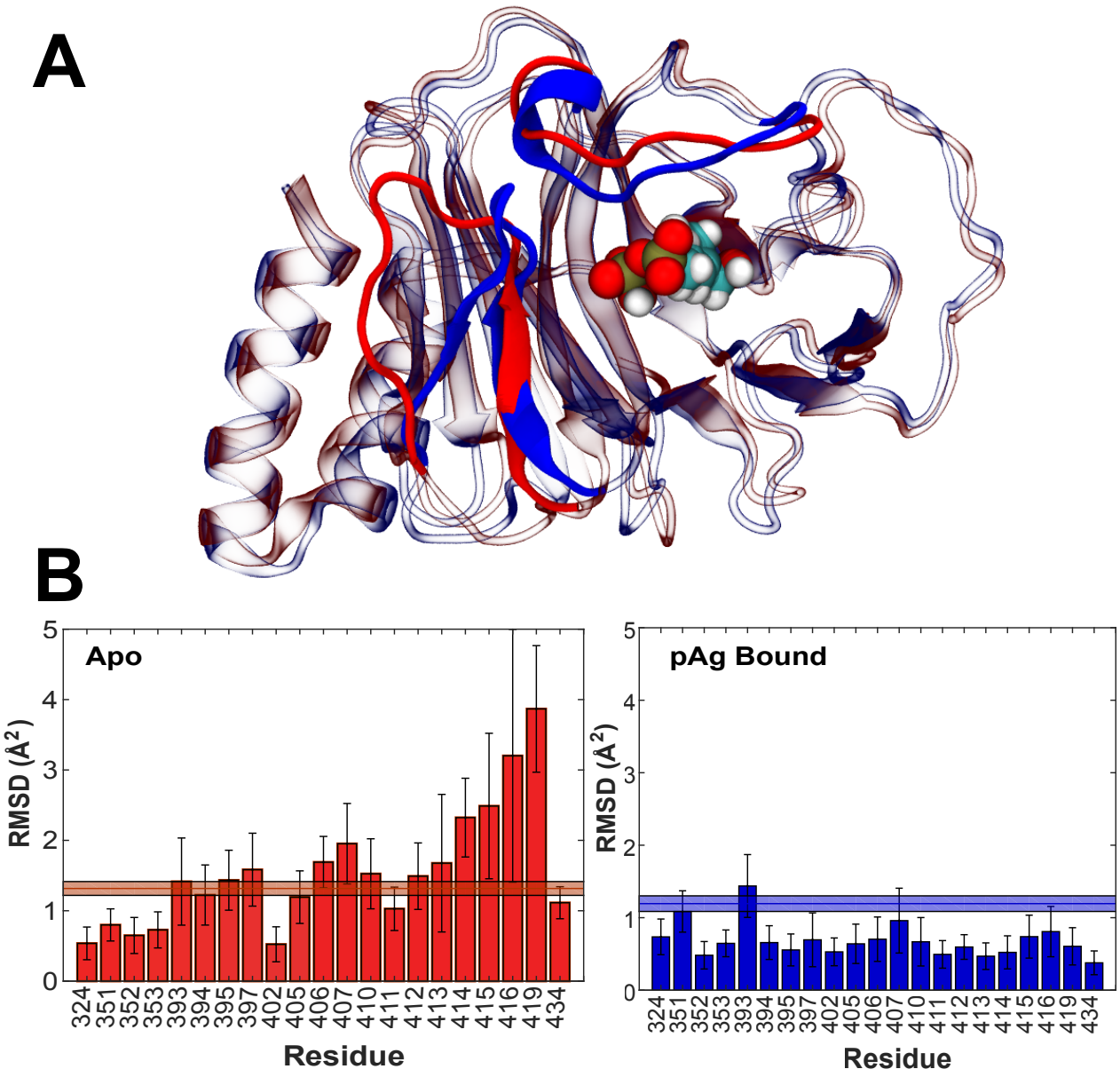


Figure 3.10: **All-atom MD simulations reveal dynamic and structural differences between B30.2 domain apo and pAg-bound state.** (A) Coordinates of the BTN3A1 intracellular domain with (blue) or without (red) pAg cHDMAPP (VDW spheres) after 500 ns of MD simulations are shown in cartoon representation. The segments 393–397 and 410–419 that exhibit major backbone RMSD shift are shown as ribbon representation in red (apo) or blue (pAg-bound). (B) Single amino acid backbone RMSD of the residues identified to experience high CSP upon pAg binding. The thick horizontal line in each plot is the mean plus or minus SD of the backbone RMSD of the rest of the protein.

These simulations have a mean deviation between crystal structure and simulation of 1.81 Å for the apo systems and 1.57 Å for the pAg-bound systems (Figure 3.10B). While the overall structure is stable, single-residue RMSD analysis of the backbone showed that the residues flanking the pAg binding pocket show significant local variation throughout the course of the simulations (Fig. 3.10B).

The specific residues showing the most prominent shifts in silico are along two distinct flexible loops encompassing residues 393–397 and residues 410–419. Chemical shift perturbation measurements comparing apo- and pAg-bound B30.2 domains using nuclear magnetic resonance highlight these same two flexible loops as the regions of strongest change upon pAg binding [116]. This strong agreement with experiment provides further confidence in our computational results. Throughout the course of the simulation, the pAg-bound B30.2 structure is rigid and stable, primarily through interactions between the pAg diphosphate moiety and two arginine residues on the most flexible loop within the binding pocket. This more stable, contracted structure remains close to the crystallographic state, whereas the apo-simulation displays flexible motions in these flexible loops flanking the binding pocket. Both effects are pronounced and persist throughout the duration of the simulations. Potentially, this stabilization of a crystal-like conformation of the B30.2 domain may prove entropically favorable for dimerization partners critical for T cell activation.

One such dimerization partner for B30.2-pAg complexes may be yet another B30.2 domain of BTN3A1. The existence of homodimer forms of BTN3A1's B30.2 domains has been suggested previously in the work of Gu et al. [116]. From crystal structures of the BTN3A1 full-length intracellular domain (BFI), which includes both the B30.2 domain as well as a portion of the juxtamembrane region, two putative dimer forms have been identified within the crystal lattice [116]. These two dimer forms are simply referred to as B30.2 Dimer I and Dimer II, with Dimer I adopting an asymmetric dimer interface overlapping with the pAg-binding pocket of one monomer (Figure 3.11A) and Dimer II utilizing the N-terminal helices to form a symmetric interface (Figure 3.11B).

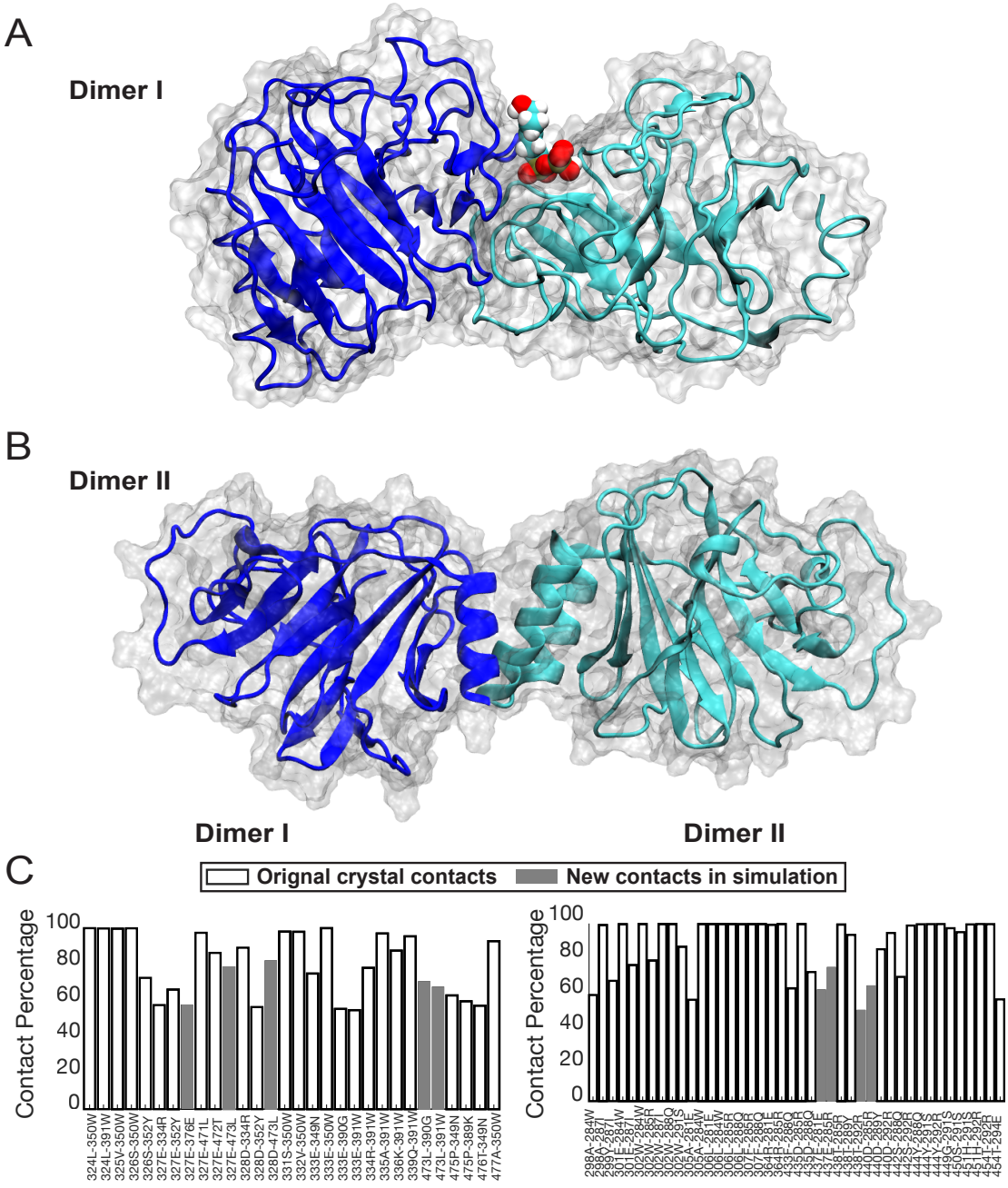


Figure 3.11: **The 1.9-Å structure of BTN3A1 BFI reveals two potential dimer interfaces.** (A) Overview of two dimer conformations observed in the crystal lattice of the BTN3A1 BFI domain. The dimer observed in the asymmetric unit (Top) is referred to as Dimer I. The other dimer (Bottom) similar to the previously published B30.2 structure is referred to as Dimer II. The pAg-binding pockets are indicated by cHDMAPP model (VDW spheres). (B) Quantification of contact persistence throughout 500ns MD simulations within the dimer interfaces of Dimer I and Dimer II. Contact percentage is defined as the number of simulation frames in which two residues are separated by 5 Å or less divided by the total number of frames in the simulation.

Using MD simulations we set out to probe the stability of these dimer forms. With the two BFI dimer interfaces as starting points for our simulations each system was equilibrated and then run in duplicate at 293.15 K in replica 1 and 303.15 K in replica 2 until an accumulated total trajectory time of 500 ns was reached (see Appendix for details).

In the case of both interfaces, multiple contacts persist throughout the entire 500 ns trajectory, suggesting that both Dimer I and Dimer II are stable over this time frame. Visualizations of the simulated trajectories and analysis highlighting persistent contacts indicate that the hydrophobic interactions of the Dimer I interface and the inter-helix hydrogen bonds and salt bridges of the Dimer II interface are both stable (Figure 3.11C). Importantly, both of these simulations are run without pAg bound in the binding pockets of either monomer within the dimeric interface. In the case of Dimer II, this lack of pAg in the simulations is unlikely to alter the dynamics significantly, as the pAg-binding pockets of each monomer are sufficiently far from the dimer interface. While the timescales tested in these simulations are too short to probe global conformational effects, this interface appears to be stable throughout the course of the simulation. Conversely, the interface of Dimer I overlaps with the pAg-binding pocket of one of the B30.2 monomers. Notably, one of the identified regions stabilized by pAg, residues 393-397, directly contact the opposing monomer in the dimeric interface, further providing an avenue for a B30.2 dimer form stabilized by pAg.

3.3.3 Quantifying the Precise Contribution of pAg to Intracellular Dimerization

To obtain the clustering suggested by our model in Figure 3.7, a driving force must somehow act upon BTN3A1's extracellular domains. Previous studies have suggested that this may occur through interactions with the cytoskeleton [130, 131] but no evidence has been provided demonstrating direct interactions between BTN3A1 and any cytoskeletal proteins. Instead, given the stabilization pAg induces upon the binding pocket of the B30.2 domain, this extracellular clustering may occur via strengthened intracellular interactions.

However, while crystal structures have previously suggested that BTN3A1's B30.2 domains may dimerize, other experimental approaches including NMR and ITC show no indication of protein dimerization induced by the addition of pAg [116]. Again, we must utilize molecular dynamics simulations to probe regimes inaccessible to current experimental approaches, specifically within the regime of weakly interacting proteins. Here, we utilize a more nuanced approach than the typical "brute force" MD simulations, and instead use a technique called the string method to calculate the binding free energy between B30.2 homodimers.

Briefly, the string method is a means of calculating the binding free energy for a molecular interaction using a more natural pathway of unbinding (Figure 3.12). In many modern calculations of binding free energy, the two interacting molecules are pulled apart from one another in one dimension, generally along some radius r, increasing the distance between the two proteins in a straight line [135,136]. At some regular interval along r, new simulated "windows" are generated (i.e. one simulation with proteins separated by $r = 0$ Å, by $r = 1$ Å… by $r = 40$ Å) until the two molecules are no longer interacting. A harmonic restraint applied to each window maintains the proteins at each separation distance, with some freedom to oscillate about that harmonic center. Over the course of simulations at each window position, a bias to one side of the harmonic well will become apparent, particularly for those windows that are closer to the original bound structure. This process, referred to generally as umbrella sampling, is a means of sampling across the entire unbinding pathway.

We can speed up the convergence of this sampling along the unbinding pathway using a replica exchange scheme, in which harmonic restraints between neighboring windows are swapped with some probability. In this way, we obtain multiple simulations traversing the full binding/unbinding pathway, building up sufficient statistics for our final calculation. This final calculation is frequently completed through the use of the potential of mean force (PMF) formulation [135].

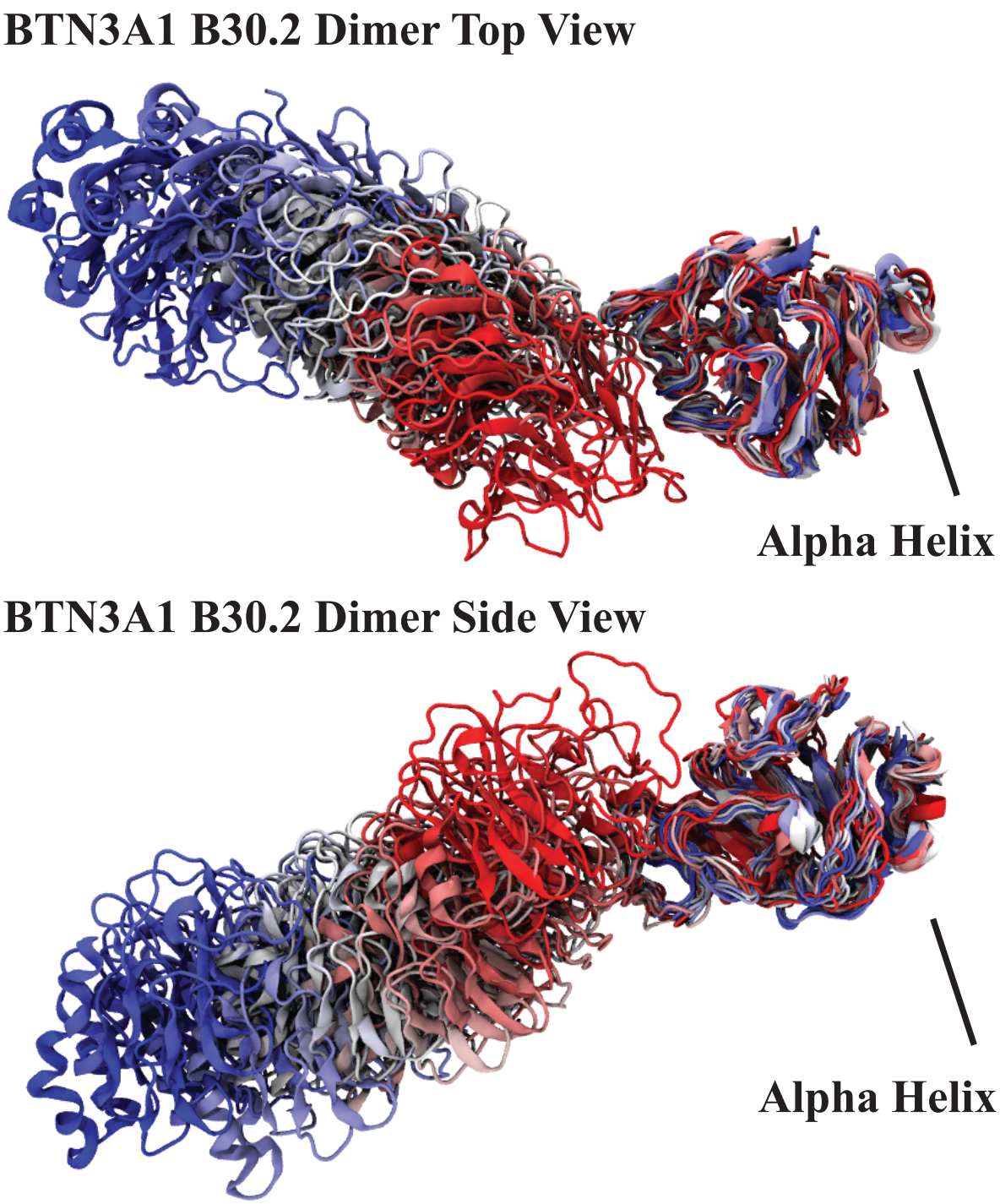


Figure 3.12: **Visualization of the most probable transition pathway generated by the string method highlights the curvilinear unbinding path of the B30.2 domains.** Each window along the string is represented as a color gradient from red to blue. Red structure represents the crystal structure, while the blue structure represents the final, unbound pose.

The general form of the PMF, $W(\xi)$ can be seen in equation 3.1:

$$W(\xi) = W(\xi^*) - k_b T \ln\left(\frac{< p(\xi) >}{< p(\xi^*) >}\right) \tag{3.1}$$

Here k_b is the Boltzmann constant, T is temperature, $< p(\xi) >$ represents the distribution function over some reaction coordinate ξ, and $*$ denotes arbitrary reference parameters. In the case of the simple one-dimensional umbrella sampling outlined above, this reaction coordinate ξ is simply r. While the PMF $W(\xi)$ is rarely calculated directly from molecular simulations, practically we can calculate the PMF based upon the probability distribution over each window along the sampled umbrella pathway (i.e. $< p(\xi) >$) [135, 136]. The string method is very slightly different from this more classical approach. In the string method formulation, we do not pull in one straight dimension, but rather along a curvilinear path allowing angles between the molecules to change. By doing this, we allow for a more natural binding and unbinding pathway, allowing the molecules to relax in to the "most probable transition pathway".

Once this pathway of windows is defined, we then simulate umbrella windows and calculate the binding free energy in a manner similar to the classical rectilinear approach. In addition to calculated binding free energies, we can visually inspect the trajectories of these simulations – the literal motions of the atoms in each B30.2 domain – and determine which individual residues or atoms are contributing the most to the interaction. Specifically, these simulations allow us to define at a resolution unattainable by experiment how pAg alters the binding between these two domains.

Applying the string method to the crystallized, asymmetric homodimeric B30.2 complex with and without pAg bound provides an estimate of the binding free energy change, or $\Delta\Delta G$. Figure 3.13 shows the converged energy landscapes for the apo and pAg-bound calculations of the binding free energy.

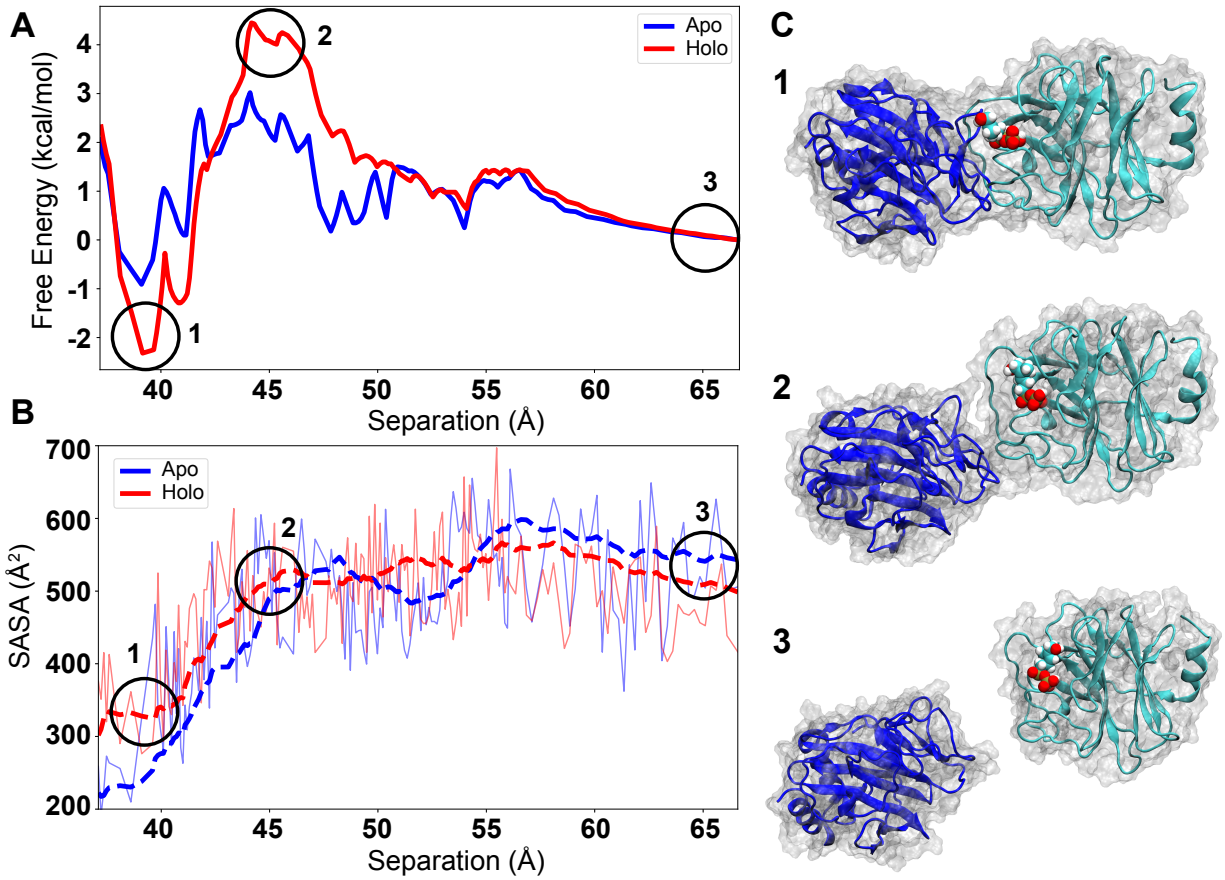


Figure 3.13: **The string method calculation of B30.2 domains with and without pAg show that pAg increases the free energy of binding of the homodimeric interaction.** (A) Free energy landscape along the most probable transition pathway of the apo and holo (pAg bound) simulations. (B) Solvent accessible surface area (SASA) of the pAg-binding pocket across the string. Identical x-axis as in (A). (C) Visualization of key points across the unbinding pathway. Numbers correspond to the circles of highlighting specific regions in (A) and (B).

Immediately, we see that the energy well of the pAg-bound system is deeper, suggesting a stronger interaction between the B30.2 monomers. Quantifying the ΔG of each system, accounting for the various restraints applied to each simulated system, we find that the apo system has a binding affinity of 4.53 kcal/mol, while the pAg-bound system has a binding affinity of 5.93 kcal/mol. ΔG_{apo} agrees well with experimental measurements of the apo binding affinity, on the order of 1mM (4 kcal/mol), using analytical ultracentrifugation [133]. ΔG_{holo} converted into molar affinity suggests a binding affinity on the order of 90μM, quite low by most standards of protein-protein interactions, but likely sufficient to

alter the dynamics of diffusive motion within the constrained two-dimensional context of transmembrane butyrophilin proteins.

Figure 3.13B highlights the change in the solvent accessible surface area (SASA) of the pAg binding pocket of each simulation through time. When comparing the peak of the energetic landscape to the running average of the SASA measurement, we see that the peaks of both traces are roughly incident at the same separation across the unbinding pathway. Figure 3.13C further breaks down this energetic landscape, providing clues into the precise role pAg plays in strengthening this interaction. We can see from the three representative structures that the lowest energy state corresponds to the fully bound structure, the proteins are completely dissociated at the furthest distance, while the proteins are still making strong contacts at the peak of the energetic landscape. As is evident from the rendered structure of this second highlighted state, which occurs at the onset of maximal binding pocket exposure and at the peak of pAg-bound free energy profile, the bound pAg is fully exposed to solvent. This, coupled with the similar energetic landscapes for the pAg-bound and apo-simulations from 50-65 Å separation suggest that the B30.2 domains may interact transiently in the apo-state, forming weakly interacting dimer interfaces. Upon introduction of activating levels of pAg in the cytosol, these transient dimers may become kinetically trapped within this dimeric state by introduction of pAg in the exposed binding pocket.

3.4 Discussion

Over the latter half of this past decade, the butyrophilin field and broader Vγ9Vδ2 T cell field have yielded few novel models for the role of BTN3A1 in the activation of T cells. Many new studies either avoid contextualizing their results explicitly into a complete activation model or attempt to tie these results into the incredibly influential work from the Adams lab [56, 57]. While this model proposed by Palakodeti et al. was consistent with the data within their study, there has been little effort expended outside of the Adams lab towards attempts to validate or contradict this model. Without a strong model for

activation, the field is feeling around in the dark, attempting to solve the puzzling case of butyrophilin-mediated T cell activation without strong guiding principles. In this work, we set out to first test the likelihood that the "inside-out" signaling initiated by pAg binding to the intracellular B30.2 domain initiates extracellular conformational change which is then recognized by Vγ9Vδ2 T cell receptors. Subsequently, we address gaps that remain in the model and propose a new form of the model that is more consistent with results obtained by other experimental researchers.

In an attempt to test the conformational change model, we first sought to experimentally determine the existence of the head-to-tail conformation of the extracellular domains of BTN3A1 outside of the constrained context of a crystal lattice. Using atomic force microscopy, a technique with nanometer-scale precision capable of probing the conformations of single molecules, we assayed full-length BTN3A1 solubilized in lipid vesicles or detergent micelles. These data are able to distinguish molecules that appear consistent with the physical properties of the v-shaped conformation, which has been previously observed using other experimental techniques and is likely the primary conformation adopted in solution [57,116]. While no evidence of the head-to-tail conformation has been observed, we cannot fully rule out its existence using these AFM results. The data using deposited lipid vesicles suffer from a low number of observed protein recognition events, likely due to a poor yield of reconstituted BTN3A1 molecules within the vesicle. While the solubilization of BTN3A1 in detergent overcomes this issue, this requires probing BTN3A1 within the less precise XY-axis of the AFM scan, adding significant uncertainty in our ability to confidently identify a conformation as v-shaped or head-to-tail.

To date, experimental approaches have proven incapable of resolving the head-to-tail conformation. Such a search-based approach makes ruling out the existence of this conformational state difficult, as the inability to find the state can be blamed on poor resolution or experimental conditions. Taking a different approach, we chose to interrogate the fundamental properties of each of the v-shaped and head-to-tail conformations of the BTN3A1

extracellular homodimer. Using molecular dynamics simulations, we are able to ascertain the relative stability of each conformation and contextualize these stabilities with the current models for BTN3A1-mediated T cell activation. Initial all-atom molecular dynamics simulations show that both conformations are stable on the hundreds of nanoseconds-scale, but that the v-shaped interface is exceptionally rigid when compared to the head-to-tail conformation. Further results from coarse-grained Upside simulations expand upon all-atom data, demonstrating that the head-to-tail conformation dissociates much sooner than the v-shaped conformation. Surprisingly, these Upside results suggest that the v-shaped dimeric interface persists even as the rest of the protein begins to melt.

The most current iteration of the extracellular conformational change activation model suggests that the head-to-tail and v-shaped conformations exist at some thermodynamic equilibrium with a shift in this equilibrium towards the v-shaped conformation induced by intracellular pAg binding. This increase in the population of v-shaped BTN3A1 homodimers then somehow confers an activating signal to Vγ9Vδ2 T cells. However, the lack of any experimental evidence for the existence of the head-to-tail conformation, in conjunction with these computational results, calls this model into question. If the v-shaped conformation is significantly more stable than the head-to-tail conformation, and the v-shape is indeed the "activating" conformation, this would represent a potentially leaky activation pathway. By chance alone, you would expect to occasionally find a majority population of v-shaped homodimers on the surface of cells, leading to activation of T cells when they should remain in a quiescent state. One would expect that the less stable state should require some trigger, forcing the protein to adopt this state, but it is highly unlikely that the head-to-tail conformation is able to activate T cells. Structural studies using the 20.1 mAb indicate that the binding of this activating antibody is sterically prevented from binding to the head-to-tail conformation [57]. These inconsistencies suggest a significantly different mechanism may be responsible for T cell activation.

We propose this new model should take the form of a clustering-based model driven

by changes of the intracellular dynamics induced by phosphoantigen binding. Biochemical and biophysical measurements have shown that the intracellular B30.2 domain of BTN3A1 undergoes significant changes upon the binding of negatively charged pAg to a positively charged binding pocket [56,116,133,137,138]. Our results outlined above provide an atomistic explanation for these experimental observations, specifically the NMR results of Gu et al. [116]. These simulations find that pAg contracts and rigidifies the B30.2 binding pocket, while the apo state is flexible and open. Immediately, this contraction and stabilization of an interface lends itself to the possibility of a more attractive binding partner for some activating ligand.

Our hypothesis is that this stabilized binding interface induces homodimerization of BTN3A1's intracellular B30.2 domains. An asymmetric homodimeric interface has been previously suggested as a key activating feature in the literature [116, 133]. In this work, we show that this asymmetric dimer does likely to exist, and that the binding free energy of this interaction is increased upon the binding of pAg within the dimeric interface. Using free energy calculations, we find good agreement between our computational results and experimental data, and further find that pAg increases the molar affinity between the B30.2 domains by two orders of magnitude. This higher affinity, on the order of 90uM, is weak enough to be consistent with the inability to observe the dimers experimentally, but strong enough to potentially alter the dynamics of BTN3A1 diffusing across the cell membrane.

These results, while significant, do raise further questions about the activation mechanism. Starting first with the question of binding interface stabilization, our molecular dynamics simulations suggest that the diphosphate moiety of pAg is primarily responsible for stabilizing the binding interface. Yet clearly, given the differential activating potential of HDMAPP, cHDMAPP, and IPP, the chemical composition of the pAg matters greatly for activation. Indeed, Salim et al. find that adenosine diphosphate (ADP), malonate, and citrate all bind to the B30.2 domain but fail to activate T cells [115]. Knowing that the asymmetric interface may be a crucial driver of T cell activation, it seems likely that the

significant bulk of the adenosine group of ADP interferes with this dimeric interface, as has been seen with synthetic bulky diphosphates [133].

Citrate and malonate, on the other hand, essentially mimic a diphosphate yet lack the acyl tail found in other phosphoantigens. While the crystal structures of these molecules bound to the B30.2 domain suggest they bind in a similar manner, NMR results show a decreased magnitude in the chemical shifts of residues K423 and R442 [115]. Perhaps the symmetry of the phosphate group is required to significantly contract the pocket, or potentially the tail group is necessary for the full contraction. As was seen in MD simulations, the aforementioned K423 is periodically contacted by the hydroxyl of cHDMAPP and HDMAPP, partially explaining why these molecules may activate while citrate and malonate do not. However, this explanation fails to describe the ability of IPP to activate T cells, given IPP's lack of a terminal hydroxyl. While the acyl chain does contact hydrophobic residues on the adjacent B30.2 domain during dimeric simulations, no single contact is long lived throughout the trajectory. Further experimental evidence is required to understand this question.

Similarly, further experiments are needed to test the hypothesized consequences of this new mechanism of activation. Our computational results suggest that upon addition of pAg, B30.2 domains should form more stable dimer interfaces, likely restricting diffusion and potentially inducing clustering on the cell surface. This clustering hypothesis has been tested recently, and no evidence for clustering was found [139] . However, a strong positive control was not included in this work, and the experiment was carried out in fixed cells, suggesting some form of immobilization upon pAg binding may be possible. While this clustering and immobilization has not yet been convincingly ruled out, localization of BTN3A1 to the immune synapse has been repeatedly observed [55, 56, 116, 139]. Is BTN3A1 trafficked to the synapse, or does the synapse form over immobilized, clustered proteins? Live cell microscopy, while challenging, is likely the most promising approach to carefully dissect this nuanced problem of dynamics.

As of this writing, we are confident that BTN3A1 is necessary but not sufficient for activation of T cells [140]. Indeed, BTN3A1 is the key sensor of pAg and in some way the extracellular domains are instrumental in conferring this activating signal to the T cell [55–57, 114, 116, 141]. For the majority of my thesis research, the missing "factor X" required for T cell activation remained a mystery. This past year, breakthrough work by Rigau et al. solved this decade-old conundrum, identifying BTN2A1 as the second target-cell molecule required for activation [132]. The pairing of BTN3A1 and BTN2A1 appear to be necessary and sufficient for T cell activation [132, 142], although this label of sufficiency has yet to be rigorously confirmed. While critical, this discovery intersects minimally with the results outlined above. Regardless of the ability of BTN2A1 to contact T cells directly [132], BTN3A1 is firmly entrenched as the pAg-sensing molecule. Working towards uncovering the downstream consequences of this pAg binding, as we have done here, represents a substantial step towards the full characterization of the pAg-butyrophilin-Vγ9Vδ2 T cell activation axis.

CHAPTER 4

COMPUTATIONAL DEVELOPMENTS IN BIOPHYSICS AND MOLECULAR IMMUNOLOGY

4.1 Introduction: The Role of Physics and Computation in Immunological Research

The previous chapters of this thesis focus primarily on the scientific insights provided by interdisciplinary computational and experimental approaches. These interdisciplinary approaches are becoming increasingly prevalent in the biological sciences across all fields. In fields such as neurobiology, population genetics, and molecular biology, there exists a strong connection to the quantitative sciences such as physics, mathematics, and chemistry. In these fields, significant efforts are expended yearly towards the design, development, and implementation of novel computational approaches, theoretical frameworks, and experimental techniques. Entire theses are spent perfecting pipelines, developing testable theories, and publishing code in proper repositories for widespread use and more open science. Yet despite what is recognized as an increasing need in immunology for this quantitative cross-disciplinary approach [143], the field remains largely reticent to fully embrace advances from researchers external to the field, save for the strong niche staked out by structural biologists and molecular biophysicists. However, some concepts from fields such as computer science, physics, and machine learning are beginning to find everyday use in the lives of experimental immunologists.

At some level, immunologists are increasingly beginning to recognize the complex, high-dimensional datasets acquired via flow or mass cytometry should be utilized to their fullest extent. Rather than discarding data through filtering strategies called "gating", immunologists are borrowing tools from physics and computer science to analyze these complex datasets using dimensionality reduction techniques such as t-distributed stochastic neighbor

embedding (tSNE) and uniform manifold approximation and projection (UMAP). In turn, this increased adoption of quantitative approaches by immunologists can create a welcoming environment enticing physicists to the field. Biophysicists are increasingly finding that immunology represents a rich, complex system that, historically, is infrequently interrogated by non-immunologists. Ideas passed between these two fields have resulted in significant contributions to our understanding of immunology as a whole.

Specifically, both supervised and unsupervised machine learning are being proposed as a means of standardizing cytometric analysis of cellular samples [144]. These cytometric datasets comprise intensities in as many as 40 dimensions per cell [145], where each dimension is a proxy for the expression level of some cellular marker. Quantitative approaches to the analysis of these datasets have led to breakthrough results in the study of nuanced classes of myeloid cells [146], the creation of a reference map for immune cells in multiple organs [147], and the implementation of a classification strategy for diagnosing hematopoietic diseases [148]. These high-profile applications of quantitative cytometric analysis can then lead to further improvements upon previous approaches [149–151].

Likewise, physicists have been able to contribute novel hypotheses for how the complex components of the immune system interact in an attempt to discriminate "self" from "non-self". Physical treatments of biological systems have provided insights into potential strategies for eliciting broadly neutralizing antibodies [152–154], an understanding of the complex biochemistry and physics occurring at the immune synapse [155, 156], and quantifications and estimations of immunoglobulin diversity resulting from V(D)J recombination [83, 157–159]. This dual approach to studying the immune system is necessary for understanding the highly complex network of signaling molecules, receptors, cell types, and coordinated reactions involved in immune responses.

Despite this germination of productive ideas between these disciplines, there is little evidence of extended collaborative work between individual physicists and immunologists.

Notable exceptions to this pattern can be found in the literature [160–162], but there exist few examples of a prolonged dialogue between the disciplines. Some of this lack of truly collaborative endeavors may simply be attributed to the recency of physicists starting to work on immunology. A testament to the potential impact of highly collaborative research between physicists and biologists can be found in neurobiology, where the joint work between two physicists, Aaron Lloyd Hodgkin and Andrew Huxley, with a neuroscientist, John Eccles, led to the 1963 Nobel Prize in Physiology or Medicine for the study of neuronal signal transmission [163, 164]. Physics and neuroscience research continue to be thoroughly intertwined; hopefully physicists and immunologists can one day find a similar common ground as research progresses.

In my own thesis research and moving forward in my scientific career, I intend to further this interconnectivity between immunology and physics through fundamental physical approaches and more modern computational methods. In this chapter, I will outline my work in developing physics-based approaches for the study of immunological systems of interest. These include the development of entirely new bioinformatic analysis pipelines and the standardization and automation of existing approaches to lower the barrier to entry for those starting out in the field. First, I will outline my work developing AIMS – An Automated Immune Molecule Separator, and the accompanying graphical user interface that holds promise as a novel analytical tool for the characterization of large immune repertoires. Then, my work adapting the well-known string method approach to calculating binding free energies into a more streamlined protocol will be outlined, with identifications of fundamental problems in the underlying computing architecture noted throughout. Lastly, my discussion will touch on the potential next steps in the continued adoption of physical approaches to the study of immunology.

4.2 AIMS – An Automated Immune Molecule Separator

This relatively recent increased interest in quantitative approaches in immunology can be explained in part by the decreased costs associated with sequencing technologies [165]. As a result, there has been an explosion of T and B cell receptor sequences published in a wide range of studies in humans and mice [166–169]. Despite the generation of these massive datasets, there is a distinct lack of commonly used analytical tools for the analysis of this data. Many studies resort to basic analyses, including simple observations of the average length of CDR loops, averages across entire structures of certain biophysical properties, and quantifications of the genes used by the sequences in the dataset. While these surface level analyses are frequently sufficient, they ignore a wealth of data that could provide deeper, exciting insights into the underlying biology at hand.

This lack of nuanced analysis in recently published research is not due to a dearth of available approaches. A large number of computational packages are currently available for the analysis of antibody and T cell receptor sequences. Some highlights include IgBLAST [84], IgOR [159], TCRdist [83], and IMGT's V-Quest [170], each of which provide their own unique contribution to the study of immune repertoires. Despite these fantastic resources, there is no one approach that has risen to prominence within the repertoire analysis literature. This lack of widespread adoption is potentially due to two distinct causes; a reliance on the user's ability to program, and the lack of diversity in analytical approaches offered by current software. This first issue is nearly insurmountable in the biological sciences, as any attempt to make a graphical user interface (GUI) takes control away from the researcher, while scripts written in programming languages require some expertise to utilize. However, more can certainly be done to diversify the approaches to data analysis taken by the available software packages.

In an attempt to address these hindrances to the widespread adoption of analytical techniques for immune repertoire analysis, I set out to generate a new analysis pipeline that

is both easy to use and orthogonal to existing software. In order to improve the usability, yet retain full control, of the software I adopted a hybrid approach to software distribution, packaging it both as a GUI and as a Python notebook. Those with little programming experience are able to take advantage of the GUI, yet still have some creative control using a feature allowing for the output of raw data. Additionally, all of the source code is provided for more adept programmers, giving them the freedom to choose which pieces they prefer to adopt into their own analysis pipelines. The first application of the software, and indeed why the software was created in the first place, is outlined in Chapter 2 [171]. In the following sections, I will discuss some of the unique approaches adopted for this novel analysis pipeline in further detail and outline a second simple, yet successful application of this approach.

4.2.1 Construction of a Bioinformatic Platform for Repertoire Analysis

The vast majority of computational tools for repertoire analysis generated to date focus primarily on the genetic characterization of antibody and T cell receptor sequences, tracing back the genetic shuffling and mutations that occur during V(D)J recombination and somatic hypermutation. While these analyses are helpful to uncover any initial biases of specific genes towards the recognition of certain pathogenic epitopes, like the V_H1-69 gene responsible for encoding the heavy chain of many influenza- and HIV-reactive antibodies [71, 72], they provide little insight into the mechanism of recognition. In this work, we focus primarily on the biophysical properties of the complementarity determining region (CDR) loops of the receptors of interest. Critically, we look at these biophysical properties not just as an average across CDR loops but instead using a position-sensitive approach. This position sensitivity, as can be seen in Figure 2.2, allows for a flexible analytical approach capable of a multitude of distinct characterizations of the data. The majority of these characterizations, outlined in Chapter 2, are relatively simple transformations of the data. More complex analyses, and the additional application of this software to non Ig-like molecules are outlined below.

4.2.2 Linear Discriminant Analysis as a Tool for Repertoire Analysis

In addition to the inherent advantages of retaining the position-sensitive information of the antibody sequences, the resulting matrix generated provides a high dimensional space that lends itself well to a machine-learning based approach. The primary machine learning algorithm we use in this work is linear discriminant analysis (LDA). LDA works in a manner conceptually similar to principal component analysis (PCA), reducing the dimensionality of a given dataset via a linear combination of the original dimensions. However, LDA takes one additional input, the label or class of each sequence. Whereas the objective of PCA is to identify the axes which maximize the variance in the dataset, LDA has the dual objective of maximizing the projected distance between two classes while minimizing the variance within a given class. While LDA is well adapted for classifying two distinct populations, it is susceptible to overfitting, unlike PCA [172]. Generally, we can sort binary classes by labeling each class in the matrix with either a "1" for the first class, or "0" for the second class. In our application of LDA we parse down the large number of input vectors first by removing highly correlated data, and then by using either PCA or an algorithm which selects the vectors with the largest average differences between the two populations. This reduction in dimensionality ensures the data are not being overfit, and the tunable number of input vectors allows us to control for overfitting in each individual application.

LDA analysis is versatile in its applications, and in this software package we utilize the method in two distinct modes. Visualizing the basic workflow in Figure 4.1, we see that the conceptualization is not overly complex. In the first mode, all of the available data is used as input with the output vector representing the features that best distinguish between the two complete populations. Plots of the data projected onto this vector (as in Figure 2.4A) represent the maximum achievable separation between the two populations for a defined number of input components from the given biophysical property matrix. In this mode, we are most interested in the weights α of Figure 4.1A, and which biophysical properties these weights are affiliated with. By doing so, we can identify the features which best delineate the

two classes, with larger weights corresponding to features that are most important for this delineation. In the second mode, we utilize LDA as a more canonical classification algorithm separating the data randomly into training and test groups. In this classification mode of operation, a combination of correlation analysis coupled with maximal average differences is used to parse input features, and a support vector machine (SVM) is used to generate the final classifier from these features. Here, we apply only the training pipeline of Figure 4.1A to the predetermined training group. Then, one by one, we apply the filters that best split this training data to test data that has not been used in any way to influence the classifier and identify whether the antibody should be polyreactive or non-polyreactive as in Figure 4.1B. Accuracy of the resultant classifier is then assessed via k-fold or leave one out cross validation.

The use of LDA in this analysis pipeline is motivated primarily by the requirement for increased transparency in the classifications of antibodies and T cell receptors of interest. While approaches like neural networks and multi-layered classification algorithms can lead to improved classification accuracies, they lack interpretability. In the biological sciences and particularly in immunology, we care not just for the proper classification for these sequences but also the biochemical, biophysical, and genetic patterns being selected upon by the classifier. What good is the proper classification of antibodies as flu-binders if we cannot use that classification to better design new and improved anti-flu antibodies? While machine learning is a powerful approach that is becoming more widely adopted by researchers in immunology, care must be taken that these techniques are not being misused, and that interpretations of these models are addressed with the appropriate caveats. Transparency in these models goes a long way towards achieving this goal, and linear discriminant analysis provides one of the most transparent machine learning approaches available.

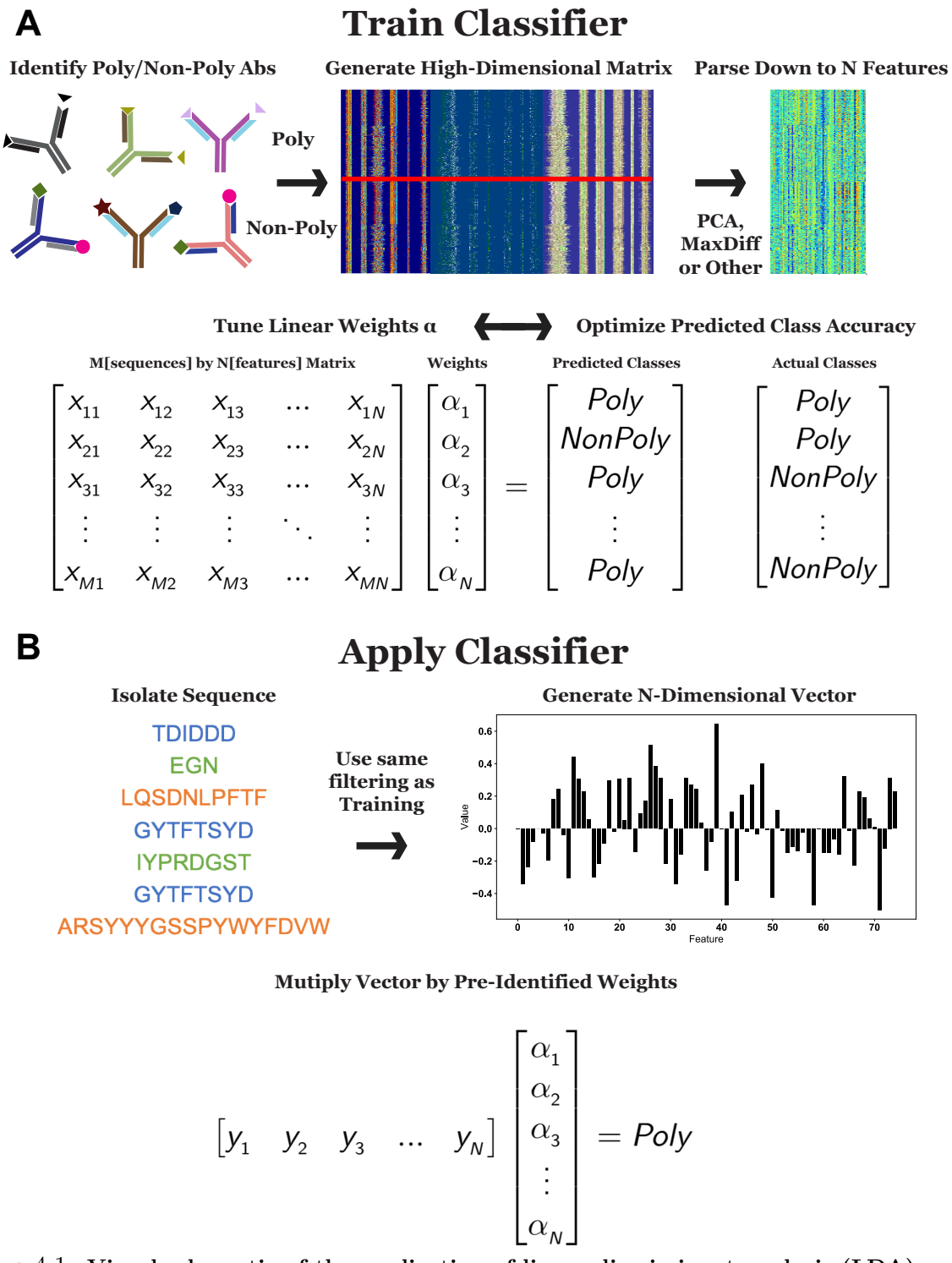


Figure 4.1: **Visual schematic of the application of linear discriminant analysis (LDA) used in this study.** (A) Representation of the training pipeline. (B) Representation of the application pipeline for future analysis.

4.2.3 Using Information Theory to Characterize Diversity and Crosstalk

While machine learning models are rising to prominence in immunology, there are other concepts from physics, mathematics, and computer science that can be readily applied to outstanding problems in the field. Information theory, a theory classically applied to communication across noisy channels, is incredibly versatile in its applications, with high potential for further applications in immunology [157, 161, 173–176]. In this bioinformatic pipeline, we utilize two powerful concepts from information theory, namely Shannon entropy and mutual information. Shannon entropy, in its simplest form, can be used as a proxy for the diversity in a given input population. This entropy, denoted as H, has the general form:

$$H(X) = -\sum_{X} p(x) \log_2 p(x) \tag{4.1}$$

Where $p(x)$ is the occurrence probability of a given event, and X is the set of all events. We can then calculate this entropy at every position along the CDR loops, where X is the set of all amino acids, and $p(x)$ is the probability of seeing a specific amino acid at the given position. In other words, we want to determine, for a given site in a CDR loop, how much diversity (or entropy) is present.

Importantly, from this entropy we can calculate an equally interesting property of the dataset, namely the mutual information. Mutual information is similar, but not identical to, correlation. Whereas correlations are required to be linear, if two amino acids vary in any linked way, this will be reflected as an increase in mutual information. In addition, due to some of the highly conserved residues in the non-CDR3H loops, high covariance can be achieved for residues that have not been specifically selected for in the germinal center. Using this information theory framework, these conserved residues have a mutual information of 0. Overall, the mutual information can be used to identify patterns in antibody sequences that were not readily evident through the previous analysis in this or other studies. If there is some coevolution or crosstalk between residues undergoing some selection pressure in the

antibody maturation process, it will be reflected as an increase in the mutual information. In this work, mutual information $I(X;Y)$ is calculated by subtracting the Shannon entropy described above from the conditional Shannon entropy $H(X|Y)$ at each given position as seen in equations 2 and 3:

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \tag{4.2}$$

$$I(X;Y) = H(X) - H(X|Y) \tag{4.3}$$

Broadly, these quantifications of information provide a strong theoretical foundation for the analysis of biophysical interactions. T cell receptors, antibodies, and peptide-major histocompatibility complexes (pMHCs) represent an ideal testing ground for this information theoretic approach, as their binding interfaces are almost perfectly conserved structurally, while the actual amino acid composition within this interface is highly variable. Future studies should aim to not just use information theory to describe specific systems, but instead to make predictions of the relative strength of interactions.

In his seminal work published in the 1940's, Claude Shannon used information theory to analyze the English language [173]. By examining the statistical nature of written English, i.e. the conditional probabilities of certain letter and word structures, Shannon was effectively able to simulate human speech and generate sentences that closely resemble natural language (Figure 4.2A). There is potential for the use of this same formalism in the study of TCR-pMHC interactions (Figure 4.2B). Rather than the 26 letters we use in the English alphabet, we can consider a replacement "alphabet" of the 20 amino acids. Similarly, the "words" in this formalism are formed by the closely interacting amino acid pairs of the cognate TCR and pMHC molecules. Each "sentence", i.e. amino acid sequence pair, then elicits either a weak, medium, or strong response. While this approach can apply to all

amino acids in this interaction interface, we can simplify our search space and focus solely on agonist, partial agonist, and antagonist peptides in a given TCR-MHC interaction. From this dictionary, we can generate predictions of the responses to *de novo* peptide sequences for a given TCR-MHC pair. The AIMS analysis pipeline provides not only an ability to analyze repertoires using information theory, but also a module to carefully characterize the amino acid frequencies of said repertoires in a position-sensitive manner. As such, AIMS may prove a useful tool in this more ambitious vision for the characterization of immune interactions.

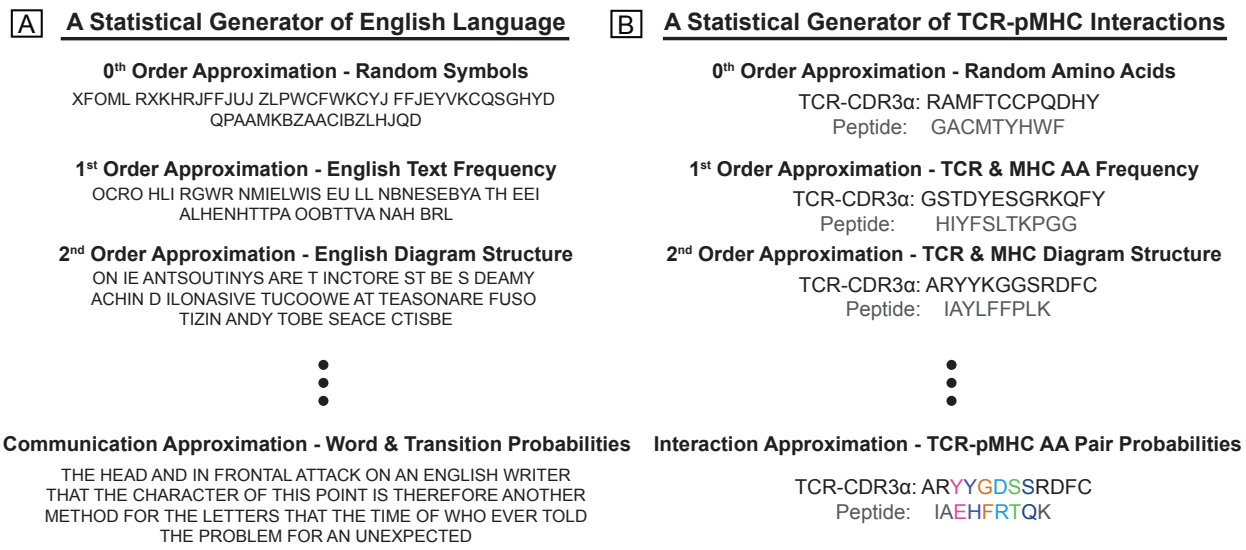


Figure 4.2: **Information theory can be used to study immune receptors in a manner similar to the analysis of natural language.** (A) Outline of the statistical generation of language as in Claude Shannon's *A Mathematical Theory of Communication*. (B) Analogous representation of a potential similar treatment of T cell receptors (TCR) and peptide-major histocompatibility complexes (pMHC).

4.2.4 An Application to the Identification of MHC-Like Molecules

Given the ability of this analysis pipeline to find nuanced differences between polyreactive and non-polyreactive antibodies, we next sought to expand the range of applications of our approach. Extension of the pipeline to the analysis of TCR sequences is trivial, due to the similar arrangement of CDR loops on the binding surface and the capability of IgBLAST

to annotate TCR sequences [84]. Instead, we sought to significantly expand the scope of this software by applying a similar approach to the analysis of MHC and MHC-like molecules. MHC molecules are encoded by a large superfamily of genes that are spread throughout the genome [177,178]. MHC and MHC-like genes are found across a wide range of divergent species, and these genes have diversified extensively over time, making the distinction between orthologs and instances of convergent evolution difficult. In some cases, the divergence is extreme enough that phylogenetics cannot provide predictions of function. Given that these MHC molecules have evolved to present different antigenic subtypes, such as lipid molecules in the case of CD1 proteins [179–181], we explored the use of our pipeline as a classifier based on biophysical properties rather than phylogeny.

As a test case, we use two example training classes; a representative list of human MHC Class I molecules, and the output from a BLAST query on CD1d [182]. We can then assess our ability to separate these two training classes, while also introducing a test class derived from the data of Almeida et al. [183]. Upon closer comparison between the structures of the classical MHC Class I molecule HLA-A and the non-classical MHC-like CD1d, we see that the two molecules look quite similar to the untrained eye (Figure 4.3A). While the CD1d structure clearly has a prominent α-helical kink, the other regions of each protein align nearly perfectly. However, a slice through each antigen binding pocket colored by amino acid biophysical property shows massive differences between the two molecules. We immediately see that the binding pocket of HLA-A (Figure 4.3B) is far more hydrophilic than the binding pocket of CD1d (Figure 4.3C), consistent with their roles as peptide and lipid binders, respectively. These two distinct classes that are strongly defined by the biophysical properties of their binding pockets represent an ideal test case for the new functionality of the AIMS software.

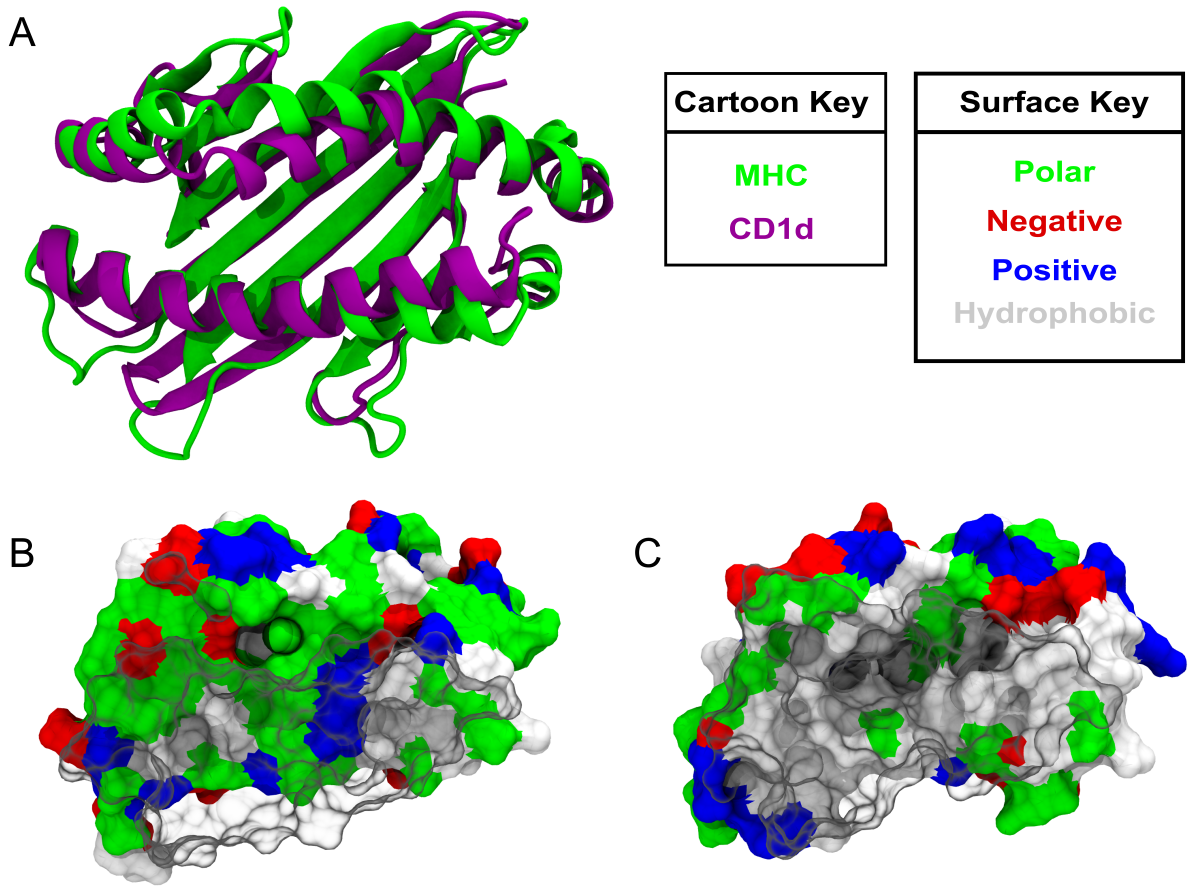


Figure 4.3: **A structural comparison of HLA-A and CD1d shows differences are primarily restricted to the biophysical properties of the amino acids in the antigen binding pocket.** (A) Cartoon representation of HLA-A and CD1d (PDB IDs: 2XPG [184], 1ZT4 [185]). (B) Surface representation of HLA-A (B) and CD1d (C) with amino acids colored by biophysical property (see key).

In achieving this new functionality, the critical step lies in the transformation of the MHC sequences into a numeric form as in Figure 2.2A. To accomplish this, we split the sequences by their most prominent structural features. For MHC and MHC-like molecules, these features are the two β-strands and α-helices of the so-called platform domain. Each sequence within a given class is globally aligned, and one representative sequence from each class is sent through the Phyre2 structural prediction server [186]. We then use these structural predictions to identify the start and end points of each major structural feature in the alignment space. These start and end points then define the boundaries that are numerically

encoded into our position-sensitive matrix, as seen in Figure 4.4A.

Once the data are in this form, all downstream analysis outlined previously can be applied in a similar manner. In this example case, we find that average biophysical properties across these sequences reveal expected differences in hydropathy, with the lipid binding CD1 molecules displaying increased hydrophobicity when compared to the peptide binding MHC class I molecules (Figure 4.4B). Interestingly, unlike in the case of the antibody analysis, a simple PCA can effectively discriminate between the two training classes in this example case. Figure 4.4C shows the projection of the biophysical property data of each class onto the first two principal components.

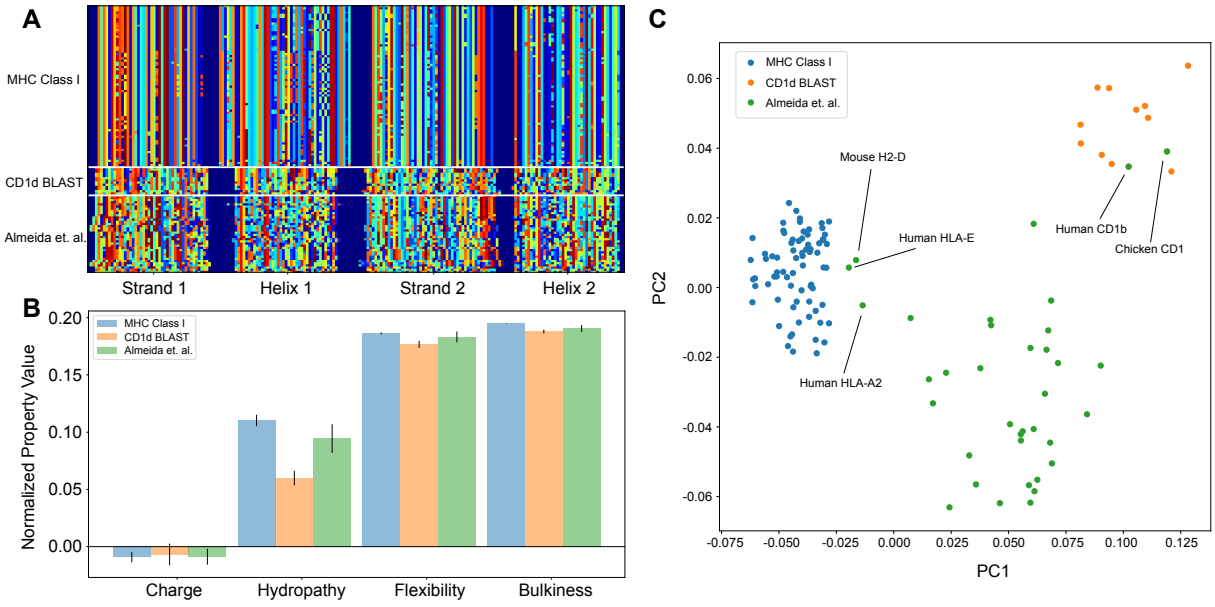


Figure 4.4: **The analysis pipeline is flexible and sufficiently identifies differences between MHC Class I and CD1-like molecules and has the potential to be used as a classifier moving forward.** (A) Similar to antibody sequences, the MHC and CD1 sequences can be encoded into a matrix. Here, we focus on human Class I MHC molecules and CD1 molecules from various organisms and use these sequences as training data to classify sequences collected in Almeida et al., 2020. (B) Comparisons of simple biophysical properties across these molecular species highlight differences between classes. (C) Projection of the biophysical properties of each class on to the first two principal components can be used to classify MHC- and CD1-like molecules. Molecules present in Almeida et al., 2020 but absent from the training data are labeled.

Here, we see that the peptide binding molecules (HLA-E, HLA-A2, H2-D) and the lipid binding molecules (Human CD1b, Chicken CD1) of the test dataset cluster with the respective peptide and lipid binding training data. The majority of the data of Almeida et al., comprised of non-classical MHC class I molecules from cartilaginous fish, clusters as its own distinct group, likely due to evolutionary distance between these molecules and those derived from mammalian and avian immune systems.

The software generated for this study is publicly available as a Python application (see Appendix). The unique aspect of this software is its hybrid approach to position-sensitive amino acid sequence analysis. Structural information is implicitly encoded by the alignment strategy employed, yet these assumptions are weaker than those imposed by explicit structural prediction. Downstream analysis from this positional encoder is streamlined and can be generalized to analyze any binary or higher order classification problems. This streamlined analysis allows for the generation of each figure in this study to be applied to thousands of sequences in a matter of minutes. The classification capabilities of the software could prove particularly useful when comparing binary classes, such as T cell receptors or antibody sequences derived from healthy and diseased tissue samples. Acceptable inputs are not restricted to CDR loops of immunoglobulins, and we have shown that the software can be adapted for the analysis of MHC-like molecules. Moving forward, this MHC analysis has the potential to classify the antigen binding properties of highly-divergent MHC sequences from a broad range of species, providing insights where phylogenetic approaches prove difficult. This software represents a strong addition to the existing toolkit for repertoire analysis of diverse molecular species.

4.3 A Refined Pipeline for Free Energy Calculations

While the development of entirely new analysis pipelines is import and gratifying work, in some ways the careful documentation and refinement of existing analytical tools can prove even more beneficial to science. For example, although the calculations necessary to

acquire the free energy of binding discussed in Chapter 3 provide succinct data that are readily interpretable, the practical implementation of these calculations are actually quite involved. Unlike the more frequently used "brute-force" simulations that are accessible to novice users, the best practices for free energy calculations are still in development. In this section, I will first provide a brief history on the development of these calculations, followed by a broad overview of the approach used to calculate the free energy of binding for protein-protein interactions. I will subsequently outline the practical workflow developed for these calculations, both as a record of my own contributions to the field, and as a guide to future potential users of this approach. This written section and the code provide on *github.com/ctboughter/string_method_automate* should allow for a more user-friendly implementation of free energy calculations moving forward.

4.3.1 An Introduction to Free Energy Calculations

Given the significant advances in computation outlined in Chapter 1, the scope and efficiency of free energy calculations are continually expanding and improving. Sampled free energy landscapes have been extended from the simple characterization of the rotomeric states adopted by dialanine dipeptide [135] to the elucidation of entire protein folding pathways and binding interactions [44, 187–191]. These improved capabilities are paving the way for applications unheard of a mere decade ago. Technological and methodological advancements in the field are such that current research is focused on the *de novo* design of drugs using computational approaches, with real hope these computationally designed therapies reach the clinic in the near future [192, 193].

Despite these modern advances, the fundamentals in the field are derived from developments in statistical physics dating back nearly 90 years. A majority of the modern approaches for calculating protein-protein binding free energies are rooted in the concept of the potential of mean force (PMF), first formulated by Kirkwood in 1935. In this work, Kirkwood generates a practical statistical approach for predicting the behavior of non-ideal fluids, with

a specific application towards the behavior of complex liquid solutions [194]. Ultimately, this formalism has been adopted for the calculation of binding free energies using molecular dynamics simulations through either a direct translation of this PMF-based approach [135] or through a slightly more nuanced procedure as is found in adaptive biasing force (ABF) methods [195].

The primary method for calculating the binding free energy in this work is based on the methods outlined by Woo and Roux, using a potential of mean force-based approach to obtain an absolute binding free energy between a protein and ligand [136]. Woo and Roux utilize an umbrella-sampling scheme to effectively pull a ligand out from the protein binding pocket. Using this formalism, the unbinding pathway is directly sampled throughout the course of the simulation, using the umbrella "windows" to direct the sampling and speed up the convergence of the calculation. This approach has a distinct advantage over so-called alchemical perturbative approaches whereby the ligand is gradually decoupled from the simulated system. While both approaches are theoretically rigorous and appear to have similar accuracies, the alchemical calculation of binding free energies is limited to small ligands that are not too strongly charged [196]. Importantly, this means that the PMF formulation can be used to calculate the binding free energies of protein-protein interactions.

Practically, little has changed in the protocol for using the PMF-based sampling approach in the 15-plus years since Woo and Roux published their methodology. The protein and the ligand, which can take the form of a small molecule or second protein, start out in their bound pose, often as determined by a crystal structure. Simple RMSD and center of mass restraints are placed on each molecule across the unbinding pathway of overlapping umbrella windows to conserve the bound protein conformations. However, these simple restraints are insufficient for calculating the binding free energy given current computational constraints. If the molecules are allowed to diffuse freely relative to each other, the energetic landscape along the unbinding pathway becomes significantly more complex, reducing the likelihood of convergence. To circumvent this problem, we constrain the two proteins relative

to each other. Woo and Roux generated these relative orientational constraints by selecting three positions within each molecule and calculating both the Euler and Polar angles using this internal coordinate system (Figure 4.5). Once the unbinding pathway is generated, each umbrella is simulated and unbiased using methods such as WHAM or MBAR [197, 198], until convergence is reached. Afterwards, any additional biases applied through various restraining potentials are accounted for in the final free energy calculation.

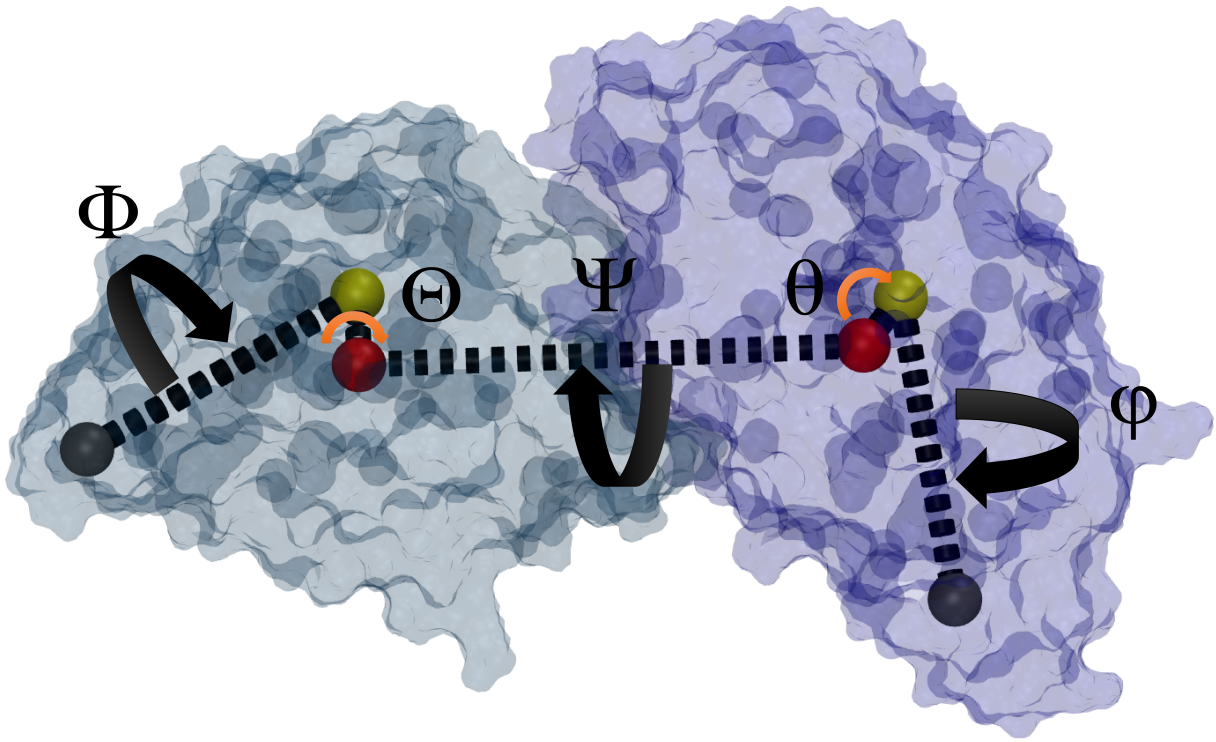


Figure 4.5: **A schematic depiction of the use of internal reference points for the calculation of Euler angles and polar angles necessary for the string method.** Highlighted structure is the B30.2 homodimeric interaction discussed in chapter 3 (PDB: 5HM7).

While the above outlined approach is still widely used for calculating binding free energies, there remains significant room for improvement in the protocol. First, the selection of arbitrary anchors within each molecule can lead to issues. Particularly, if certain relative orientations of the molecules are sampled throughout the course of the simulation, it can lead

to the so-called "Gimbal lock", collapsing a degree of freedom of the Euler angles, which can subsequently reduce the accuracy of the free energy calculation [199]. Further, the determination of the optimal unbinding pathway remains an open question in the field. When pulling the molecules apart from each other, a given straight path is frequently suboptimal. The identification of the "most probably transition pathway", which can be thought of essentially as the path of least resistance to unbinding, can help speed up convergence. However, the computational resources required to identify this pathway should not exceed the costs associated with slower convergence of the energetic calculations. Lastly, the many modules and applications required for running these calculations remain burdensome, so significant effort must be expended to streamline the process making it more use friendly. These outstanding issues are addressed, in order, in the following sections.

4.3.2 Troubleshooting Toolkits for Implementation

Once one has committed to the calculation of the binding free energy of a protein-protein or protein-ligand interaction, there are a multitude of preprocessing steps necessary before any sampling of the unbinding pathway can be undertaken. First and foremost, the identification of specific reference points within each molecule for the generation of the aforementioned internal coordinate system remains a nontrivial step that must be completed prior to the generation of inputs for the string method. These internal coordinates must be carefully chosen to avoid Gimbal lock, ideally selecting them in such a way that this degree of freedom collapse is avoided throughout the generation of the unbinding pathway, which can be difficult to predict *a priori*. This choice is further complicated by the manner in which NAMD's collective variables module (colvars) enforces the angular restraints throughout the course of the simulation. To maintain the relative orientations of each binding partner across all trajectories, corrective forces are applied throughout duration of the simulation. However, due to the instantaneous nature and nontrivial magnitude of these forces, if they are applied directly to singular atoms during dynamic steps they can deform atomic bonds, causing the simulation to crash. Thus practically we must select entire clusters of atoms spanning stable

structures such as parallel β-sheets or α-helices to define our internal coordinates.

In an attempt to overcome this issue, Fu et al. set out to automate the process of the definition of the restraints necessary to maintain the relative orientation of each binding partner using the quaternion [199]. The Euler angles that were previously defined by the arbitrarily selected positions can instead be defined using the four quaternions: the scalar q_0, and the three quaternion units q_1, q_2, and q_3. These quaternions are sufficient for defining the Euler angles, but require careful calculations throughout the duration of the simulation rather than the simple measurements taken from the arbitrary reference points. While a complete implementation for this module has been provided along with the published work of Fu et al., limited practical testing was published alongside this module.

Considering the additional rigor contributed by the use of the quaternion method for defining Euler angles, I worked to test and implement this method into the existing workflow for the string method, which will be discussed further in Section 4.3.3. The string constraining the unbinding path of the BTN3A1 B30.2 homodimer system in Chapter 3 was generated using both the quaternion and the internal user-defined reference points for the calculation of the Euler angles. In the initial stages of string generation, when each protein is still in close proximity, the two systems exhibit nearly identical behavior, with the unbinding pathways following similar routes through Euclidean space. However, as the two proteins begin to separate and more extreme deviations from the initial angles are sampled, the quaternion simulations abruptly crash. The data describing the first instance of this crash can be found in Figure 4.6.

We see from the instantaneous measurements of the collective variables of the proteins (Figure 4.6A) that all metrics start close to their harmonic restraints to begin the simulations. Quickly, some instability, seemingly starting in Euler Φ strongly pushes the protein away from the harmonic center. It is important to note, all of the Euler angles have some relative dependence on one another, i.e. this issue in Euler Φ may be responsible for the drift in

the measured Euler Θ and Ψ. We see from the RMSD trace of Protein A that the forces generated during this glitch are sufficient to deform the protein, providing confirmation that this error is a major concern moving forward. Interestingly, despite the harmonic center of Euler Φ being set to -20 Degrees, we see that it appears to reach a new, stable state at -100 Degrees. Strangely, this newly stable point increases the total energy of the system, suggesting this new configuration is relatively unfavorable (Figure 4.6B).

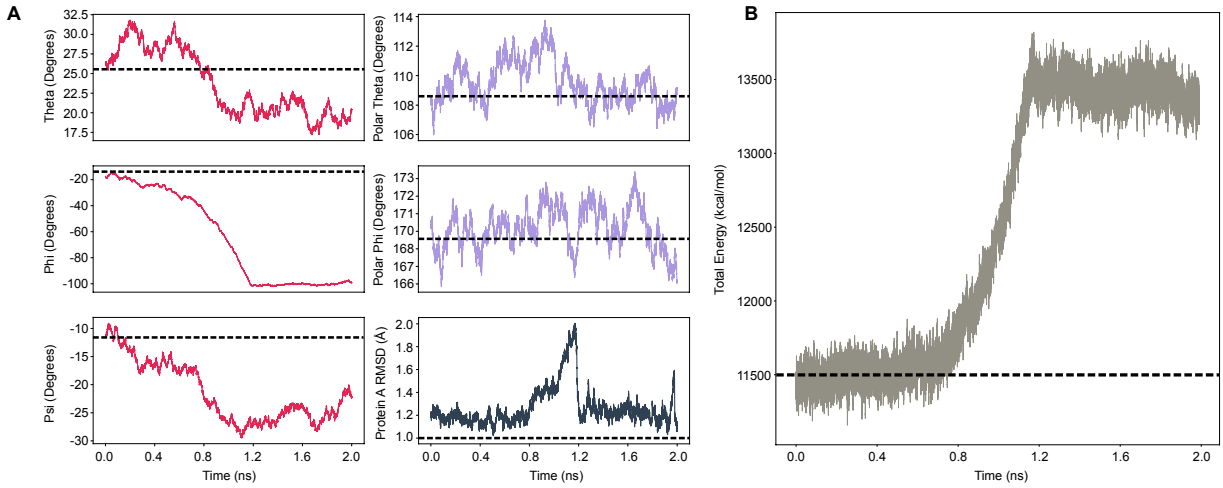


Figure 4.6: **The quaternion method for defining Euler angles introduces instabilities into simulated systems.** (A) Instantaneous measurements of the collective variables in a protein-protein interaction system highlights unexpected drift in the Euler Phi variable. (B) Measurement of the total energy of the simulations in (A) show spontaneous increase in the total energy of the system.

To identify the source of the problem, we generated a simple toy system comprised of two triads, i.e. eight total particles used to generate two trirectangular tetrahedrons with unit bond lengths. The tetrahedrons are generated using CHARMM and simulated in vacuum. These triads were rotated through all possible polar and Euler angles, as seen in Figure 4.7. These simple toy systems, as first noted by Benoît Roux, reduce the degrees of freedom in the system significantly. At the outset, it was not clear if this issue in the implementation of the quaternion was a result of protein-protein interactions or complications due to the extended conformations of each protein. These concerns were quickly dismissed thanks to the toy model.

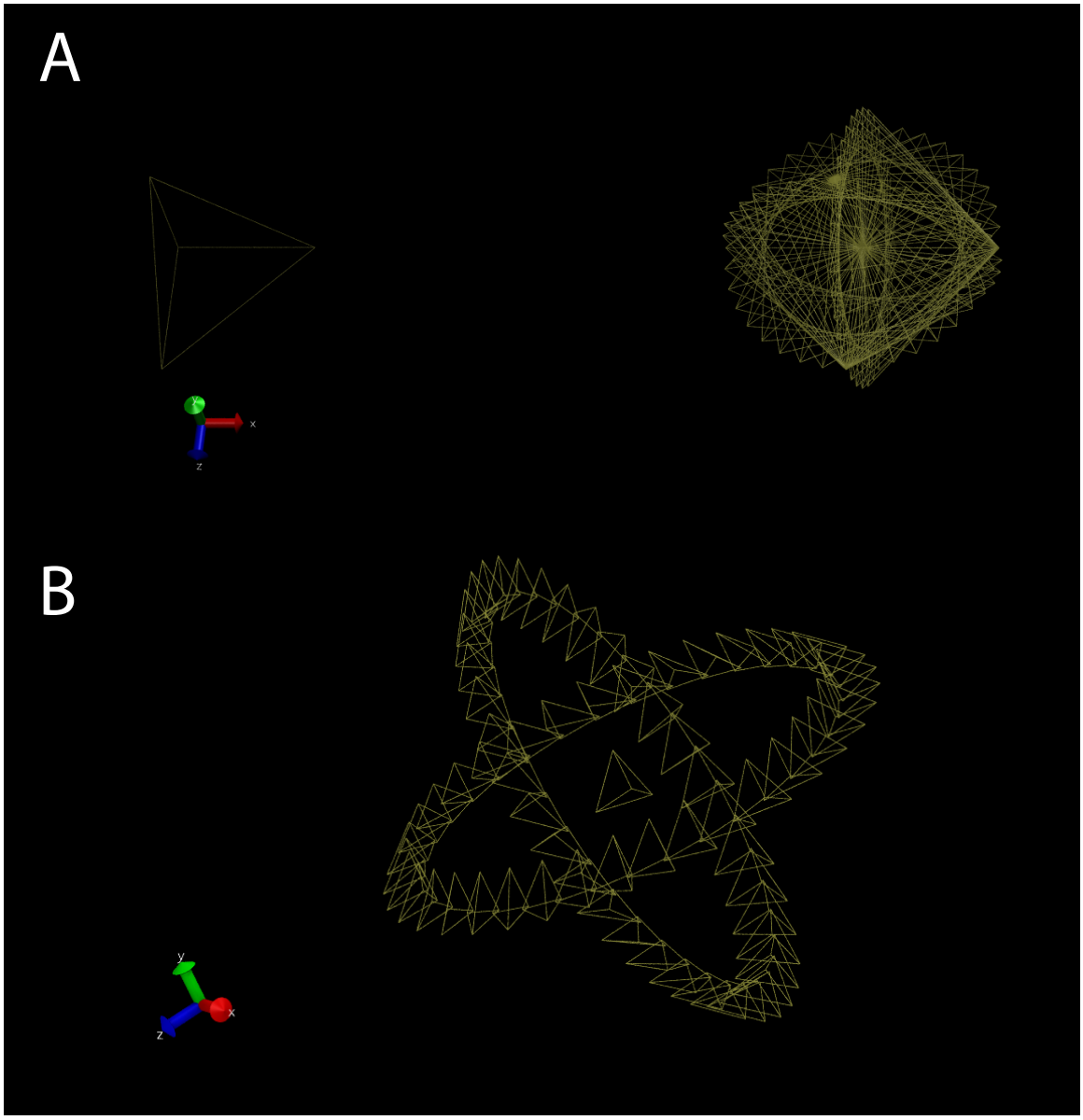


Figure 4.7: **Visualization of the toy model system for diagnosis of issues with the quaternion implementation.** (A) Rotations of the toy triad through Euler angles. (B) Rotations of the toy triad through polar angles.

By positioning the toy system in precisely the same conformation as the protein system, we could quickly recreate the same issues with the quaternion, suggesting the existence of some inherent issue within the module. We next set out to identify precisely what in the implementation of the quaternion module produces these errors. Are the energies increasing due to some issue with all of the collective variable restraints interfering with each other in

some kind of positive feedback loop, or are issues with the individual restraints to blame?

Using this toy model, we further dissect the problem by turning on each restraint one by one and assaying the relative stability of each toy system at the same angular positions and the same harmonic restraints as the protein system. We find that the polar angle constraints behave as expected (Figure 4.8A), diffusing freely in whichever dimension is not constrained while closely adhering to the harmonic center when the restraint is turned on. Likewise, the total energy in these simulations is exquisitely conserved, gradually moving towards a very slightly lower energy state, as we might expect (Figure 4.8B). Conversely, the Euler angles all appear to have significant issues around these angular positions (Figure 4.8C). Again, only one harmonic restraint is applied to the system at a given time.

Of the three Euler angles, the Euler Θ restraint appears to be the most well behaved, yet significant deviations from the harmonic center are observed throughout the duration of the simulation. The Euler Φ and Ψ restraints each induce significant deviations from expected behavior, forcing each toy particle to wildly rotate through space when these restraints are turned on. Consistent with these angular traces, we see that the Euler Φ and Ψ restraints cause the total energy of the system to grow uncontrollably and unpredictably (Figure 4.8D) and lead to the deformation of the toy particles (Figure 4.8E).

Far from the additional restraints introducing issues with the implementation of the quaternion, it appears when all restraints are turned on they in fact prevent the Euler angle constraints from plunging the system into complete disarray. We find that the Euler Φ and Ψ angles are the most volatile, perhaps due to the use of the arctangent function for the determination of these Euler angles. However, it should be noted that Euler Θ, which is calculated from the quaternion using the arcsine function, is unstable but less violently so. Importantly, issues arise not only where we should expect singularities to exist in these functions.

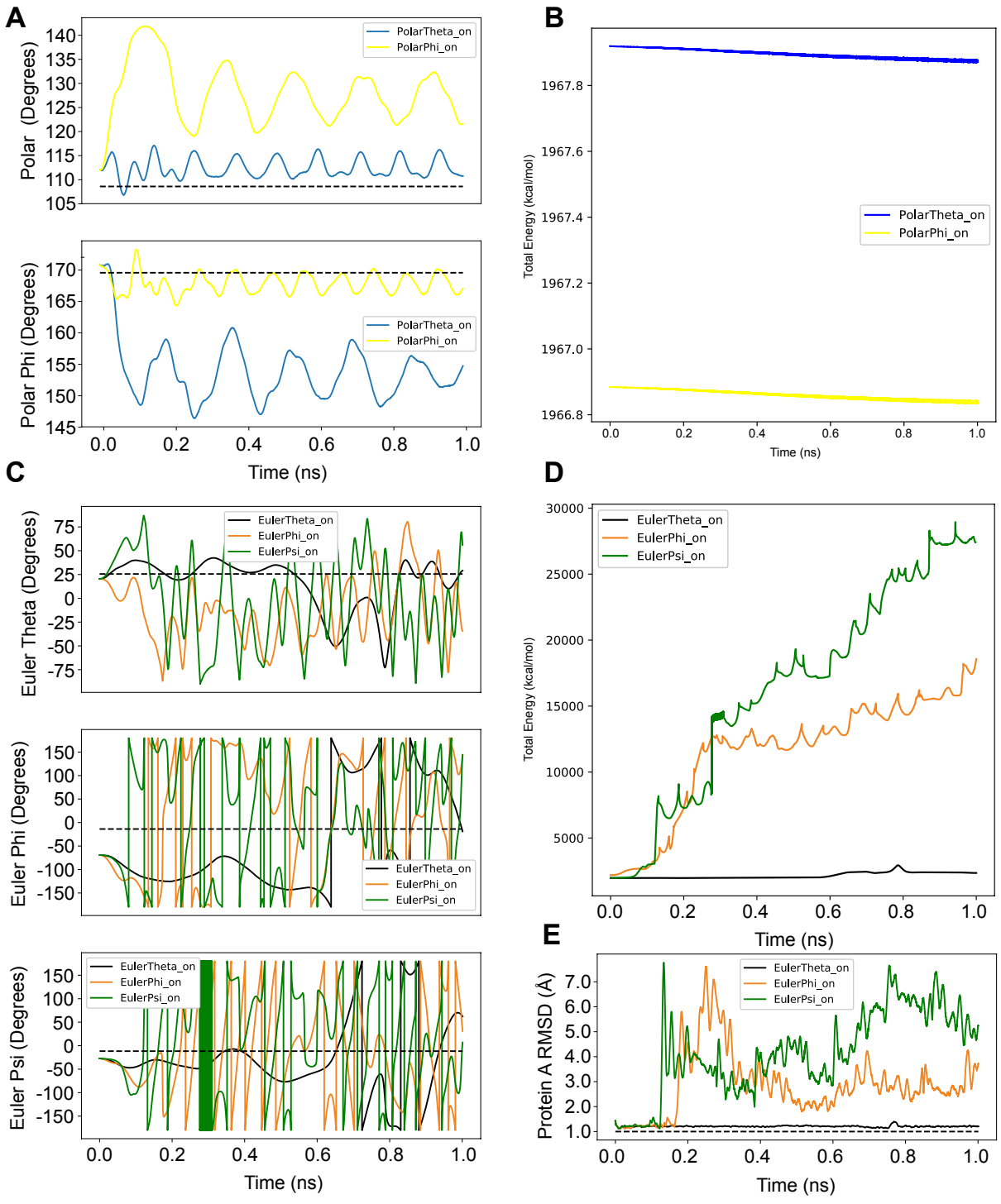


Figure 4.8: **The toy model conclusively highlights the violent instabilities of the quaternion calculation of Euler angles.** (A) Collective variable traces of the polar angles. (B) Total energy plot of tested polar angle restraints. (C) Collective variable traces of systems with individually applied Euler angle restraints. (D) Total energy of the systems in (C). (E) RMSD measurements of the toy particle as simulated in (C).

When testing the toy model in all orientations and all harmonic restraint locations, we find that despite starting the simulations with the toy particles in the exact center of the harmonic well, the system becomes rapidly unstable across a wide range of tested angles, as measured by the total energy (Figure 4.9). These instabilities persist not only at singularities that we might expect, but at all positions between these possible singularities. This suggests some kind of fundamental issue with the code, and current efforts are underway to solve this issue. However, without these strenuous tests on new implementations such as this, bugs would alter the results of carefully crafted free energy calculations for years to come.

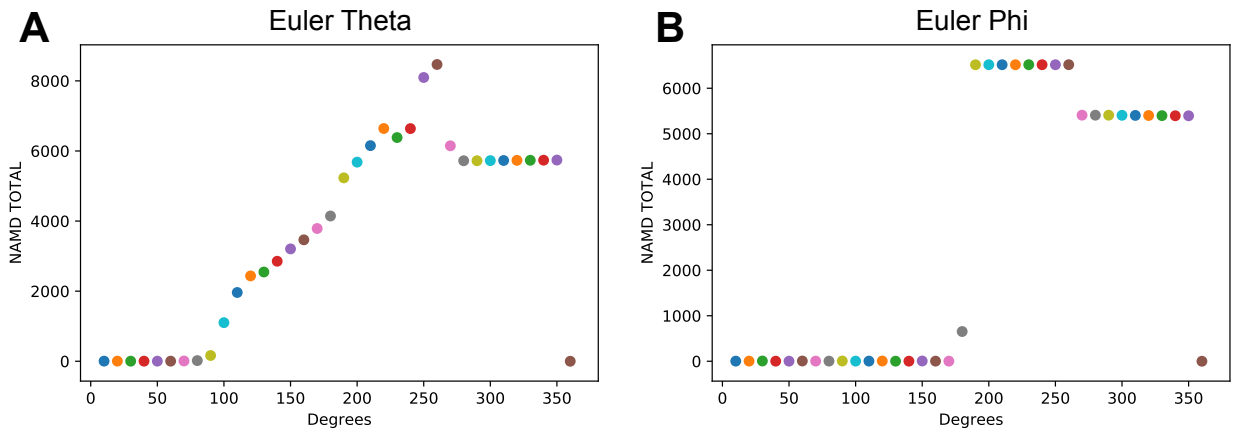


Figure 4.9: **Issues with the quaternion arise independent of expected singularities in the arctangent and arcsine functions.** (A) Total energy of the toy model system when testing all possible positions and restraints in Euler Theta. (B) Total energy of the toy model system when testing all possible positions and restraints in Euler Phi.

4.3.3 A Streamlined Method for Identifying the Most Probable Transition Pathway

Upon our determination that the quaternion was not yet ready for direct applications to the string method, we continued to press forward with our binding free energy calculations using the existing method for calculating Euler angles. However, I continued to work to improve the string method in other ways. Specifically, I have made improvements to the way

that the initial string is generated, and in the optimal placement of replica exchange windows across the energy landscape spanned by the string. In highlighting these improvements, I will discuss all steps proceeding the selection of anchor points within each protein for the calculation of Euler angles, using the specific example of the BTN3A1 B30.2 homodimer system discussed in Chapter 3. Novel contributions to this workflow will be explicitly referred to as such.

Starting from the bound crystal structure of the B30.2 domain (PDB: 5HM7) [116], we initialize a simulation not in a fully hydrated periodic box, but instead in implicit solvent. Practically, we generate the string by slowly pulling the protein away from the bound pose, allowing this newly displaced protein to equilibrate somewhere near a local energy minimum until each monomer is fully separated. At each incremental step away from the bound pose, we allow the system to relax while maintaining restraints on the separation between each monomer. However, in our generation of the string, we are updating six of the collective variables in the system, the five orientational restraints and the single center to center distance restraint. The evolution of this latter restraint is straightforward, as the Euclidean distance between each window can be arbitrarily chosen so long as the energy difference between each window is constrained. A protocol for guaranteeing this energetic constraint is satisfied is discussed in Section 4.3.4.

While the determination of the physical separation of each window along the string pathway is nearly trivial, the careful tuning of the orientational restraints along the string is critical for separating the two complexes along the most probable transition pathway. The two molecules can separate across a three-dimensional space swept out by a cone emanating from the bound pose, thus we adopt a systematic approach to identify the lowest energy path through this cone. Previously, an algorithm referred to as the dynamic histogram analysis method (DHAM) has been used to generate new windows based on the energetic landscape sampled in the previous, nearby windows [191, 200]. In this way the string can be iteratively generated based on the local energetic landscape. However, this approach is computationally

inefficient, requiring roughly 8 hours of computing time to generate 10 new windows on a single standard computing node. We reason that similar performance can be achieved using a simple averaging method, whereby the local energy minimum at the next window location can be estimated by a simple averaging scheme as in equation 4.4:

$$\xi_{cen}^{n+1} = (1 - S) * \xi_{avg}^{n} + S * \xi_{cen}^{n} \tag{4.4}$$

Where ξ_{cen}^{n} gives the center of the harmonic potential applied at window n, and ξ_{avg} is the real averaged value of the collective variable throughout the course of the simulation. In this way, we allow the string to evolve across the energy landscape while restricting each window to regions of the conformational landscape near the previous window. We evolve the parameter S based upon the minimum distance between each protein to slowly halt updates in the evolution of the restraints when the dimer completely dissociates:

$$\begin{cases} S = 1 - \frac{(r_{min} - a)^2}{a^2} & r_{min} \leq a \\ S = 1 & r_{min} > a \end{cases} \tag{4.5}$$

The variable a is then a tunable degree of freedom that allows us to determine at what minimum distance we stop updating the orientational restraints applied to the protein and begin pulling the two proteins along a straight pathway. Figure 4.10 shows comparisons between DHAM and the averaging algorithm of Equations 4.4 and 4.5 across three distinct conditions; $a = 12$, $a = 5$, or $S = 0.5$, also referred to as "no updates". In Figure 4.10, we simplify the nomenclature of the orientational restraints by referring to Euler Θ, Φ, and Ψ as Alpha, Beta, and Gamma. We see that while Alpha, Beta, and Polar Phi restraints all display some qualitative agreements in their individually evolved paths, the Gamma and Polar Theta restraints diverge quickly. Despite this divergence, visual inspection of the proteins through these paths show good agreement between DHAM and the averaging

algorithm with $a = 12$ (data not shown).

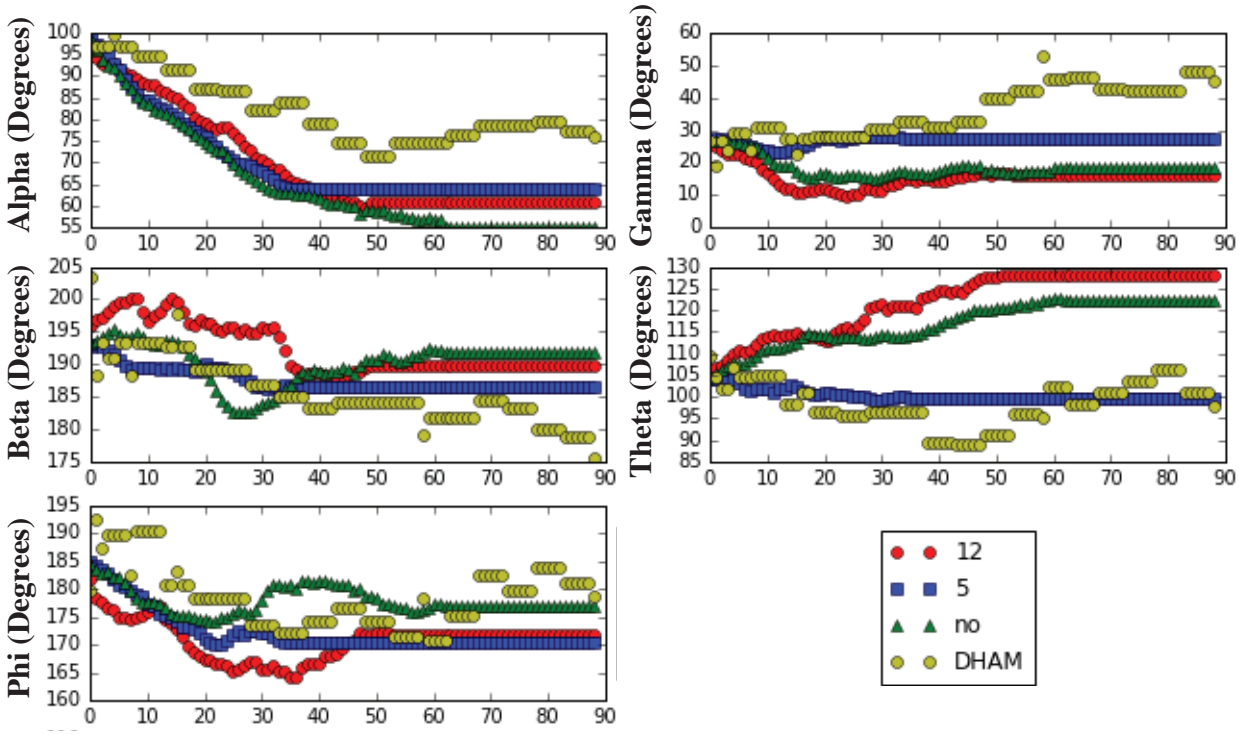


Figure 4.10: **A simple, more computationally efficient averaging procedure for the generation of novel unbinding pathways for the string method performs similarly to the more robust DHAM-based approach.** The determined harmonic centers of the most probable transition pathway generated using DHAM or three distinct averaging methods.

With this novel update to the protocol for generating a curvilinear pathway for string method calculations, we are able to substantially improve the computational efficiency of this step. Rather than the roughly 48 hours required to generate 60 windows using the DHAM approach, the adaptive averaging method I developed needs only 30 minutes for all 60 windows, with the slowest step being the determination of the minimum distance r_{min} between the two proteins for the calculation of S. While this represents a substantial improvement in the determination of the most probable transition pathway, both the DHAM-based approach and the averaging method outlined here require multiple equilibration simulations to sufficiently sample the local energy landscape. For more complicated energy landscapes, these initial simulations in implicit solvent represent a significant investment of computational resources, although these resources pale in comparison to those necessary for the final

replica exchange simulations.

Once the initial string is fully generated in implicit solvent, we then solvate the system following the standard CHARMM-GUI protocol [123, 124]. Each newly generated string window is its own independent molecular system, and is treated as such throughout the hydration process to add explicit solvent. However, upon initiation of replica exchange simulations, each window must have an identical number of atoms and atom types for proper function of the exchange steps. As such, we must carefully control the solvation steps for each window. This step is likewise nontrivial, as there exists stochasticity in the deletion of waters clashing with protein atoms during the solvation step of the standard CHARMM protocol when using the online GUI. This stochasticity can lead to differing numbers of water atoms in each window along the string. To circumvent this issue, I have generated custom scripts provided on the accompanying GitHub page that allow for the direct hydration of each window on a local machine, guaranteeing uniformity in the number of atoms across all windows.

Once all windows are hydrated, we allow the string to equilibrate in the explicit water environment, as we expect the energetic landscape should change throughout the transition from implicit to explicit solvent. Additionally, due to the periodicity of the simulation box, we apply a restraint to prevent each monomer from crossing over the periodic boundaries, which can complicate the final free energy calculation. These restraints are defined with respect to the center of mass of each protein and both the XY and YZ planes. Effectively, this creates a cylindrical region in which the proteins are restrained (Figure 4.11). Once the string appears relatively stable in the energetic landscape of the hydrated system, the replica exchange umbrella sampling steps can begin.

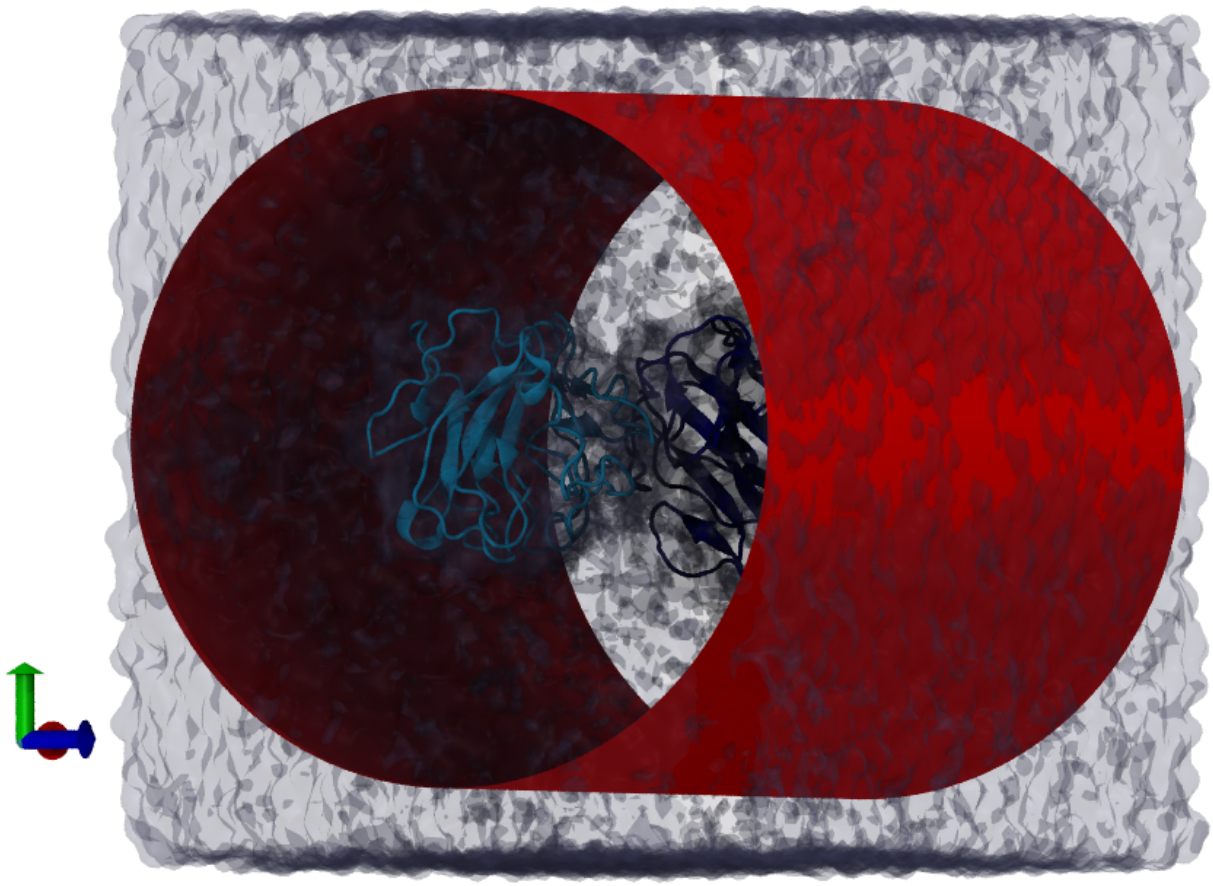


Figure 4.11: **Visualization of the angular restraints applied to proteins throughout the string method calculation to prevent crossing of the periodic boundary.** Proteins from the simulated string in Chapter 3, the B30.2 domains, are fully hydrated in a rectangular box. To prevent crossing the periodic boundary, we apply restraints that prevent crossing the cylindrical plane visualized here in red.

4.3.4 A Novel Approach for Optimizing Replica Exchange Performance

Importantly, before immediately initiating the production runs of the replica exchange molecular dynamics umbrella sampling (REMD-US) simulations, we must guarantee that the windows are evenly spaced across the energetic landscape of the unbinding pathway. Replica exchange molecular dynamics simulations are meant to improve sampling of MD simulations. When combined with umbrella sampling, as in the REMD-US approach used for the string method calculations, one can speed up convergence of calculations by sampling

over the barriers between each window across the reaction coordinate of interest. However, if the energy difference between each window is too large, exchange will not occur, and the unbinding pathway will not be fully sampled, likely introducing errors into the final free energy calculation.

To circumvent this issue, I set out to create a simple yet efficient approach for guaranteeing that windows are properly spaced throughout the unbinding pathway. Standard replica exchange algorithms currently used in the field operate on a simple Metropolis–Hastings criterion to make decisions regarding whether a potential exchange move should be accepted. As such, the exchange probability P for REMD-US simulations is generally defined as:

$$P = \begin{cases} 1 & \text{if } \delta \leq 0 \\ e^{-\beta\delta} & \text{if } \delta > 0 \end{cases} \tag{4.6}$$

If δ is negative, i.e. an energetically favorable exchange, the move will occur with 100% probability. However, should the move be energetically unfavorable, with positive δ, then the probability of exchange decays exponentially with the difference in energy. We calculate this δ for all adjacent windows i and j using equation 4.7:

$$\delta = (\Delta E_i + \Delta E_j) \tag{4.7}$$

Generally, we consider an exchange rate of 20% across neighboring windows as ideal for proper sampling. We can use the above equations 4.6 and 4.7 to back calculate and determine the average energy difference between two neighboring windows required to attain this 20% exchange rate:

$$e^{-\beta\bar{\delta}} = 0.2 \tag{4.8}$$

$$\bar{\delta} \approx \frac{1.61}{\beta} \tag{4.9}$$

$$\bar{\delta} \approx 0.959 \ ^{kcal/mol} \quad at \quad 300K \tag{4.10}$$

In other words, to guarantee exchange at or near 20% acceptance across the length of the simulation, there should be an energy difference between windows of about 1 kcal/mol. We can see the dramatic changes introduced when applying this methodology to initial short replica exchange runs. By visualizing the current restraint applied to each window, whereby window 1 is initialized with restraint 1, window 2 is initialized with restraint 2, etc., we can track how each individual restraint is swapped throughout the replica exchange trajectory. Figure 4.12A visualizes these exchanges before re-optimization enforcing the 1 kcal/mol window distance rule, whereas Figure 4.12B displays the drastic improvement in exchange rate after the addition of new windows to decrease the energy difference between windows to less than 1 kcal/mol.

Critically, no further simulations are required to identify these gaps between windows. To achieve this significant improvement in exchange, we use the output collective variable data from the final equilibration run of the hydrated string. We can then calculate the energy difference across the windows using the multi-state Bennett acceptance ratio (MBAR) algorithm [198]. Figure 4.12C shows these energy differences, both before and after the interpolation of windows between large energy gaps across the unbinding pathway. Interpolation of windows results in near ideal sampling across the entire string (Figure 4.12D). The final step in this entire process is to then run the REMD-US simulations, as in the original formulation of replica exchange by Sugita et al. [201], and calculate the resulting free energies as in Gumbart et al. and Suh et al. [191, 202].

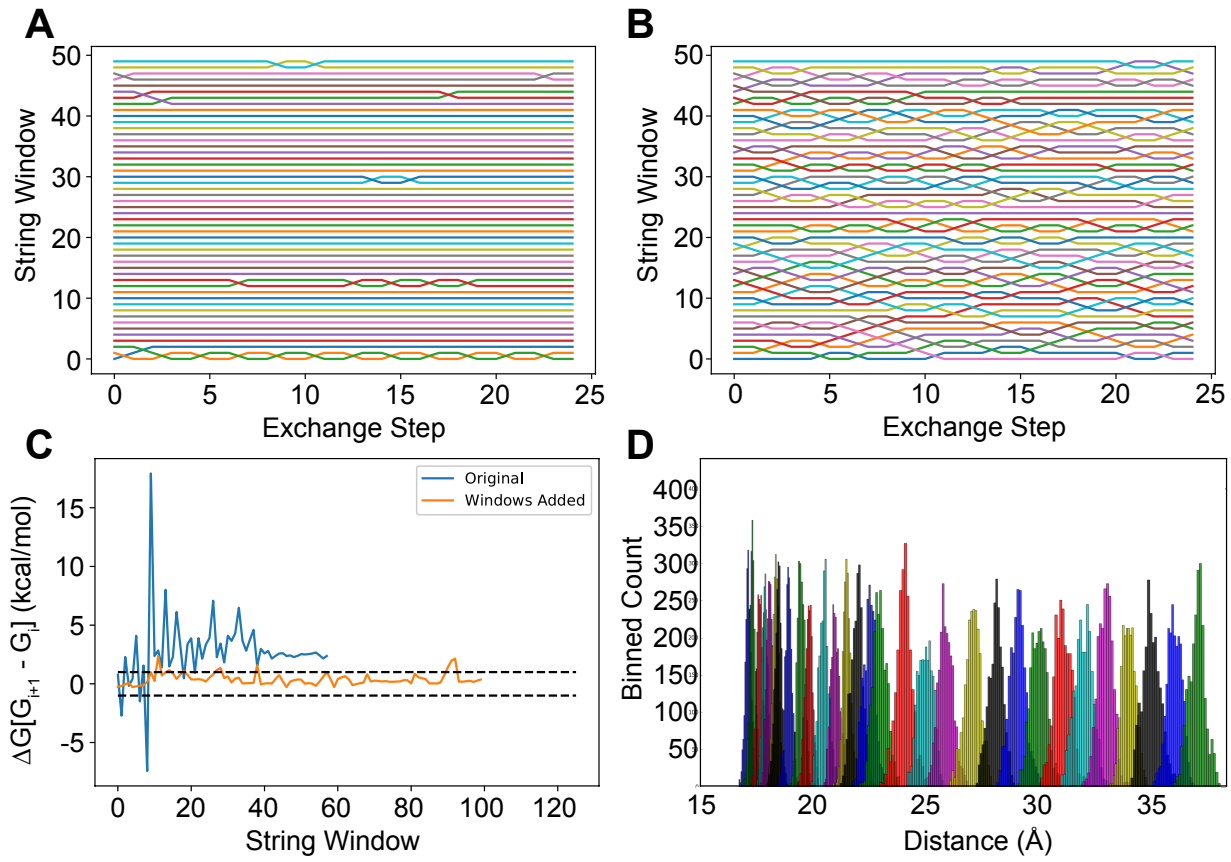


Figure 4.12: **A novel algorithm for the identification of gaps in replica exchange simulations shows significant improvement in exchange percentage.** (A) Exchange before application of algorithm. (B) Exchange after application of algorithm. (C) Energy difference between windows of for the systems highlighted in (A) and (B). (D) Histogram representation of improved exchange system shows exceptional sampling across the region of interest.

4.4 Discussion

The computational tools developed throughout the duration of this dissertation have all contributed to the generation of basic knowledge of the fundamental biological mechanisms of adaptive immune responses, but also represent substantial contributions towards a movement of more open and reproducible science. All of the code discussed in this chapter is publicly available online, with accompanying instruction for novice users and, in the case of AIMS,

a graphical user interface for those completely unfamiliar with computer programming. As we progress forward in science, especially as computation is becoming increasingly prevalent in all fields of science, there must be a consistent effort to develop and maintain software that is freely shared, so the focus of the field can remain results driven, without time wasted across multiple groups struggling with identical issues.

With the development of AIMS, we are providing not only a powerful platform for the analysis of immune molecules, but a novel way of thinking about immune interactions and machine learning in biological applications. Rather than considering antibodies and T cell receptors as complex products of processes like V(D)J recombination and somatic hypermutation, we can boil these molecules down to their simplest components to find patterns that describe their behavior. Frequently, the aspect of the antibody or T cell receptor's behavior we are most interested in is their ability to bind specific molecular targets. Barring some metaphysical memory of its evolutionary path, this behavior should be entirely encoded in the amino acid sequence.

We use this fact as motivation to strip away all excess information in this problem, focusing only on the fundamental biophysical properties of the protein-ligand interaction interface. Artificially imposed restrictions are likewise discarded in the AIMS analysis pipeline, particularly the common approach in the field of segregating sequenced complementarity determining regions or MHC regions by length, thereby simplifying the issue of alignment. Frequently, the edges of regions identified as CDR loops by software such as IgBLAST [84] or IMGT's V-Quest [170] are buried deep in the framework regions of the antibody or T cell receptor, thus contributing little to the recognition of binding partners. Similarly, the residues at the ends of the α-helices and β-strands of MHC molecules are rarely contacted by the TCR. We thus reason that the majority of critical regions of interest across the surface of immune molecules occurs near the center of each structure.

With this rationale, we can analyze CDR and MHC regions of unequal lengths together,

removing the aforementioned artificial restriction. We align all sequences by their center, under the assumption that this region is the most likely to be solvent exposed, and the most likely to interact with binding partners. We expect errors introduced by this approach to be averaged out with large enough datasets. Importantly, we accomplish this while maintaining a decreased emphasis on structural information when crystal structures are unavailable. Computational prediction of loop conformation is difficult, and drawing inferences from incorrect models regarding side-chain interactions and positioning could be misleading. In the high-throughput analysis of immune molecule sequences, our approach strikes a careful balance of the structural assumptions that should apply consistently across antibody and TCR populations.

Furthermore, the use of interpretable machine learning algorithms and information theory in the AIMS analysis pipeline may promote inquiry into the application of these approaches in immunology. Machine learning algorithms are frequently judged by their performance, with high performing algorithms judged primarily by their percent accuracy. However, an ability to break down these algorithms and understand the components responsible for their success, as is possible with linear discriminant analysis, further strengthens these machine learning approaches. Unlike machine learning, the application of information theory to immunology is still in its infancy. The utilization of concepts from information theory outlined herein represent only a glimpse at what may be possible as this niche space is further explored.

Whereas AIMS represents an entirely new analytical approach, the contributions to the string method for the calculation of binding free energies discussed above involve more modest contributions that are nonetheless critical to the advancement of the field. From the simple scripts aimed at automating specific complicated steps throughout the generation of inputs to the more advanced alterations in the implementation of the final simulated system, each new module is aimed at improved usability. As computational resources continue to become more readily available, more users may be inclined to attempt complicated calculations

of binding free energies using approaches such as the string method. We can prepare for this influx of users by making gradual improvements to the ways that we distribute software and communicate results. This Chapter serves as a testament to the importance that should be placed upon achieving these goals.

CHAPTER 5

CONCLUSIONS & PERSPECTIVES

5.1 Expanding the Definition of Canonical Immune Recognition

In our quest to broaden our understanding of the adaptive immune system, we are continually finding that what we hope to identify as hard and fast rules frequently have critical exceptions. These exceptions force us to step back and reconsider the way we think about our models and the roles of individual immune cells. Unlike in physics, there are no fundamental "first principles" in immunology from which we can build our functional theories. Instead, we must take a top-down approach, expanding our line of inquiry towards the periphery of our understanding. This periphery is precisely where my dissertation is focused, pushing our boundaries of understanding of the adaptive immune system.

As we continue to push these boundaries, we can redefine what is thought of as the canonical representation of the immune system. At present, some believe that properly functioning antibodies should bind to their target, and only their target, with exceptional affinity. Yet the prevalence and persistence of polyreactivity across all stages of antibody maturation seems to suggest a broader role for these promiscuous binders. Similarly, we classically expect that in a natural T cell response, T cell receptors should bind directly to an MHC or MHC-like molecule presenting a stimulating antigen in order to achieve T cell activation. While some may currently believe the activation of Vγ9Vδ2 T cells is an outlier case, the possibility remains that the elucidation of the precise mechanism of activation shares strong similarities with other existing activation pathways in non-immune contexts.

The data presented herein help build upon the existing literature in the field to understand non-canonical forms of immune recognition, providing a more robust base for future researchers. Antibody polyreactivity and Vγ9Vδ2 T cell activation each have lines of inquiry tracing back decades, with many of the most critical discoveries occurring in the past few

years. Each field has progressed rapidly, with much excitement currently surrounding newly published results. The opportunity to report my own findings and hypotheses in this space and contribute in a meaningful way has been incredibly gratifying, yet major questions still remain in these fields. In this chapter, I will provide my broader perspectives in each space, and speculate on the future directions that could drastically advance each field.

5.2 On the Biological and Biophysical Implications of Polyreactivity

While the results of Chapter 2 convincingly identify the fundamental properties responsible for making an antibody polyreactive, larger questions concerning antibody polyreactivity as a whole persist. The dataset tested is expansive, yet not comprehensive, leaving open the possibility for different pathways towards polyreactivity. Likewise, the coverage of antigen space is low, suggesting we may be selecting for a very precise form of polyreactivity. While recent effort has been extended to confirm that polyreactive antibodies also bind to additional ligands such as lysozyme and ubiquitin [91], higher throughput biochemical assays will be necessary to truly answer these questions. These shortcomings should be readily addressed as the field continues to expand, however larger questions remain in both the protein biophysics and B cell biology of polyreactive antibodies.

At the root of antibody polyreactivity lies a fundamental question: what qualifies as a true "interaction" between biomolecules? Developers of small molecule drugs might be inclined to ignore any interaction weaker than a femto- to picomolar affinity [203]. Those studying typical protein-protein interactions are concerned primarily with affinities on the order of a few hundred nanomolar or stronger [204]. T cell immunologists are primarily concerned with occasionally difficult to measure micromolar affinities of TCR-pMHC interactions [205]. At what point can we draw a hard cutoff and determine that two molecules are "non-interacting". Some may argue that the high micromolar to low millimolar affinities of polyreactive antibodies [91] are merely a sign of an unstable antibody, yet the ability of

polyreactivity to survive the affinity maturation process suggests there may be cases where a selection pressure acts to select for polyreactivity.

B cells are exceptional models for evolution, undergoing multiple rounds of genetic shuffling, mutation, and selection in a few short days. To understand the role of this selection process on the polyreactivity status of a given antibody, we must improve our overall understanding of the germinal center where this rapid evolution occurs. The germinal center remains something of a black box [8], with the specifics of how each antibody "sees" its cognate ligand unclear. Nuanced differences in this antigen exposure could hold the key for understanding whether polyreactivity is directly selected for, or if this feature merely survives selection. As we expand our understanding of the biological role of antibody polyreactivity, we can begin to more definitively answer whether antibodies interacting with polyreactive ligands represents a "true" interaction.

Should polyreactivity play a nearly insignificant role in the biology of antibody binding, a possibility that is becoming less likely given recent findings [47, 67], understanding how polyreactivity arises in antibodies still represents a massive contribution to the generation of novel therapeutic antibodies. While localized patches of positive and negative charge may directly match a given cognate ligand, maximizing the enthalpic interaction, our results suggest that care must be taken to avoid potential inter-loop interactions of these oppositely charged residues. The resulting rigid, charge-sequestered binding surface represents a blank canvas for weak, nonspecific interactions with a wide range of ligands. Broadly, such a binding interface may generally be key to promiscuity in biophysical interactions.

5.3 Towards Identifying the Mechanism Behind Vγ9Vδ2 T Cell Activation

Chapter 3 concerns the systematic breakdown and careful rebuilding of models of the butyrophilin-mediated activation of Vγ9Vδ2 T cells. Throughout the course of my disser-

tation research, I have carefully deconstructed a prevalent model in the field and provided a novel explanation for Vγ9Vδ2 T cell activation, whereby intracellular dimerization of BTN3A1's B30.2 domain cluster and immobilize BTN3A1 on the cell surface. However, this model lacks the capability to explain the role of BTN2A1 or the other BTN3A family members in this activation process. This latter shortcoming appears less critical to address immediately, as BTN3A1 appears to be capable of activating T cells independent of the presence of BTN3A2 and BTN3A3 [132].

Understanding how BTN3A1 and BTN2A1 interact, if they interact at all, is key in unravelling the mystery of T cell activation. Both Rigau et al. and Karunakaran et al. suggest that BTN3A1 and BTN2A1 directly associate on the surface of cells, but the evidence is indirect at best [132, 142]. Do BTN2A1 and BTN3A1 potentially heterodimerize? If so, does this interaction occur via the intracellular or extracellular domains? While BTN2A1 is unable to bind pAg [132], could the stabilization of BTN3A1's B30.2 domain by pAg increase the binding free energy of a heterodimeric 3A1-2A1 interaction in the same way it does for the homodimeric 3A1-3A1 interaction? Such an interaction could allow BTN3A1 to act as a chaperone for BTN2A1, either increasing the effective local concentration of BTN2A1 for direct interactions with T cell receptors or trafficking it to the immune synapse.

While such a proposed mechanism is highly speculative, we can more concretely say that it appears both BTN3A1 and BTN2A1 interact with molecules on the surface of the T cell [139]. While BTN2A1 is known to interact with the T cell, the identity of the proposed BTN3A1 interaction partner remains a mystery. It could be possible that pAg acts to immobilize BTN3A1 on the cell surface, forming the initial synapse. From there BTN2A1, and potentially another, heretofore unidentified additional "factor Z" engage the T cell receptor to confer the final activating signal. In this way, BTN3A1 would act as a sort of co-receptor with a unique role in synapse initialization.

Overall, the most critical outstanding questions in solving the mystery of Vγ9Vδ2 T

cell activation appear to be dynamic in nature. Understanding the order of the cascade of events initiated by intracellular pAg binding and concluding in immune synapse formation and T cell stimulation has the potential to lock all of these identified pieces into a coherent model for activation. However, due to the sensitivity of this system to perturbations such as overexpression or cytosolic tags [206], clever experimental designs will be necessary to probe these dynamic phenomena. Moving forward, I have the utmost confidence in the incredible researchers in this space to crack this persistent problem.

5.4 Perspectives on a Cross Disciplinary Interrogation of Adaptive Immunology

Throughout the entirety of this dissertation research, careful documentation and publication of all relevant code and analysis has been provided to maximize the reproducibility of this work. For biology and the biophysical sciences to advance, we must strive towards maximal transparency in our methods as well as our complete reporting of our thought processes throughout the entire progression of discovery. This is most critical in the space of interdisciplinary research, where the boundaries between each discipline blur and the background expertise of the audience for subsequent publications is ill-defined.

It is important to note that the role of the researcher is not just to blaze new trails in one's field and dazzle peers with sheer brilliance, but also to carefully teach both experts and non-experts alike the nuances and implications of this exciting research. Indeed, these teaching moments are of incredible value to science, both for the enlightenment of the mentee and the external perspective provided to the mentor. Science, ultimately, is an endeavor towards improving not just our understanding of the machinations of natural phenomena, but also towards improving the education and understanding of those around us.

"[This] progress in learning about the world of nature has changed rather profoundly not only what we know of nature, but some of the things that we know about ourselves as knowers" – J. Robert Oppenheimer, *The Flying Trapeze* [207]

APPENDIX - MATERIALS & METHODS

Computational Methods

AIMS Software Development

All analysis was performed in Python, with code tested and finalized using Jupyter Notebooks [208]. Figures were generated with MatPlotLib [209] or seaborn [210], while the majority of data analysis was carried out using Pandas [211], SciPy [212], and SciKit-learn [213]. All code and data is available at *https : //github.com/ctboughter/AIMS*, including the original Jupyter Notebooks used to generate the data in this manuscript as well as generalized Notebooks and a Python-based GUI application for analysis of novel datasets.

Statistical Tests in Polyreactivity Analysis

Error bars in all plots are provided by the standard deviation of 1000 bootstrap iterations. Statistical significance is calculated using either a two-sided nonparametric Studentized bootstrap or a two-sided nonparametric permutation test as outlined in "Bootstrap Methods and Their Application" [214]. For the Studentized bootstrap, the bootstrapped data are drawn from a resampling of the null distribution of the data, with replacement. Practically, this entails combining the polyreactive and non-polyreactive antibodies into a single matrix, without labels, and using the Scikit-learn resample module to randomly separate this matrix into two classes, preserving the number of sequences in each population. To calculate bootstrapped averages, we draw from the empirical rather than null distribution. Statistical significance is estimated by calculating the p-value using the relation:

$$p = \frac{1 + \sharp(z^2 \geq z_0^2)}{R + 1} \tag{5.1}$$

Here, we calculate the p-value by counting the number of bootstrap iterations where z^2 is greater than or equal to z_0^2. z^2 and z_0^2 are Studentized test statistics taken from the null and empirical and distributions, respectively. R is the number of times this bootstrapping process is repeated. The general form of z is given by:

$$z = \frac{\bar{Y}_2 - \bar{Y}_1}{(\frac{\sigma_2^2}{n_2} - \frac{\sigma_1^1}{n_1})^{1/2}} \tag{5.2}$$

Where $\bar{Y}$ represents the bootstrapped sample mean of each population, σ is the bootstrapped sample standard deviation, and n is the number of samples. Populations 1 and 2 in this case correspond to polyreactive and non-polyreactive antibodies. To calculate z for the empirical distribution (z_0), all values correspond to the empirical rather than bootstrapped values.

To calculate p-values for differences in mutual information, the permutation test was used rather than the Studentized bootstrap. Here, the test statistic t is set to a simple difference of means, and rather than sampling with replacement from the empirical or null distributions with replacement, we randomly permute the data into "polyreactive" or "non-polyreactive" bins. We then count the number of permutations where the randomly permuted test statistic is greater than or equal to the empirical test statistic. This count then replaces the count ($\sharp$) in the above equation for p.

Specifics for Generating Simulated Systems

All simulations performed were prepared using the CHARMM-GUI Input Generator [123, 124, 215]. Generally, each all-atom simulated system was fully hydrated with TIP3P water molecules and neutralized with 0.15 M KCl. All simulations were carried out in simulation boxes with periodic boundary conditions using the additive PARAM36 force field from the CHARMM (Chemistry at HARvard Macromolecular Mechanics) [123]. Simulation

specifics varied for each system and but used a combination of NAMD, OpenMM, and AMBER [124, 216–218]. For all brute force simulated systems run on the Midway Computing Cluster at the University of Chicago, at least two replicas were run to confirm independence of results on initial velocity assignments. Larger replica exchange umbrella sampling simulations for the calculation of the binding free energy using the string method were run on the Blue Waters high performance computing cluster at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign. Simulation specifics for each system are described below.

pAg, which has not been parameterized previously for molecular simulation, was modeled with custom parameter files based upon other well-characterized diphosphate parameters. BFI Dimer I and Dimer II simulations were run in rectangular simulation boxes to minimize the number of atoms in the system and were run using the AMBER16 GPU software [218]. To get fast and robust sampling, hydrogen mass repartitioning of both the protein and solvent, a frictional coefficient of 0.3 to lower effective viscosity, and a 4-fs time step were used [219]. To ensure the dimers did not float out of the rectangular simulation box and interact with periodic images a spatial constraint was applied to non-interfacial residues on one of the monomers in the simulation box. A third, unconstrained replica set of dimer simulations was run using the NAMD software package with a 2-fs time step for 250 ns to confirm that the steps taken to enhance sampling did not affect the outcome of the simulations [216]. BFI monomer simulations with and without pAg were run in cubic simulation boxes on graphical processing units (GPUs) using the Amber toolkit with a 4-fs time step at 303.15 K. Each replica was equilibrated with a 1-fs time step for 250 ps and run for 500 ns of simulation time. Data were analyzed using Python once the simulation was fully equilibrated, when the backbone RMSD reached a stable value.

Equilibrated systems use an NVT ensemble and production runs use an NPT ensemble, with the temperature kept constant using Langevin dynamics [220]. The simulations were kept at constant pressure at one bar with the Nosé–Hoover Langevin piston by allowing the

cell box size to change semi-isotropically [221]. van derWaals interactions were computed using a Lennard-Jones force-switching function over 10–12 Å while long-range electrostatics used particle mesh Ewald [222]. Production runs for non-GPU simulations used a 2-fs time step and the SHAKE algorithm to constrain the bonds having hydrogen atoms [223].

Raw simulation data was processed using PyEMMA, a Python library for the generation of Markov models [101] to extract relevant protein parameters. Structural visualizations and alignments were carried out using VMD [224]. Hydrogen bonds were determined using a 3.2 Å distance cutoff and 20° angle cutoff in VMD. Solvent accessible surface area (SASA), root mean square deviation (RMSD), and root mean square fluctuation (RMSF) was calculated using the Python package MDTraj [225]. SASA calculation have a probe radius of 1.4 Å, with complete sidechains included in the calculations. Further data processing used custom Python scripts with trajectory featurization and data handling provided by the MDTraj library [225]. In our analysis, we perform all tICA decompositions using a fixed choice of time-lag at 1 ns to make the analysis more comparable across decompositions. We then calculate distances in this tICA space while clustering using the standard Euclidean metric after projecting the data frames onto the first 4 tICA degrees of freedom. All clustering was done using the K-centers algorithm.

Experimental Methods

Protein Purification

The full-length BTN3A1 was expressed from the pAcGP67A baculovirus transfer vector (BD Biosciences) with a C-terminal 12x-polyhistidine-tag (12xHis). This construct was used for the production of recombinant baculoviruses using BestBac linearized baculovirus DNA (Expression Systems). High Five insect cell culture was infected with baculoviruses encoding BTN3A1-12xHis, incubated at 27 degrees C for 72 hours and spun down for 15 min at 4 degrees C and 1,700×g. Cells were washed in 20 mM Hepes buffer, pH 7.2, with 150 mM

NaCl and spun down for 15 min at 4 degrees C and 1,700×g. Pellet was resuspended in 10 mM Tris buffer, pH 7.9, with 1 mM EDTA and protease inhibitors mixture (PIC; Sigma). Cells were lysed in a glass homogenizer. Lysate was spun down for 30 min at 4 degrees C and 40,000×g and the pellet containing membrane fraction was collected. BTN3A1-12xHis was extracted from the membrane with 50 mM Tris buffer, pH7.9, with 150 mM NaCl, 1% (vol/vol) Triton X-100 (TX-100; AcrosOrganics) and PIC, rotated at 4 °C for 1 h. The suspension was spun down for 30 min at 4 degrees C and 40,000×g. Detergent soluble fraction was collected and incubated with Ni-NTA resin (Qiagen) in the presence of 30 mM imidazole and 5 mM 2-mercaptoethanol for 1 hour at 4 degrees C. The resin was washed with 50 mM Tris buffer, pH 7.9, with 150 mM NaCl, 30 mM imidazole, 0.1% (vol/vol) TX-100, and 5 mM 2-mercaptoethanol. Protein was eluted with 50 mM Tris buffer, pH 7.9, 150 mM NaCl, 500 mM imidazole, 0.1% (vol/vol) TX-100, and 5 mM 2-mercaptoethanol. Sample containing BTN3A1-12xHis homodimer was concentrated using Amicon Ultra filter with 100-kDa molecular cutoff (Millipore) and purified by gel filtration on Superdex 200 10/300 GL column (GE Healthcare) in 10 mM Tris buffer, pH 7.9, with 50 mM NaCl and 0.03% (vol/vol) TX-100. Reconstitution of detergent-solubilized His-tagged BTN3A1 dimers into liposomes was carried out as in Garten et al. [226]

Atomic Force Microscopy Preparation

The POPC (Avanti) lipids purchased in chloroform were dried in glass vials under a stream of a nitrogen gas. Residual chloroform was additionally removed by overnight incubation under vacuum desiccator; 10 mM of lipids were then dissolved in buffer A (50 mM Hepes, pH 7.5, and 200 mM NaCl)supplemented with 20 mM Triton X-100 (TX-100) and sonicated until clear using a digital sonicator (Branson) with a tapered microtip. Lipids were extruded through a 100nm filter membrane to generate large unilamellar vesicles of similar diameter.

All AFM experiments were conducted using an Asylum Research Cypher AFM, equipped

with the Environmental Scanner (ES). The protein-reconstituted lipid vesicle solution was then applied to freshly cleaved mica (Ted Pella, Product 50) and allowed to incubate for 30 minutes. The Olympus BL-AC40TS cantilever (Tip Radius 8nm, Spring Constant 0.09 N/m) a small, soft cantilever used for biological measurements was used to acquire data on the Cypher ES. Free amplitudes of tips were kept low (3nm) to avoid disruption of the deposited membrane surfaces. All data were acquired using tapping mode AFM, with oscillation of the cantilever controlled by the Asylum blueDrive™ photothermal excitation method. Data analysis was done using Igor Pro's built in "imageThreshold" particle picking function, with a 1nm cutoff for the selection mask. Once particles were chosen using the "imageThreshold" function, properties such as particle height could be acquired for either single particles or the population mean. Further data analysis and visualization was carried out using home built Python scripts.

REFERENCES

[1] Jennifer A. Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 2014.

[2] Martin F. Flajnik and Louis Du Pasquier. Evolution of innate and adaptive immunity: Can we draw a line? *Trends in Immunology*, 2004.

[3] Nadia Danilova. The evolution of immune mechanisms. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 2006.

[4] Caitlin D. Castro, Yuko Ohta, Helen Dooley, and Martin F. Flajnik. Noncoordinate expression of J-chain and Blimp-1 define nurse shark plasma cell populations during ontogeny. *European Journal of Immunology*, 2013.

[5] Jeannine A. Ott, Caitlin D. Castro, Thaddeus C. Deiss, Yuko Ohta, Martin F. Flajnik, and Michael F. Criscitiello. Somatic hypermutation of T cell receptor α chain contributes to selection in nurse shark thymus. *eLife*, 2018.

[6] Kenneth Murphy and Casey Weaver. *Janeway's Immunobiology*. 2016.

[7] David G. Schatz and Yanhong Ji. Recombination centres and the orchestration of V(D)J recombination. *Nature Reviews Immunology*, 2011.

[8] Gabriel D. Victora and Michel C. Nussenzweig. Germinal Centers. *Annual Review of Immunology*, 2012.

[9] Herman N. Eisen and Gregory W. Siskind. Variations in Affinities of Antibodies during the Immune Response. *Biochemistry*, 1964.

[10] D. McKean, K. Huppi, M. Bell, L. Staudt, W. Gerhard, and M. Weigert. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, 1984.

[11] I Correa, M Bix, N S Liao, M Zijlstra, R Jaenisch, and D Raulet. Most gamma delta T cells develop normally in beta 2-microglobulin-deficient mice. *Proceedings of the National Academy of Sciences of the United States of America*, 1992.

[12] M Bigby, J S Markowitz, P A Bleicher, M J Grusby, S Simha, M Siebrecht, M Wagner, C Nagler-Anderson, and L H Glimcher. Most gamma delta T cells develop normally in the absence of MHC class II molecules. *Journal of immunology (Baltimore, Md. : 1950)*, 1993.

[13] Rachel West, Amanda Kobokovich, Nancy Connell, and Gigi Kwik Gronvall. COVID-19 Antibody Tests: A Valuable Public Health Tool with Limited Relevance to Individuals. *Trends in Microbiology*, 2020.

[14] David M. Weinreich, Sumathi Sivapalasingam, Thomas Norton, Shazia Ali, Haitao Gao, Rafia Bhore, Bret J. Musser, Yuhwen Soo, Diana Rofail, Joseph Im, Christina Perry, Cynthia Pan, Romana Hosain, Adnan Mahmood, John D. Davis, Kenneth C. Turner, Andrea T. Hooper, Jennifer D. Hamilton, Alina Baum, Christos A. Kyratsous, Yunji Kim, Amanda Cook, Wendy Kampman, Anita Kohli, Yessica Sachdeva, Ximena Graber, Bari Kowal, Thomas DiCioccio, Neil Stahl, Leah Lipsich, Ned Braunstein, Gary Herman, and George D. Yancopoulos. REGN-COV2, a Neutralizing Antibody Cocktail, in Outpatients with Covid-19. *New England Journal of Medicine*, 2020.

[15] Dana R. Leach, Matthew F. Krummel, and James P. Allison. Enhancement of antitumor immunity by CTLA-4 blockade. *Science*, 1996.

[16] Y. Ishida, Y. Agata, K. Shibahara, and T. Honjo. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO Journal*, 1992.

[17] Padmanee Sharma and James P. Allison. Immune checkpoint targeting in cancer therapy: Toward combination strategies with curative potential. *Cell*, 2015.

[18] Tushar Jain, Todd Boland, Asparouh Lilov, Irina Burnina, Michael Brown, Yingda Xu, and Maximiliano Vásquez. Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. *Bioinformatics*, 2017.

[19] Matthew I.J. Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P. Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America*, 2019.

[20] Vikas K. Sharma, Thomas W. Patapoff, Bruce Kabakoff, Satyan Pai, Eric Hilario, Boyan Zhang, Charlene Li, Oleg Borisov, Robert F. Kelley, Ilya Chorny, Joe Z. Zhou, Ken A. Dill, and Trevor E. Swartz. In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[21] Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H. Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T. Gray, Eric M. Krauland, Yingda Xu, Maximiliano Vásquez, and K. Dane Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 2017.

[22] Olga Obrezanova, Andreas Arnell, Ramón Gómez De La Cuesta, Maud E. Berthelot, Thomas R.A. Gallagher, Jesús Zurdo, and Yvette Stallwood. Aggregation risk prediction for antibodies and its application to biotherapeutic development. *mAbs*, 2015.

[23] Charles G. Starr and Peter M. Tessier. Selecting and engineering monoclonal antibodies with drug-like specificity. *Current Opinion in Biotechnology*, 2019.

[24] Isidro Hötzel, Frank Peter Theil, Lisa J. Bernstein, Saileta Prabhu, Rong Deng, Leah Quintana, Jeff Lutman, Renuka Sibia, Pamela Chan, Daniela Bumbaca, Paul Fielder, Paul J. Carter, and Robert F. Kelley. A strategy for risk mitigation of antibodies with fast clearance. *mAbs*, 2012.

[25] Ryan L. Kelly, Tingwan Sun, Tushar Jain, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Maximiliano Vásquez, K. Dane Wittrup, and Yingda Xu. High throughput cross-interaction measures for human IgG1 antibodies correlate with clearance rates in mice. *mAbs*, 2015.

[26] Ryan L. Kelly, Doris Le, Jessie Zhao, and K. Dane Wittrup. Reduction of Nonspecificity Motifs in Synthetic Antibody Libraries. *Journal of Molecular Biology*, 2018.

[27] Amita Datta-Mannan, Jirong Lu, Derrick R. Witcher, Donmienne Leung, Ying Tang, and Victor J. Wroblewski. The interplay of non-specific binding, targetmediated clearance and FcRn interactions on the pharmacokinetics of humanized antibodies. *mAbs*, 2015.

[28] Carolin Daniel, Jens Nolting, and Harald Von Boehmer. Mechanisms of self-nonself discrimination and possible clinical relevance. *Immunotherapy*, 2009.

[29] Jedd D. Wolchok. Altered self: The not-so-neo-antigens. *Nature Reviews Immunology*, 2018.

[30] Patrick A. Ott, Zhuting Hu, Derin B. Keskin, Sachet A. Shukla, Jing Sun, David J. Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, Christina Chen, Oriol Olive, Todd A. Carter, Shuqiang Li, David J. Lieb, Thomas Eisenhaure, Evisa Gjini, Jonathan Stevens, William J. Lane, Indu Javeri, Kaliappanadar Nellaiappan, Andres M. Salazar, Heather Daley, Michael Seaman, Elizabeth I. Buchbinder, Charles H. Yoon, Maegan Harden, Niall Lennon, Stacey Gabriel, Scott J. Rodig, Dan H. Barouch, Jon C. Aster, Gad Getz, Kai Wucherpfennig, Donna Neuberg, Jerome Ritz, Eric S. Lander, Edward F. Fritsch, Nir Hacohen, and Catherine J. Wu. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 2017.

[31] Ugur Sahin, Evelyna Derhovanessian, Matthias Miller, Björn Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, Arbel D. Tadmor, Ulrich Luxemburger, Barbara Schrörs, Tana Omokoko, Mathias Vormehr, Christian Albrecht, Anna Paruzynski, Andreas N. Kuhn, Janina Buck, Sandra Heesch, Katharina H. Schreeb, Felicitas Müller, Inga Ortseifer, Isabel Vogler, Eva Godehardt, Sebastian Attig, Richard Rae, Andrea Breitkreuz, Claudia Tolliver, Martin Suchan, Goran Martic, Alexander Hohberger, Patrick Sorn, Jan Diekmann, Janko Ciesla, Olga Waksmann, Alexandra Kemmer Brück, Meike Witt, Martina Zillgen, Andree Rothermel, Barbara Kasemann, David

Langer, Stefanie Bolte, Mustafa Diken, Sebastian Kreiter, Romina Nemecek, Christoffer Gebhardt, Stephan Grabbe, Christoph Höller, Jochen Utikal, Christoph Huber, Carmen Loquai, and Özlem Türeci. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, 2017.

[32] Colby S. Shemesh, Joy C. Hsu, Iraj Hosseini, Ben Quan Shen, Anand Rotte, Patrick Twomey, Sandhya Girish, and Benjamin Wu. Personalized Cancer Vaccines: Clinical Landscape, Challenges, and Opportunities. *Molecular Therapy*, 2020.

[33] Matteo D'Antonio, Joaquin Reyna, David Jakubosky, Margaret K.R. Donovan, Marc Jan Bonder, Hiroko Matsui, Oliver Stegle, Naoki Nariai, Agnieszka D'antonio-Chronowska, and Kelly A. Frazer. Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. *eLife*, 2019.

[34] Gerald P. Linette, Edward A. Stadtmauer, Marcela V. Maus, Aaron P. Rapoport, Bruce L. Levine, Lyndsey Emery, Leslie Litzky, Adam Bagg, Beatriz M. Carreno, Patrick J. Cimino, Gwendolyn K. Binder-Scholl, Dominic P. Smethurst, Andrew B. Gerry, Nick J. Pumphrey, Alan D. Bennett, Joanna E. Brewer, Joseph Dukes, Jane Harper, Helen K. Tayton-Martin, Bent K. Jakobsen, Namir J. Hassan, Michael Kalos, and Carl H. June. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood*, 2013.

[35] Richard A. Morgan, Nachimuthu Chinnasamy, Daniel Abate-Daga, Alena Gros, Paul F. Robbins, Zhili Zheng, Mark E. Dudley, Steven A. Feldman, James C. Yang, Richard M. Sherry, Giao Q. Phan, Marybeth S. Hughes, Udai S. Kammula, Akemi D. Miller, Crystal J. Hessman, Ashley A. Stewart, Nicholas P. Restifo, Martha M. Quezado, Meghna Alimchandani, Avi Z. Rosenberg, Avindra Nath, Tongguang Wang, Bibiana Bielekova, Simone C. Wuest, Nirmala Akula, Francis J. McMahon, Susanne Wilde, Barbara Mosetter, Dolores J. Schendel, Carolyn M. Laurencot, and Steven A. Rosenberg. Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *Journal of Immunotherapy*, 2013.

[36] Federico Garrido, Natalia Aptsiauri, Elien M. Doorduijn, Angel M. Garcia Lora, and Thorbald van Hall. The urgent need to recover MHC class I in cancers for effective immunotherapy. *Current Opinion in Immunology*, 2016.

[37] Jordan D. Dimitrov, Cyril Planchais, Lubka T. Roumenina, Tchavdar L. Vassilev, Srinivas V. Kaveri, and Sebastien Lacroix-Desmazes. Antibody Polyreactivity in Health and Disease: Statu Variabilis. *The Journal of Immunology*, 2013.

[38] Adrian F. Ochsenbein, Thomas Fehr, Claudia Lutz, Mark Suter, Frank Brombacher, Hans Hengartner, and Rolf M. Zinkernagel. Control of early viral and bacterial distribution and disease by natural antibodies. *Science*, 1999.

[39] Hedda Wardemann, Sergey Yurasov, Anne Schaefer, James W. Young, Eric Meffre, and Michel C. Nussenzweig. Predominant autoantibody production by early human B

cell precursors. *Science*, 2003.

[40] Thomas Tiller, Makoto Tsuiji, Sergey Yurasov, Klara Velinzon, Michel C. Nussenzweig, and Hedda Wardemann. Autoreactivity in Human IgG+ Memory B Cells. *Immunity*, 2007.

[41] Hugo Mouquet, Johannes F. Scheid, Markus J. Zoller, Michelle Krogsgaard, Rene G. Ott, Shetha Shukair, Maxim N. Artyomov, John Pietzsch, Mark Connors, Florencia Pereyra, Bruce D. Walker, David D. Ho, Patrick C. Wilson, Michael S. Seaman, Herman N. Eisen, Arup K. Chakraborty, Thomas J. Hope, Jeffrey V. Ravetch, Hedda Wardemann, and Michel C. Nussenzweig. Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. *Nature*, 2010.

[42] Julie Prigent, Valérie Lorin, Ayrin Kök, Thierry Hieu, Salomé Bourgeau, and Hugo Mouquet. Scarcity of autoreactive human blood IgA+ memory B cells. *European Journal of Immunology*, 2016.

[43] Kristi Koelsch, Nai Ying Zheng, Qingzhao Zhang, Andrew Duty, Christina Helms, Melissa D. Mathias, Mathew Jared, Kenneth Smith, J. Donald Capra, and Patrick C. Wilson. Mature B cells class switched to IgD are autoreactive in healthy individuals. *Journal of Clinical Investigation*, 2007.

[44] Fabian Paul and Thomas R. Weikl. How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates. *PLoS Computational Biology*, 2016.

[45] Gordon G. Hammes, Yu Chu Chang, and Terrence G. Oas. Conformational selection or induced fit: A flux description of reaction mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 2009.

[46] James E. Crooks, Christopher T. Boughter, L. Ridgway Scott, and Erin J. Adams. The hypervariable loops of free TCRs sample multiple distinct metastable conformations in solution. *Frontiers in Molecular Biosciences*, 2018.

[47] Jenna J. Guthmiller, Linda Yu-Ling Lan, Monica L. Fernández-Quintero, Julianna Han, Henry A. Utset, Dalia J. Bitar, Natalie J. Hamel, Olivia Stovicek, Lei Li, Micah Tepora, Carole Henry, Karlynn E. Neu, Haley L. Dugan, Marta T. Borowska, Yao-Qing Chen, Sean T.H. Liu, Christopher T. Stamper, Nai-Ying Zheng, Min Huang, Anna-Karin E. Palm, Adolfo García-Sastre, Raffael Nachbagauer, Peter Palese, Lynda Coughlan, Florian Krammer, Andrew B. Ward, Klaus R. Liedl, and Patrick C. Wilson. Polyreactive Broadly Neutralizing B cells Are Selected to Provide Defense against Pandemic Threat Influenza Viruses. *Immunity*, 2020.

[48] Lewis L. Lanier, Joyce Ruitenberg, Reinder L.H. Bolhuls, Jannie Borst, Joseph H. Phillis, and Roberto Testi. Structural and serological heterogeneity of γ/δ T cell antigen receptor expression in thymus and peripherar blood. *European Journal of Immunology*, 1988.

[49] V. Groh, S. Porcelli, M. Fabii, L. L. Lanier, L. J. Picker, T. Anderson, R. A. Warnke, A. K. Bhan, J. L. Strominger, and M. B. Brenner. Human lymphocytes bearing T cell receptor γ/δ are phenotypically diverse and evenly distributed throughout the lymphoid system. *Journal of Experimental Medicine*, 1989.

[50] Toufic Mayassi and Bana Jabri. Human intraepithelial lymphocytes. *Mucosal Immunology*, 2018.

[51] S. Meraviglia, E. Lo Presti, M. Tosolini, C. La Mendola, V. Orlando, M. Todaro, V. Catalano, G. Stassi, G. Cicero, S. Vieni, J. J. Fourniè, and F. Dieli. Distinctive features of tumor-infiltrating $\gamma\delta$ T lymphocytes in human colorectal cancer. *OncoImmunology*, 2017.

[52] Caitlin D. Castro, Christopher T. Boughter, Augusta E. Broughton, Amrita Ramesh, and Erin J. Adams. Diversity in recognition and function of human $\gamma\delta$ T cells. *Immunological Reviews*, 2020.

[53] W. W. Franke, H. W. Heid, C. Grund, S. Winter, C. Freudenstein, E. Schmid, E. D. Jarasch, and T. W. Keenan. Antibodies to the major insoluble milk fat globule membrane-associated protein: Specific location in apical regions of lactating epithelial cells. *Journal of Cell Biology*, 1981.

[54] Ian H. Mather and Lucinda J.W. Jacks. A Review of the Molecular and Cellular Biology of Butyrophilin, the Major Protein of Bovine Milk Fat Globule Membrane. *Journal of Dairy Science*, 1993.

[55] Christelle Harly, Yves Guillaume, Steven Nedellec, Cassie Marie Peigné, Hannu Mönkkönen, Jukka Mönkkönen, Jianqiang Li, Jürgen Kuball, Erin J. Adams, Sonia Netzer, Julie Déchanet-Merville, Alexandra Léger, Thomas Herrmann, Richard Breathnach, Daniel Olive, Marc Bonneville, and Emmanuel Scotet. Key implication of CD277/butyrophilin-3 (BTN3A) in cellular stress sensing by a major human $\gamma\delta$ T-cell subset. *Blood*, 2012.

[56] Andrew Sandstrom, Cassie Marie Peigné, Alexandra Léger, James E. Crooks, Fabienne Konczak, Marie Claude Gesnel, Richard Breathnach, Marc Bonneville, Emmanuel Scotet, and Erin J. Adams. The intracellular B30.2 domain of butyrophilin 3A1 binds phosphoantigens to mediate activation of human Vγ9Vδ2T Cells. *Immunity*, 2014.

[57] Aparna Palakodeti, Andrew Sandstrom, Lakshmi Sundaresan, Christelle Harly, Steven Nedellec, Daniel Olive, Emmanuel Scotet, Marc Bonneville, and Erin J. Adams. The molecular basis for modulation of human Vγ9Vδ2 T cell responses by CD277/butyrophilin-3 (BTN3A)-specific antibodies. *Journal of Biological Chemistry*, 2012.

[58] Lisa Starick, Felipe Riano, Mohindar M. Karunakaran, Volker Kunzmann, Jianqiang Li, Matthias Kreiss, Sabine Amslinger, Emmanuel Scotet, Daniel Olive, Gennaro De

Libero, and Thomas Herrmann. Butyrophilin 3A (BTN3A, CD277)-specific antibody 20.1 differentially activates Vγ9Vδ2 TCR clonotypes and interferes with phosphoantigen activation. *European Journal of Immunology*, 2017.

[59] D. A. Rhodes, M. Stammers, G. Malcherek, S. Beck, and J. Trowsdale. The cluster of BTN genes in the extended major histocompatibility complex. *Genomics*, 2001.

[60] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 1958.

[61] M. F. Perutz, M. G. Rossmann, Ann F. Cullis, Hilary Muirhead, Georg Will, and A. C.T. North. Structure of Hæmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by X-ray analysis. *Nature*, 1960.

[62] Stephen K. Burley, Helen M. Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, Ken Dalenberg, Jose M. Duarte, Shuchismita Dutta, Zukang Feng, Sutapa Ghosh, David S. Goodsell, Rachel K. Green, Vladimir Guranović, Dmytro Guzenko, Brian P. Hudson, Tara Kalro, Yuhe Liang, Robert Lowe, Harry Namkoong, Ezra Peisach, Irina Periskova, Andreas Prlić, Chris Randle, Alexander Rose, Peter Rose, Raul Sala, Monica Sekharan, Chenghua Shao, Lihua Tan, Yi Ping Tao, Yana Valasatava, Maria Voigt, John Westbrook, Jesse Woo, Huanwang Yang, Jasmine Young, Marina Zhuravleva, and Christine Zardecki. RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, 2019.

[63] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 2005.

[64] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 1977.

[65] Max D. Cooper. The early history of B cells. *Nature Reviews Immunology*, 2015.

[66] Jacques F.A.P. Miller. The function of the thymus and its impact on modern medicine. *Science*, 2020.

[67] Jeffrey J. Bunker, Steven A. Erickson, Theodore M. Flynn, Carole Henry, Jason C. Koval, Marlies Meisel, Bana Jabri, Dionysios A. Antonopoulos, Patrick C. Wilson, and Albert Bendelac. Natural polyreactive IgA antibodies coat the intestinal microbiota. *Science*, 2017.

[68] Cyril Planchais, Ayrin Kök, Alexia Kanyavuz, Valérie Lorin, Timothée Bruel, Florence Guivel-Benhassine, Tim Rollenske, Julie Prigent, Thierry Hieu, Thierry Prazuck, Laurent Lefrou, Hedda Wardemann, Olivier Schwartz, Jordan D. Dimitrov, Laurent Hocqueloux, and Hugo Mouquet. HIV-1 Envelope Recognition by Polyreactive and

Cross-Reactive Intestinal B Cells. *Cell Reports*, 2019.

[69] Barton F. Haynes, Judith Fleming, E. William St. Clair, Herman Katinger, Gabriela Stiegler, Renate Kunert, James Robinson, Richard M. Scearce, Kelly Plonk, Herman F. Staats, Thomas L. Ortel, Hua Xin Liao, and S. Munir Alam. Immunology: Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. *Science*, 2005.

[70] Hugo Mouquet, Florian Klein, Johannes F. Scheid, Malte Warncke, John Pietzsch, Thiago Y.K. Oliveira, Klara Velinzon, Michael S. Seaman, and Michel C. Nussenzweig. Memory B cell antibodies to HIV-1 gp140 cloned from individuals infected with clade A and B viruses. *PLoS ONE*, 2011.

[71] Sarah F. Andrews, Yunping Huang, Kaval Kaur, Lyubov I. Popova, Irvin Y. Ho, Noel T. Pauli, Carole J.Henry Dunand, William M. Taylor, Samuel Lim, Min Huang, Xinyan Qu, Jane Hwei Lee, Marlene Salgado-Ferrer, Florian Krammer, Peter Palese, Jens Wrammert, Rafi Ahmed, and Patrick C. Wilson. Immune history profoundly affects broadly protective B cell responses to influenza. *Science Translational Medicine*, 2015.

[72] Julie Prigent, Annaëlle Jarossay, Cyril Planchais, Caroline Eden, Jérémy Dufloo, Ayrin Kök, Valérie Lorin, Oxana Vratskikh, Thérèse Couderc, Timothée Bruel, Olivier Schwartz, Michael S. Seaman, Oliver Ohlenschläger, Jordan D. Dimitrov, and Hugo Mouquet. Conformational Plasticity in Broadly Neutralizing HIV-1 Antibodies Triggers Polyreactivity. *Cell Reports*, 2018.

[73] Barton F. Haynes, Dennis R. Burton, and John R. Mascola. Multiple roles for HIV broadly neutralizing antibodies. *Science Translational Medicine*, 2019.

[74] Trevor A. Crowell, Donn J. Colby, Suteeraporn Pinyakorn, Carlo Sacdalan, Amélie Pagliuzza, Jintana Intasan, Khunthalee Benjapornpong, Kamonkan Tangnaree, Nitiya Chomchey, Eugène Kroon, Mark S. de Souza, Sodsai Tovanabutra, Morgane Rolland, Michael A. Eller, Dominic Paquin-Proulx, Diane L. Bolton, Andrey Tokarev, Rasmi Thomas, Hiroshi Takata, Lydie Trautmann, Shelly J. Krebs, Kayvon Modjarrad, Adrian B. McDermott, Robert T. Bailer, Nicole Doria-Rose, Bijal Patel, Robert J. Gorelick, Brandie A. Fullmer, Alexandra Schuetz, Pornsuk V. Grandin, Robert J. O'Connell, Julie E. Ledgerwood, Barney S. Graham, Randall Tressler, John R. Mascola, Nicolas Chomont, Nelson L. Michael, Merlin L. Robb, Nittaya Phanuphak, Jintanat Ananworanich, Julie A. Ake, Siriwat Akapirat, Meera Bose, Evan Cale, Phillip Chan, Sararut Chanthaburanun, Nampueng Churikanont, Peter Dawson, Netsiri Dumrongpisutikul, Saowanit Getchalarat, Surat Jongrakthaitae, Krisada Jongsakul, Sukalaya Lerdlum, Sopark Manasnayakorn, Corinne McCullough, Mark Milazzo, Bessara Nuntapinit, Kier On, Madelaine Ouellette, Praphan Phanuphak, Eric Sanders-Buell, Nongluck Sangnoi, Shida Shangguan, Sunee Sirivichayakul, Nipattra Tragonlugsana, Rapee Trichavaroj, Sasiwimol Ubolyam, Sandhya Vasan, Phandee Wattanaboonyongcharoen, and Thipvadee Yamchuenpong. Safety and efficacy of VRC01 broadly neutralising antibodies in adults with acutely treated HIV (RV397): a phase 2, randomised,

double-blind, placebo-controlled trial. *The Lancet HIV*, 2019.

[75] Gui Mei Li, Christopher Chiu, Jens Wrammert, Megan McCausland, Sarah F. Andrews, Nai Ying Zheng, Jane Hwei Lee, Min Huang, Xinyan Qu, Srilatha Edupuganti, Mark Mulligan, Suman R. Das, Jonathan W. Yewdell, Aneesh K. Mehta, Patrick C. Wilson, and Rafi Ahmed. Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells. *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[76] Joshua S. Klein and Pamela J. Bjorkman. Few and far between: How HIV may be evading antibody avidity. *PLoS Pathogens*, 2010.

[77] Maxime Lecerf, Alexia Kanyavuz, Sébastien Lacroix-Desmazes, and Jordan D. Dimitrov. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. *Molecular Immunology*, 2019.

[78] Lilia A. Rabia, Yulei Zhang, Seth D. Ludwig, Mark C. Julian, and Peter M. Tessier. Net charge of antibody complementarity-determining regions is a key predictor of specificity. *Protein engineering, design & selection : PEDS*, 2018.

[79] Sara Birtalan, Yingnan Zhang, Frederic A. Fellouse, Lihua Shao, Gabriele Schaefer, and Sachdev S. Sidhu. The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies. *Journal of Molecular Biology*, 2008.

[80] Yasaman Karami, Julien Rey, Guillaume Postic, Samuel Murail, Pierre Tufféry, and Sjoerd J. De Vries. DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Research*, 2019.

[81] Karlynn E. Neu, Jenna J. Guthmiller, Min Huang, Jennifer La, Marcos C. Vieira, Kangchon Kim, Nai Ying Zheng, Mario Cortese, Micah E. Tepora, Natalie J. Hamel, Karla Thatcher Rojas, Carole Henry, Dustin Shaw, Charles L. Dulberger, Bali Pulendran, Sarah Cobey, Aly A. Khan, and Patrick C. Wilson. Spec-seq unveils transcriptional subpopulations of antibody-secreting cells following influenza vaccination. *Journal of Clinical Investigation*, 2019.

[82] Jens Wrammert, Dimitrios Koutsonanos, Gui Mei Li, Srilatha Edupuganti, Jianhua Sui, Michael Morrissey, Megan McCausland, Ioanna Skountzou, Mady Hornig, W. Ian Lipkin, Aneesh Mehta, Behzad Razavi, Carlos Del Rio, Nai Ying Zheng, Jane Hwei Lee, Min Huang, Zahida Ali, Kaval Kaur, Sarah Andrews, Rama Rao Amara, Youliang Wang, Suman Ranjan Das, Christopher David O'Donnell, Jon W. Yewdell, Kanta Subbarao, Wayne A. Marasco, Mark J. Mulligan, Richard Compans, Rafi Ahmed, and Patrick C. Wilson. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *Journal of Experimental Medicine*, 2011.

[83] Pradyot Dash, Andrew J. Fiore-Gartland, Tomer Hertz, George C. Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E. Bridie Clemens, Thi H.O. Nguyen, Katherine Kedzierska, Nicole L. La Gruta, Philip Bradley, and Paul G. Thomas. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 2017.

[84] Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research*, 2013.

[85] Akinori Kidera, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A. Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 1985.

[86] Quanya Liu, Peng Chen, Bing Wang, Jun Zhang, and Jinyan Li. Hot spot prediction in protein-protein interactions by an ensemble system. *BMC Systems Biology*, 2018.

[87] Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 2003.

[88] Marli Tenório Cordeiro, Ulisses Braga-Neto, Rita Maria Ribeiro Nogueira, and Ernesto T.A. Marques. Reliable classifier to differentiate primary and secondary acute dengue infection based on IgG ELISA. *PLoS ONE*, 2009.

[89] Yuqian Ma, David Vilanova, Kerem Atalar, Olivier Delfour, Jonathan Edgeworth, Marlies Ostermann, Maria Hernandez-Fuentes, Sandrine Razafimahatratra, Bernard Michot, David H. Persing, Ingrid Ziegler, Bianca Törös, Paula Mölling, Per Olcén, Richard Beale, and Graham M. Lord. Genome-Wide Sequencing of Cellular microRNAs Identifies a Combinatorial Expression Signature Diagnostic of Sepsis. *PLoS ONE*, 2013.

[90] James R.R. Whittle, Ruijun Zhang, Surender Khurana, Lisa R. King, Jody Manischewitz, Hana Golding, Philip R. Dormitzer, Barton F. Haynes, Emmanuel B. Walter, M. Anthony Moody, Thomas B. Kepler, Hua Xin Liao, and Stephen C. Harrison. Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, 2011.

[91] Marta T. Borowska. Molecular Insights into Unconventional Immune Recognition: A Case of Commensal Superantigens, Polyreactive Antibodies and Butyrophilin Signaling. *The University of Chicago*, 2020.

[92] Fang Chen, Netanel Tzarum, Ian A. Wilson, and Mansun Law. V H 1-69 antiviral broadly neutralizing antibodies: genetics, structures, and relevance to rational vaccine design. *Current Opinion in Virology*, 2019.

[93] Monica L. Fernández-Quintero, Johannes R. Loeffler, Johannes Kraml, Ursula Kahler,

Anna S. Kamenik, and Klaus R. Liedl. Characterizing the diversity of the CDR-H3 loop conformational ensembles in relationship to antibody binding properties. *Frontiers in Immunology*, 2019.

[94] Aaron G. Schmidt, Huafeng Xu, Amir R. Khan, Timothy O'Donnell, Surender Khurana, Lisa R. King, Jody Manischewitz, Hana Golding, Pirada Suphaphiphat, Andrea Carfi, Ethan C. Settembre, Philip R. Dormitzer, Thomas B. Kepler, Ruijun Zhang, M. Anthony Moody, Barton F. Haynes, Hua Xin Liao, David E. Shaw, and Stephen C. Harrison. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

[95] Tong Li, Malgorzata B. Tracka, Shahid Uddin, Jose Casas-Finet, Donald J. Jacobs, and Dennis R. Livesay. Rigidity Emerges during Antibody Evolution in Three Distinct Antibody Systems: Evidence from QSFR Analysis of Fab Fragments. *PLoS Computational Biology*, 2015.

[96] Jeliazko R. Jeliazkov, Adnan Sljoka, Daisuke Kuroda, Nobuyuki Tsuchimura, Naoki Katoh, Kouhei Tsumoto, and Jeffrey J. Gray. Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification. *Frontiers in Immunology*, 2018.

[97] Boaz Musafia, Virginia Buchner, and Dorit Arad. Complex salt bridges in proteins: Statistical analysis of structure and function. *Journal of Molecular Biology*, 1995.

[98] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 1994.

[99] Yusuke Naritomi and Sotaro Fuchigami. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *Journal of Chemical Physics*, 2011.

[100] Christian R. Schwantes and Vijay S. Pande. Modeling molecular kinetics with tICA and the kernel trick. *Journal of Chemical Theory and Computation*, 2015.

[101] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11:5525–5542, October 2015.

[102] Eric H. Sasso, Ko Willems Van Dijk, Andrew P. Bull, and Eric C.B. Milner. A fetally expressed immunoglobulin V H1 gene belongs to a complex set of alleles. *Journal of Clinical Investigation*, 1993.

[103] Francesco Forconi, Kathleen N. Potter, Isla Wheatley, Nikos Darzentas, Elisa Sozzi,

Kostas Stamatopoulos, C. Ian Mockridge, Graham Packham, and Freda K. Stevenson. The normal IGHV1-69-derived B-cell repertoire contains stereotypic patterns characteristic of unmutated CLL. *Blood*, 2010.

[104] Lei Zhang, Adriana Irimia, Lingling He, Elise Landais, Kimmo Rantalainen, Daniel P. Leaman, Thomas Vollbrecht, Armando Stano, Daniel I. Sands, Arthur S. Kim, George Miiro, Jennifer Serwanga, Anton Pozniak, Dale McPhee, Oliver Manigart, Lawrence Mwananyanda, Etienne Karita, André Inwoley, Walter Jaoko, Jack DeHovitz, Linda Gail Bekker, Punnee Pitisuttithum, Robert Paris, Susan Allen, Pascal Poignard, Dennis R. Burton, Ben Murrell, Andrew B. Ward, Jiang Zhu, Ian A. Wilson, and Michael B. Zwick. An MPER antibody neutralizes HIV-1 using germline features shared among donors. *Nature Communications*, 2019.

[105] Mauno Vihinen, Esa Torkkila, and Pentti Riikonen. Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics*, 1994.

[106] Dhruv K. Sethi, Anupriya Agarwal, Venkatasamy Manivel, Kanury V.S. Rao, and Dinakar M. Salunke. Differential Epitope Positioning within the Germline Antibody Paratope Enhances Promiscuity in the Primary Immune Response. *Immunity*, 2006.

[107] Michael B. Brenner, Joanne McLean, Deno P. Dialynas, Jack L. Strominger, John A. Smith, Frances L. Owen, J. G. Seidman, Stephen Ip, Fred Rosen, and Michael S. Krangel. Identification of a putative second T-cell receptor. *Nature*, 1986.

[108] Robert L. Modlin, Claude Pirmez, Florence M. Hofman, Victoria Torigian, Koichi Uyemura, Thomas H. Rea, Barry R. Bloom, and Michael B. Brenner. Lymphocytes bearing antigen-specific $\gamma\delta$ T-cell receptors accumulate in human infectious disease lesions. *Nature*, 1989.

[109] Eric M. Janis, Stefan H.E. Kaufmann, Ronald H. Schwartz, and Drew M. Pardoll. Activation of $\gamma\delta$ T cells in the primary immune response to mycobacterium tuberculosis. *Science*, 1989.

[110] P Constant, F Davodeau, M a Peyrat, Y Poquet, G Puzo, M Bonneville, and J J Fournié. Stimulation of human gamma delta T cells by nonpeptidic mycobacterial ligands. *Science (New York, N.Y.)*, 1994.

[111] Y Tanaka, C T Morita, E Nieves, M B Brenner, and B R Bloom. Natural and synthetic non-peptide antigens recognized by human gamma delta T cells. *Nature*, 1995.

[112] Craig T. Morita, Evan M. Beckman, Jack F. Bukowski, Yoshimasa Tanaka, Hamid Band, Barry R. Bloom, David E. Golan, and Michael B. Brenner. Direct presentation of nonpeptide prenyl pyrophosphate antigens to human $\gamma\delta$ T cells. *Immunity*, 1995.

[113] Stefano Vavassori, Anil Kumar, Gan Siok Wan, Gundimeda S. Ramanjaneyulu, Marco Cavallari, Sary El Daker, Travis Beddoe, Alex Theodossis, Neal K. Williams, Emma

Gostick, David A. Price, Dinish U. Soudamini, Kong Kien Voon, Malini Olivo, Jamie Rossjohn, Lucia Mori, and Gennaro De Libero. Butyrophilin 3A1 binds phosphorylated antigens and stimulates human $\gamma\delta$ T cells. *Nature Immunology*, 2013.

[114] Hong Wang, Olivier Henry, Mark D. Distefano, Yen-Chih Wang, Johanna Räikkönen, Jukka Mönkkönen, Yoshimasa Tanaka, and Craig T. Morita. Butyrophilin 3A1 Plays an Essential Role in Prenyl Pyrophosphate Stimulation of Human Vγ2Vδ2 T Cells. *The Journal of Immunology*, 2013.

[115] Mahboob Salim, Timothy J. Knowles, Alfie T. Baker, Martin S. Davey, Mark Jeeves, Pooja Sridhar, John Wilkie, Carrie R. Willcox, Hachemi Kadri, Taher E. Taher, Pierre Vantourout, Adrian Hayday, Youcef Mehellou, Fiyaz Mohammed, and Benjamin E. Willcox. BTN3A1 Discriminates $\gamma\delta$ T Cell Phosphoantigens from Nonantigenic Small Molecules via a Conformational Sensor in Its B30.2 Domain. *ACS Chemical Biology*, 2017.

[116] Siyi Gu, Joseph R. Sachleben, Christopher T. Boughter, Wioletta I. Nawrocka, Marta T. Borowska, Jeffrey T. Tarrasch, Georgios Skiniotis, Benoît Roux, and Erin J. Adams. Phosphoantigen-induced conformational change of butyrophilin 3A1 (BTN3A1) and its implication on Vγ9Vδ2 T cell activation. *Proceedings of the National Academy of Sciences*, 2017.

[117] Sunil Kumar, Koel Chaudhury, Prasenjit Sen, and Sujoy K. Guha. Atomic force microscopy: A powerful tool for high-resolution imaging of spermatozoa. *Journal of Nanobiotechnology*, 2005.

[118] David Alsteens, Hermann E. Gaub, Richard Newton, Moritz Pfreundschuh, Christoph Gerber, and Daniel J. Müller. Atomic force microscopy-based characterization and design of biointerfaces. *Nature Reviews Materials*, 2017.

[119] Hongda Wang, Ralph Bash, Jiya G. Yodh, Gordon L. Hager, D. Lohr, and Stuart M. Lindsay. Glutaraldehyde modified mica: A new surface for atomic force microscopy of chromatin. *Biophysical Journal*, 83(6):3619 – 3625, 2002.

[120] Luda S. Shlyakhtenko, Alexander A. Gall, and Yuri L. Lyubchenko. Mica Functionalization for Imaging of DNA and Protein-DNA Complexes with Atomic Force Microscopy. *Cell Imaging Techniques*, 2012.

[121] Josep Canet-Ferrer, Eugenio Coronado, Alicia Forment-Aliaga, and Elena Pinilla-Cienfuegos. Correction of the tip convolution effects in the imaging of nanostructures studied through scanning force microscopy. *Nanotechnology*, 2014.

[122] Francisco Marques-Moros, Alicia Forment-Aliaga, Elena Pinilla-Cienfuegos, and Josep Canet-Ferrer. Mirror effect in atomic force microscopy profiles enables tip reconstruction. *Scientific Reports*, 2020.

[123] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 2008.

[124] Jumin Lee, Xi Cheng, Jason M. Swails, Min Sun Yeom, Peter K. Eastman, Justin A. Lemkul, Shuai Wei, Joshua Buckner, Jong Cheol Jeong, Yifei Qi, Sunhwan Jo, Vijay S. Pande, David A. Case, Charles L. Brooks, Alexander D. MacKerell, Jeffery B. Klauda, and Wonpil Im. 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*, 2016.

[125] Scott A. Hollingsworth and Ron O. Dror. Molecular Dynamics Simulation for All. *Neuron*, 2018.

[126] Efrem Braun, Justin Gilmer, Heather B. Mayes, David L. Mobley, Jacob I. Monroe, Samarjeet Prasad, and Daniel M. Zuckerman. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*, 2019.

[127] Stewart A. Adcock and J. Andrew McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, 2006.

[128] John M. Jumper, Nabil F. Faruk, Karl F. Freed, and Tobin R. Sosnick. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. *PLoS Computational Biology*, 2018.

[129] John M. Jumper, Nabil F. Faruk, Karl F. Freed, and Tobin R. Sosnick. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLoS Computational Biology*, 2018.

[130] David A. Rhodes, Hung-Chang Chen, Amanda J. Price, Anthony H. Keeble, Martin S. Davey, Leo C. James, Matthias Eberl, and John Trowsdale. Activation of Human $\gamma\delta$ T Cells by Cytosolic Interactions of BTN3A1 with Soluble Phosphoantigens and the Cytoskeletal Adaptor Periplakin. *The Journal of Immunology*, 2015.

[131] Zsolt Sebestyen, Wouter Scheper, Anna Vyborova, Siyi Gu, Zuzana Rychnavska, Marleen Schiffler, Astrid Cleven, Coraline Chéneau, Martje Van Noorden, Cassie Marie Peigné, Daniel Olive, Robert Jan Lebbink, Rimke Oostvogels, Tuna Mutis, Gerrit Jan Schuurhuis, Erin J. Adams, Emmanuel Scotet, and Jürgen Kuball. RhoB Mediates Phosphoantigen Recognition by Vγ9Vδ2 T Cell Receptor. *Cell Reports*, 2016.

[132] Marc Rigau, Simone Ostrouska, Thomas S. Fulford, Darryl N. Johnson, Katherine Woods, Zheng Ruan, Hamish E.G. McWilliam, Christopher Hudson, Candani Tutuka, Adam K. Wheatley, Stephen J. Kent, Jose A. Villadangos, Bhupinder Pal, Christian Kurts, Jason Simmonds, Matthias Pelzing, Andrew D. Nash, Andrew Hammet, Anne M. Verhagen, Gino Vairo, Eugene Maraskovsky, Con Panousis, Nicholas A.

Gherardin, Jonathan Cebon, Dale I. Godfrey, Andreas Behren, and Adam P. Uldrich. Butyrophilin 2A1 is essential for phosphoantigen reactivity by gd T cells. *Science*, 2020.

[133] Yunyun Yang, Liping Li, Linjie Yuan, Xiaoying Zhou, Jianxin Duan, Hongying Xiao, Ningning Cai, Shuai Han, Xianqiang Ma, Weidong Liu, Chun Chi Chen, Lingle Wang, Xin Li, Jiahuan Chen, Ning Kang, Jing Chen, Zhixun Shen, Satish R. Malwal, Wanli Liu, Yan Shi, Eric Oldfield, Rey Ting Guo, and Yonghui Zhang. A Structural Change in Butyrophilin upon Phosphoantigen Binding Underlies Phosphoantigen-Mediated Vγ9Vδ2 T Cell Activation. *Immunity*, 2019.

[134] Angélique Boëdec, Hélène Sicard, Jean Dessolin, Gaëtan Herbette, Sophie Ingoure, Cédric Raymond, Christian Belmant, and Jean Louis Kraus. Synthesis and biological activity of phosphonate analogues and geometric isomers of the highly potent phosphoantigen (E)-1-hydroxy-2-methylbut-2-enyl 4-diphosphate. *Journal of Medicinal Chemistry*, 2008.

[135] Benoît Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 1995.

[136] Hyung-june Woo and Benoît Roux. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 2005.

[137] Hong Wang and Craig T. Morita. Sensor Function for Butyrophilin 3A1 in Prenyl Pyrophosphate Stimulation of Human Vγ2Vδ2 T Cells. *The Journal of Immunology*, 2015.

[138] Michael M. Poe, Sherry S. Agabiti, Caroline Liu, Victoria Li, Kelly A. Teske, Chia Hung Christine Hsiao, and Andrew J. Wiemer. Probing the ligand-binding pocket of BTN3A1. *Journal of Medicinal Chemistry*, 2019.

[139] Anna Vyborova, Dennis X. Beringer, Domenico Fasci, Froso Karaiskaki, Eline van Diest, Lovro Kramer, Aram de Haas, Jasper Sanders, Anke Janssen, Trudy Straetemans, Daniel Olive, Jeanette H.W. Leusen, Lola Boutin, Steven Nedellec, Samantha L. Schwartz, Michael J. Wester, Keith A. Lidke, Emmanuel Scotet, Diane Lidke, Albert J.R. Heck, Zsolt Sebestyen, and Jurgen Kuball. γ9δ2T cell diversity and the receptor interface with tumor cells. *Journal of Clinical Investigation*, 2020.

[140] Felipe Riãno, Mohindar M. Karunakaran, Lisa Starick, Jianqiang Li, Claus J. Scholz, Volker Kunzmann, Daniel Olive, Sabine Amslinger, and Thomas Herrmann. Vγ9Vδ2 TCR-activation by phosphorylated antigens requires butyrophilin 3 A1 (BTN3A1) and additional genes on human chromosome 6. *European Journal of Immunology*, 2014.

[141] Alina S. Fichtner, Mohindar M. Karunakaran, Siyi Gu, Christopher T. Boughter, Marta T. Borowska, Lisa Starick, Anna Nöhren, Thomas W. Göbel, Erin J. Adams,

and Thomas Herrmann. Alpaca (Vicugna pacos), the first nonprimate species with a phosphoantigen-reactive Vγ9Vδ2 T cell subset. *Proceedings of the National Academy of Sciences of the United States of America*, 2020.

[142] Mohindar M. Karunakaran, Carrie R. Willcox, Mahboob Salim, Daniel Paletta, Alina S. Fichtner, Angela Noll, Lisa Starick, Anna Nöhren, Charlotte R. Begley, Katie A. Berwick, Raphaël A.G. Chaleil, Vincent Pitard, Julie Déchanet-Merville, Paul A. Bates, Brigitte Kimmel, Timothy J. Knowles, Volker Kunzmann, Lutz Walter, Mark Jeeves, Fiyaz Mohammed, Benjamin E. Willcox, and Thomas Herrmann. Butyrophilin-2A1 Directly Binds Germline-Encoded Regions of the Vγ9Vδ2 TCR and Is Essential for Phosphoantigen Sensing. *Immunity*, 2020.

[143] Joachim L. Schultze. Teaching 'big data' analysis to young immunologists. *Nature Immunology*, 2015.

[144] Yvan Saeys, Sofie Van Gassen, and Bart N. Lambrecht. Computational flow cytometry: Helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 2016.

[145] Matthew H. Spitzer and Garry P. Nolan. Mass Cytometry: Single Cells, Many Features. *Cell*, 2016.

[146] Burkhard Becher, Andreas Schlitzer, Jinmiao Chen, Florian Mair, Hermi R. Sumatoh, Karen Wei Weng Teng, Donovan Low, Christiane Ruedl, Paola Riccardi-Castagnoli, Michael Poidinger, Melanie Greter, Florent Ginhoux, and Evan W. Newell. High-dimensional analysis of the murine myeloid cell system. *Nature Immunology*, 2014.

[147] Matthew H. Spitzer, Pier Federico Gherardini, Gabriela K. Fragiadakis, Nupur Bhattacharya, Robert T. Yuan, Andrew N. Hotson, Rachel Finck, Yaron Carmi, Eli R. Zunder, Wendy J. Fantl, Sean C. Bendall, Edgar G. Engleman, and Garry P. Nolan. An interactive reference framework for modeling a dynamic immune system. *Science*, 2015.

[148] Albert G. Tsai, David R. Glass, Marisa Juntilla, Felix J. Hartmann, Jean S. Oak, Sebastian Fernandez-Pol, Robert S. Ohgami, and Sean C. Bendall. Multiplexed single-cell morphometry for hematopathology diagnostics. *Nature Medicine*, 2020.

[149] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 2019.

[150] Lin Lin, Greg Finak, Kevin Ushey, Chetan Seshadri, Thomas R. Hawn, Nicole Frahm, Thomas J. Scriba, Hassan Mahomed, Willem Hanekom, Pierre Alexandre Bart, Giuseppe Pantaleo, Georgia D. Tomaras, Supachai Rerks-Ngarm, Jaranit Kaewkungwal, Sorachai Nitayaphan, Punnee Pitisuttithum, Nelson L. Michael, Jerome H. Kim, Mer-

lin L. Robb, Robert J. O'Connell, Nicos Karasavvas, Peter Gilbert, Stephen C. De Rosa, M. Juliana McElrath, and Raphael Gottardo. COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nature Biotechnology*, 2015.

[151] Karthik Shekhar, Petter Brodin, Mark M. Davis, and Arup K. Chakraborty. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[152] Vedant Sachdeva, Thierry Mora, Aleksandra M. Walczak, and Stephanie Palmer. Optimal prediction with resource constraints using the information bottleneck. *bioRxiv*, 2020.

[153] J. Scott Shaffer, Penny L. Moore, Mehran Kardar, and Arup K. Chakraborty. Optimal immunization cocktails can promote induction of broadly neutralizing Abs against highly mutable pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[154] Shenshen Wang, Jordi Mata-Fink, Barry Kriegsman, Melissa Hanson, Darrell J. Irvine, Herman N. Eisen, Dennis R. Burton, K. Dane Wittrup, Mehran Kardar, and Arup K. Chakraborty. Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell*, 2015.

[155] Miloš KneŽević, Hongda Jiang, and Shenshen Wang. Active Tuning of Synaptic Patterns Enhances Immune Discrimination. *Physical Review Letters*, 2018.

[156] Andreas Carlson and L. Mahadevan. Elastohydrodynamics and Kinetics of Protein Patterning in the Immunological Synapse. *PLoS Computational Biology*, 2015.

[157] Thierry Mora, Aleksandra M. Walczak, William Bialek, and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 2010.

[158] Kristian Davidsen, Branden J. Olson, William S. DeWitt, Jean Feng, Elias Harkins, Philip Bradley, and Frederick A. Matsen. Deep generative models for T cell receptor protein sequences. *eLife*, 2019.

[159] Quentin Marcou, Thierry Mora, and Aleksandra M. Walczak. High-throughput immune repertoire analysis with IGoR. *Nature Communications*, 2018.

[160] Alon Oyler-Yaniv, Jennifer Oyler-Yaniv, Benjamin M. Whitlock, Zhiduo Liu, Ronald N. Germain, Morgan Huse, Grégoire Altan-Bonnet, and Oleg Krichevsky. A Tunable Diffusion-Consumption Mechanism of Cytokine Propagation Enables Plasticity in Cell-to-Cell Communication in the Immune System. *Immunity*, 2017.

[161] Raymond Cheong, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko. Information transduction capacity of noisy biochemical signaling networks.

Science, 2011.

[162] Kenneth K.H. Ng, Mary A. Yui, Arnav Mehta, Sharmayne Siu, Blythe Irwin, Shirley Pease, Satoshi Hirose, Michael B. Elowitz, Ellen V. Rothenberg, and Hao Yuan Kueh. A stochastic epigenetic switch controls the dynamics of T-cell lineage commitment. *eLife*, 2018.

[163] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 1952.

[164] John Carew Eccles. *The Physiology of Synapses*. 1964.

[165] Katharina Schwarze, James Buchanan, Jilles M. Fermont, Helene Dreau, Mark W. Tilley, John M. Taylor, Pavlos Antoniou, Samantha J.L. Knight, Carme Camps, Melissa M. Pentony, Erika M. Kvikstad, Steve Harris, Niko Popitsch, Alistair T. Pagnamenta, Anna Schuh, Jenny C. Taylor, and Sarah Wordsworth. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, 2020.

[166] Harlan Robins. Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, 2013.

[167] Neha Chaudhary and Duane R. Wesemann. Analyzing immunoglobulin repertoires. *Frontiers in Immunology*, 2018.

[168] Claire Marks and Charlotte M. Deane. How repertoire data are changing antibody science. *Journal of Biological Chemistry*, 2020.

[169] Christoph Schultheiß, Lisa Paschold, Donjete Simnica, Malte Mohme, Edith Willscher, Lisa von Wenserski, Rebekka Scholz, Imke Wieters, Christine Dahlke, Eva Tolosa, Daniel G. Sedding, Sandra Ciesek, Marylyn Addo, and Mascha Binder. Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. *Immunity*, 2020.

[170] Xavier Brochet, Marie Paule Lefranc, and Véronique Giudicelli. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic acids research*, 2008.

[171] Christopher T. Boughter, Marta T. Borowska, Jenna J. Guthmiller, Albert Bendelac, Patrick C. Wilson, Benoit Roux, and Erin J. Adams. Biochemical patterns of antibody polyreactivity revealed through a bioinformatics-based analysis of CDR loops. *eLife*, 2020.

[172] Zhihua Qiao, Lan Zhou, and Jianhua Z. Huang. Sparse linear discriminant analysis with applications to high dimensional low sample size data. *IAENG International Journal of Applied Mathematics*, 2009.

[173] Claude E. Shannon. The Mathematical Theory of Communication. *The Bell System Technical Journal*, 1948.

[174] Ramón Román-Roldán, Pedro Bernaola-Galván, and José L. Oliver. Application of information theory to DNA sequence analysis: A review. *Pattern Recognition*, 1996.

[175] Susana Vinga. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 2014.

[176] Anand Murugan, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[177] Erin J. Adams and Peter Parham. Species-specific evolution of MHC class I genes in the higher primates. *Immunological Reviews*, 2001.

[178] S. B. Piertney and M. K. Oliver. The evolutionary ecology of the major histocompatibility complex. *Heredity*, 2006.

[179] Natalie A. Borg, Kwok S. Wun, Lars Kjer-Nielsen, Matthew C.J. Wilce, Daniel G. Pellicci, Ruide Koh, Gurdyal S. Besra, Mandvi Bharadwaj, Dale I. Godfrey, James McCluskey, and Jamie Rossjohn. CD1d-lipid-antigen recognition by the semi-invariant NKT T-cell receptor. *Nature*, 2007.

[180] Adrienne M. Luoma, Caitlin D. Castro, Toufic Mayassi, Leslie A. Bembinster, Li Bai, Damien Picard, Brian Anderson, Louise Scharf, Jennifer E. Kung, Leah V. Sibener, Paul B. Savage, Bana Jabri, Albert Bendelac, and Erin J. Adams. Crystal Structure of Vδ1T Cell Receptor in Complex with CD1d-Sulfatide Shows MHC-like Recognition of a Self-Lipid by Human $\gamma\delta$ T Cells. *Immunity*, 2013.

[181] Erin J. Adams. Lipid presentation by human CD1 molecules and the diverse T cell populations that respond to them. *Current Opinion in Immunology*, 2014.

[182] Eric W. Sayers, Jeff Beck, J. Rodney Brister, Evan E. Bolton, Kathi Canese, Donald C. Comeau, Kathryn Funk, Anne Ketter, Sunghwan Kim, Avi Kimchi, Paul A. Kitts, Anatoliy Kuznetsov, Stacy Lathrop, Zhiyong Lu, Kelly McGarvey, Thomas L. Madden, Terence D. Murphy, Nuala O'Leary, Lon Phan, Valerie A. Schneider, Françoise Thibaud-Nissen, Bart W. Trawick, Kim D. Pruitt, and James Ostell. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 2020.

[183] Tereza Almeida, Pedro J. Esteves, Martin F. Flajnik, Yuko Ohta, and Ana Veríssimo. An Ancient, MHC-Linked, Nonclassical Class I Lineage in Cartilaginous Fish. *The Journal of Immunology*, 2020.

[184] Róisín M. McMahon, Lone Friis, Christian Siebold, Manuel A. Friese, Lars Fugger, and E. Yvonne Jones. Structure of HLA-A*0301 in complex with a peptide of proteolipid protein: Insights into the role of HLA - A alleles in susceptibility to multiple sclerosis. *Acta Crystallographica Section D: Biological Crystallography*, 2011.

[185] Michael Koch, Victoria S. Stronge, Dawn Shepherd, Stephan D. Gadola, Bini Mathew, Gerd Ritter, Alan R. Fersht, Gurdyal S. Besra, Richard R. Schmidt, E. Yvonne Jones, and Vincenzo Cerundolo. The crystal structure of human CD1d with and without α-galactosylceramide. *Nature Immunology*, 2005.

[186] Lawrence A. Kelley, Stefans Mezulis, Christopher M. Yates, Mark N. Wass, and Michael J.E. Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 2015.

[187] Fabian Paul, Christoph Wehmeyer, Esam T. Abualrous, Hao Wu, Michael D. Crabtree, Johannes Schöneberg, Jane Clarke, Christian Freund, Thomas R. Weikl, and Frank Noé. Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nature Communications*, 2017.

[188] Yilin Meng, Cen Gao, David K. Clawson, Shane Atwell, Marijane Russell, Michal Vieth, and Benoît Roux. Predicting the Conformational Variability of Abl Tyrosine Kinase using Molecular Dynamics Simulations and Markov State Models. *Journal of Chemical Theory and Computation*, 2018.

[189] Fabian Paul, Hao Wu, Maximilian Vossel, Bert L. De Groot, and Frank Noé. Identification of kinetic order parameters for non-equilibrium dynamics. *Journal of Chemical Physics*, 2019.

[190] Fabian Paul, Yilin Meng, and Benoît Roux. Identification of Druggable Kinase Target Conformations Using Markov Model Metastable States Analysis of apo-Abl. *Journal of Chemical Theory and Computation*, 2020.

[191] Donghyuk Suh, Sunhwan Jo, Wei Jiang, Chris Chipot, and Benoît Roux. String Method for Protein-Protein Binding Free-Energy Calculations. *Journal of Chemical Theory and Computation*, 2019.

[192] David W. Borhani and David E. Shaw. The future of molecular dynamics simulations in drug discovery. *Journal of Computer-Aided Molecular Design*, 2012.

[193] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science Advances*, 2018.

[194] John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 1935.

[195] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *Journal*

of Chemical Physics, 2001.

[196] James C. Gumbart, Benoît Roux, and Christophe Chipot. Standard binding free energies from computer simulations: What is the best strategy? *Journal of Chemical Theory and Computation*, 2013.

[197] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules: I. The method. *Journal of Computational Chemistry*, 1992.

[198] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics*, 2008.

[199] Haohao Fu, Wensheng Cai, Jérôme Hénin, Benoît Roux, and Christophe Chipot. New Coarse Variables for the Accurate Determination of Standard Binding Free Energies. *Journal of Chemical Theory and Computation*, 2017.

[200] Edina Rosta and Gerhard Hummer. Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *Journal of Chemical Theory and Computation*, 2015.

[201] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 1999.

[202] James C. Gumbart, Benoiît Roux, and Christophe Chipot. Efficient determination of protein-protein standard binding free energies from first principles. *Journal of Chemical Theory and Computation*, 2013.

[203] Eugenia V. Gurevich and Vsevolod V. Gurevich. Therapeutic potential of small molecules and engineered proteins. *Handbook of Experimental Pharmacology*, 2014.

[204] Ariel Erijman, Eran Rosenthal, and Julia M. Shifman. How structure defines affinity in protein-protein interactions. *PLoS ONE*, 2014.

[205] Leah V. Sibener, Ricardo A. Fernandes, Elizabeth M. Kolawole, Catherine B. Carbone, Fan Liu, Darren McAffee, Michael E. Birnbaum, Xinbo Yang, Laura F. Su, Wong Yu, Shen Dong, Marvin H. Gee, Kevin M. Jude, Mark M. Davis, Jay T. Groves, William A. Goddard, James R. Heath, Brian D. Evavold, Ronald D. Vale, and K. Christopher Garcia. Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding. *Cell*, 2018.

[206] Siyi Gu, Marta T. Borowska, Christopher T. Boughter, and Erin J. Adams. Butyrophilin3A proteins and Vγ9Vδ2 T cell activation. *Seminars in Cell and Developmental Biology*, 2018.

[207] J. Robert Oppenheimer and Raymond J. Seeger. The Flying Trapeze: Three Crises for Physicists. *American Journal of Physics*, 1966.

[208] Thomas Kluyver, Benjamin Ragan-kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter Development Team. Jupyter Notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016.

[209] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 2007.

[210] Erik Ziegler, Yury V. Zaytsev, Michael T. Waskom, Olga Botvinnik, Paul Hobson, John B. Cole, Yaroslav Halchenko, Stephan Hoyer, Alistair Miles, Tom Augspurger, Tal Yarkoni, Tobias Megies, Luis Pedro Coelho, Daniel Wehner, and Michael Waskom. seaborn: v0.5.0. *zenodo*, 2014.

[211] Wes McKinney and PyData Development Team. Pandas - Powerful Python Data Analysis Toolkit. *Pandas - Powerful Python Data Analysis Toolkit*, 2015.

[212] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 2020.

[213] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand

Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[214] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. 1997.

[215] Sunhwan Jo, Taehoon Kim, and Wonpil Im. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS ONE*, 2007.

[216] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 2005.

[217] Mark S. Friedrichs, Peter Eastman, Vishal Vaidyanathan, Mike Houston, Scott Legrand, Adam L. Beberg, Daniel L. Ensign, Christopher M. Bruns, and Vijay S. Pande. Accelerating molecular dynamic simulation on graphics processing units. *Journal of Computational Chemistry*, 2009.

[218] Andreas W. Götz, Mark J. Williamson, Dong Xu, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. generalized born. *Journal of Chemical Theory and Computation*, 2012.

[219] Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of Chemical Theory and Computation*, 2015.

[220] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 1985.

[221] Shuichi Nosé and Michael L. Klein. A study of solid and liquid carbon tetrafluoride using the constant pressure molecular dynamics technique. *The Journal of Chemical Physics*, 1983.

[222] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, 1993.

[223] Jean Paul Ryckaert, Giovanni Ciccotti, and Herman J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 1977.

[224] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 1996.

[225] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang,

Thomas J. Lane, and Vijay S. Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, 2015.

[226] Matthias Garten, Sophie Aimon, Patricia Bassereau, and Gilman E.S. Toombes. Reconstitution of a transmembrane protein, the voltage-gated ion channel, KvAP, into giant unilamellar vesicles for microscopy and patch clamp studies. *Journal of Visualized Experiments*, 2015.