1-1-2013

Navigating the Extremes of Biological Datasets for Reliable Structural Inference and Design

Brett Thomas Hannigan
University of Pennsylvania, brettth@mail.med.upenn.edu

Navigating the Extremes of Biological Datasets for Reliable Structural Inference and Design

Abstract
Structural biologists currently confront serious challenges in the effective interpretation of experimental data due to two contradictory situations: a severe lack of structural data for certain classes of proteins, and an incredible abundance of data for other classes. The challenge with small data sets is how to extract sufficient information to draw meaningful conclusions, while the challenge with large data sets is how to curate, categorize, and search the data to allow for its meaningful interpretation and application to scientific problems. Here, we develop computational strategies to address both sparse and abundant data sets. In the category of sparse data sets, we focus our attention on the problem of transmembrane (TM) protein structure determination. As X-ray crystallography and NMR data is notoriously difficult to obtain for TM proteins, we develop a novel algorithm which uses low-resolution data from protein cross-linking or scanning mutagenesis studies to produce models of TM helix oligomers and show that our method produces models with an accuracy on par with X-ray crystallography or NMR for a test set of known TM proteins. Turning to instances of data abundance, we examine how to mine the vast stores of protein structural data in the Protein Data Bank (PDB) to aid in the design of proteins with novel binding properties. We show how the identification of an anion binding motif in an antibody structure allowed us to develop a phosphate binding module that can be used to produce novel antibodies to phosphorylated peptides - creating antibodies to 7 novel phospho-peptides to illustrate the utility of our approach. We then describe a general strategy for designing binders to a target protein epitope based upon recapitulating protein interaction geometries which are over-represented in the PDB. We follow this by using data describing the transition probabilities of amino acids to develop a novel set of degenerate codons to create more efficient gene libraries. We conclude by describing a novel, real-time, all-atom structural search engine, giving researchers the ability to quickly search known protein structures for a motif of interest and providing a new interactive paradigm of protein design.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
William F. DeGrado

Second Advisor
Jeff G. Saven

Keywords
computational biology, degenerate codons, gene libraries, protein design, protein engineering, structural search

Subject Categories
Bioinformatics | Biophysics

NAVIGATING THE EXTREMES OF BIOLOGICAL DATASETS FOR RELIABLE STRUCTURAL INFERENCE AND DESIGN

Brett T. Hannigan

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

2013

Supervisor of Dissertation

_____

William F. DeGrado

Professor of Pharmaceutical Chemistry, University of California, San Francisco

Graduate Group Chairperson

_____

Maja Bucan, Professor of Genetics, University of Pennsylvania

Dissertation Committee

William F. DeGrado, Professor of Pharmaceutical Chemistry, University of California San Francisco
Jeffery G. Saven, Associate Professor of Chemistry, University of Pennsylvania
Kathryn M. Ferguson, Associate Professor of Physiology, University of Pennsylvania
Shane T. Jensen, Associate Professor of Statistics, University of Pennsylvania
Roland L. Dunbrack, Jr., Professor, Institute for Cancer Research, Fox Chase Cancer Center
Ahmet Sacan, Assistant Professor, School of Biomedical Engineering, Science & Health Systems
Drexel University

NAVIGATING THE EXTREMES OF BIOLOGICAL DATASETS FOR RELIABLE STRUCTURAL INFERENCE AND DESIGN

To my family.

ACKNOWLEDGMENT

No man is an island,

Entire of itself,

Every man is a piece of the continent,

A part of the main.

-John Donne

I'd like to begin by thanking my mentor and advisor, Bill DeGrado. I came to graduate school without a fully formed idea of what I'd like to study – only the somewhat vague notion that I wanted to put my background in computers and computation to use studying problems of a biological nature, and if possible, make the world a little better place. Bill showed me the exciting ways computationalists contribute to the field of structural biology and had a million ideas for projects I could pursue. He is a quintessential polymath, equally at home discussing chemical synthesis, biologic pathways, and methods for structure minimization. Importantly for me, he was always eager to explain the details of any topic in which my knowledge was deficient, and I have the distinct impression that it is in these teaching moments that he is happiest. Moreover, the work he chooses to focus on in his lab holds great promise to make the world a better place – from designing new drugs to combat evolving flu strains, to developing a vaccine for HIV.

One way to recognize a great leader is to survey with whom he chooses to surround himself. Bill's lab was filled with such excellent scientists that even if you were not to know him, you'd have to conclude that Bill must have serious talent. I'd like to especially thank Gevorg Grigoryan and Jason Donald, two computational post-docs who were always eager to discuss research in my first few years as a DeGrado lab member. Cinque Soto, another outstanding post-doc, took a special interest in mentoring me, making sure I always thought about the big picture of the science we were working on and was the driving force behind completing the paper that makes up chapter 2 of this thesis. I'd also like to thank Ilan Samish for sharing his vast knowledge of the literature and providing context for many of my early projects. Dan Kulp was a

PhD student when I joined the lab, got me started on my first project, and continued to provide useful guidance long after he graduated. Kathleen Molnar and Chaim Schramm are both very talented and provided great advice over the years, but more importantly, they were also great running partners. I had the great fortune of working with Michelle McCully and Gözde Ulas on an incredibly interesting HIV project. This project allowed Michelle and me to become “expert” molecular biologists together and our experiments were some of the most fun I had during my time in the lab. Both Jun Wang and Hyunil Jo were very generous with their time, teaching me all that I know with regards to peptide synthesis and purification, and generally making sure I was safe in the chemistry lab. Nate Joh and Jenny Hu were always willing to help me set up experiments and Yibing Wu walked me through NMR experiments a number of times. Lisa Span helped me with my first forays into protein expression. Paul Billings was always eager to help troubleshoot protein expression issues. Bruk Mensa helped immensely whenever I had questions about molecular biology, and also provided constant musical entertainment throughout the lab. Zac Kornberg will someday make a great doctor, but until then, I’m glad he chose to spend time in our lab. Gabriel Gonzalez had an infectiously positive attitude in the lab and was a generous collaborator. Shaoqing Zhang was a great friend and motivator. I’ll never forget having the feeling of someone standing over me, turning around, and being greeted with an enthusiastic “Read anything interesting?” His devotion to science is inspiring. Thanks also to Jessica Thomaston, Manasi Bhate, Mimi Nick, and Leo Gendelev for listening to my presentations and always providing useful feedback.

I’d also like to take the time to thank the Jim Wells lab for allowing me to spend time in their lab learning phage display techniques. J.T. Koerber graciously devoted significant time to helping me put together the phage display experiments for my hemagglutinin binder project and was a fantastic teacher.

In 2011 Bill moved our lab from Penn to UCSF. When I was unable to make the move initially, Dr. Casim Sarkar generously offered me space in his lab at Penn. During my time in his

lab, Casim took the time to meet regularly with me, talk science, and offer advice. He treated me with the same care as a full member of his lab, and for this I am very grateful.

I'd like to thank my graduate group, GCB, and especially our chair, Maja Bucan, for giving me the opportunity to pursue my studies at Penn. I'd also like to thank Dr. Harold Riethman for taking the time to talk with me and providing me with very useful advice upon my acceptance to the program, Dr. Junhyong Kim for allowing me to T.A. his Introduction to Computational Biology class, and Dr. Carlo Maley for his friendship.

Graduate school can be very stressful and therefore I think it is important to have an outside interest that can act as a release. For me, that is running. I'd like to thank the folks at Pretzel City Sports who put on excellent races throughout the year all over the Delaware Valley.

I'd be remiss if I failed to mention the incredible debt of gratitude I owe to my parents Tom and Ann and my sister Bonnie. No person could ask for a more loving and encouraging family.

Thanks to my two sons. Patrick, it's been the most joyous experience of my life to watch you grow up over the past two and a half years. I'm so excited to see what a curious, generous, and funny personality you have. And Wesley, you're only 3 weeks so I'll excuse the excessive crying, but I can tell that you will be a fun and precocious child.

And finally thanks so much to my wife, Monique, for supporting me throughout graduate school. First by agreeing to move clear across the country just so I could go back to school and make almost no money for 6+ years and then for quitting your awesome job to move back across the country when our lab moved. Thanks for keeping me upbeat when I was feeling low. Thanks for being a great partner.

ABSTRACT

NAVIGATING THE EXTREMES OF BIOLOGICAL DATASETS FOR RELIABLE STRUCTURAL INFERENCE AND DESIGN

Brett T. Hannigan

William F. DeGrado

Structural biologists currently confront serious challenges in the effective interpretation of experimental data due to two contradictory situations: a severe lack of structural data for certain classes of proteins, and an incredible abundance of data for other classes. The challenge with small data sets is how to extract sufficient information to draw meaningful conclusions, while the challenge with large data sets is how to curate, categorize, and search the data to allow for its meaningful interpretation and application to scientific problems. Here, we develop computational strategies to address both sparse and abundant data sets. In the category of sparse data sets, we focus our attention on the problem of transmembrane (TM) protein structure determination. As X-ray crystallography and NMR data is notoriously difficult to obtain for TM proteins, we develop a novel algorithm which uses low-resolution data from protein cross-linking or scanning mutagenesis studies to produce models of TM helix oligomers and show that our method produces models with an accuracy on par with X-ray crystallography or NMR for a test set of known TM proteins. Turning to instances of data abundance, we examine how to mine the vast stores of protein structural data in the Protein Data Bank (PDB) to aid in the design of proteins with novel binding properties. We show how the identification of an anion binding motif in an antibody structure allowed us to develop a phosphate binding module that can be used to produce novel antibodies to phosphorylated peptides – creating

antibodies to 7 novel phospho-peptides to illustrate the utility of our approach. We then describe a general strategy for designing binders to a target protein epitope based upon recapitulating protein interaction geometries which are over-represented in the PDB. We follow this by using data describing the transition probabilities of amino acids to develop a novel set of degenerate codons to create more efficient gene libraries. We conclude by describing a novel, real-time, all-atom structural search engine, giving researchers the ability to quickly search known protein structures for a motif of interest and providing a new interactive paradigm of protein design.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1

1.1 Introduction

Knowledge of protein structure is essential in our quest to understand the complex signaling and interaction networks that make life possible. The past two decades have seen an explosion in the number of protein structures solved to atomic level, with only around 500 such structures deposited in the Protein Data Bank (PDB) in 1990 to nearly 100,000 today (1). This amazing growth belies the fact that not all categories of proteins have been solved with equal success. In particular, membrane proteins make up fewer than 2% of solved structures in the PDB (2) despite comprising an estimated 25% of the human proteome (3). This discrepancy is largely due to the difficulty in obtaining atomic level data for proteins embedded in their native lipid environment. To compensate for this difficulty, researchers have developed a number of experimental techniques to obtain low-resolution data on membrane protein structure, including the TOXCAT assay (4), the ToxR assay (5), reversal potential assay (6), cysteine cross-linking (7), and scanning mutagenesis studies (8). The data obtained from these experiments is extremely sparse, typically on the order of a single data point per residue. Despite this paucity, in chapter 2, we describe a computational approach which uses this sparse data to infer the structure of homo-oligomeric, transmembrane proteins to an accuracy rivaling that of X-ray crystallography and NMR.

The wealth of protein structural data that is now available carries with it the potential to inform efforts in designing novel protein-protein interactions. Through extensive mining of protein structures, countless groups have identified and categorized structural motifs that are frequently involved in various protein interactions, from the GxxxG motif involved in transmembrane helix association (9) to the leucine zipper interaction motif found in many DNA

binding proteins (10). In chapter 3, we leverage this data to engineer an antibody which binds to novel phosphorylated peptides. Whereas, traditionally, phospho-specific antibodies have been generated through immunization (11), we developed a rational, structure-based approach paired with high-throughput screening. We exploited a previously identified structural motif called a "nest" (12) which forms a cationic hole and reasoned that the phosphate group of a phosphorylated peptide might be a perfect anion to fill such a hole. Scanning through the set of solved antibody/peptide structures, we found an example of an antibody in which one of the complementarity determining regions (CDR) involved in peptide binding formed a perfect nest. In this structure, the nest was involved in an interaction with the carboxyl group of an aspartic acid residue on the peptide. Using a technique called phage display (13), we selected mutants based on the original antibody that preferentially bound peptides whose aspartic acid was replaced with a phosphorylated residue (phospho-serine, phospho-threonine, or phospho-tyrosine.) More impressively, we were then able to use this modified, phospho-specific motif as a "phosphate-binder" module and isolated 51 phospho-specific antibodies against seven different phosphorylated peptides unrelated to the original peptide. This technique offers the promise of a much more efficient means of generating antibodies to recognize specific post-translationally modified peptides and highlights the power of leveraging our knowledge of known structural binding motifs to engineer new interactions.

In chapter 4 we extend the idea of mining structural data for protein interacting motifs in order to design a binder to the influenza fusion protein hemagglutinin. A recent report details the successful design of two protein binders to hemagglutinin with nanomolar affinity (14). In this work, the designers first docked individual amino acids to a solved structure of hemagglutinin in order to identify "hot-spot" residues that contribute a large fraction of the

overall binding energy. Once two or three of such residues were identified, the authors searched through a database of small, easily expressible proteins which they could use as scaffolds to hold these hot-spot residues. These proteins were in turn redesigned to accommodate the "hot-spot" residues and provide a level of complementarity to the hemagglutinin molecule. Eighty-eight designs were expressed and tested for binding activity. Two designs were found to bind, and after a round of affinity maturation, both designs produced variants with dissociation constants in the single-digit nanomolar range. We hypothesize that an approach which explicitly attempts to create binding interfaces that mimic those found frequently in nature would have a higher success rate. To that end, we developed a design methodology based upon the concept of "designability" (15), the idea that out of the vast ensemble of possible packing arrangements of protein secondary structure, only a limited subset is ever observed, and some of those arrangements can accommodate a wide variety of amino acid sequences. We used computational techniques to search a non-redundant database of known protein packing arrangements to identify designable motifs which were a good match to our hemagglutinin epitope of interest. These designable motifs provide a scaffold upon which we can then computationally design amino acids to drive binding to hemagglutinin. We apply our approach to design short helical peptides to bind to hemagglutinin, experimentally characterize their behavior, and attempt to isolate modified versions with enhanced binding characteristics through phage display. Although our first-generation peptide designs do not bind at levels detectable by our assays, we are able to draw valuable conclusions regarding the difficulty of designing peptide binders and propose a modified protocol to provide a protein scaffold to accommodate the proposed designable motif.

As shown in chapters 3 and 4, while the use of large sets of structural data can help in the design of novel protein interactions, to date our computational models are not robust enough to produce high-affinity designs directly. Typically, designs with lower binding affinity are produced and then used as starting points for affinity maturation in gene library experiments such as phage display. In these experiments, a library is produced by introducing mutations to the initial design, and variants are screened for enhanced binding activity through high-throughput assays. One popular method of introducing mutations is through the use of degenerate codons (16), essentially mixtures of nucleic acids which form defined amino acid distributions. However, while widely used, traditional degenerate codons have a number of shortcomings. First, many experiments will mutate more residues than can be exhaustively sampled in the library. Consequently, shaping what regions of sequence space are explored by the library can dramatically affect how likely it is that improved designs will be found. Traditional degenerate codons sample space without regards to what amino acid was present in the initial, albeit weak, design. A more efficient approach would sample the original amino acid with a higher frequency, as it is already known that that amino acid is at least compatible with binding. Second, sequencing projects have provided an incredible source of data on the amino acid transition probabilities seen during evolution. It would be useful to use these probabilities to direct the mutations introduced in gene libraries rather than settle for an entirely random approach. Finally, many degenerate codons introduce stop codons at significant levels. Any decrease in the probability of introducing a stop codon has the potential to greatly increase the effective size of the gene library, and thus significantly improve the likelihood of finding improved designs. In chapter 5 we describe the development of a novel algorithm that produces four mixtures of nucleotides to form “Super Codons” which address each of these

deficiencies. The use of these mixtures in gene library experiments should focus the sampling of sequence space to regions closer to the original design, increasing the percentage of mutants that still fold properly and retain some of the initial binding interactions. Additionally, sequence space will be explored more intelligently by explicitly making use of our knowledge of amino acid substitution rates seen through evolution and by decreasing the introduction of stop codons compared to the most popular degenerate codon. Gene libraries offer the incredible potential to find enhanced binders from a pool of billions, but will only be successful if enhanced binders make it into the pool in the first place. By intelligently directing how the pool samples sequence space around the initial design, we will greatly improve our chances of success.

Finally, one of the difficulties that comes with an incredible abundance of data is how to efficiently search it and retrieve only that information which is relevant to the query of interest. Imagine the internet without Google. A vast sea of data would be present with almost every fact known to humankind, and yet there would be no practical way to make sense of it or to search through it. The utility of the internet would be dramatically curtailed. In many ways, this is the present situation regarding the immense store of protein structural data we have amassed. As shown in chapters 3 and 4 we are able to mine structural data to improve our efforts in protein design, but the search tools currently available have severe limitations. First, the search algorithms focus only on the backbone structure of proteins, while neglecting the details of side-chain interactions. As it is the side-chains which are typically involved in protein packing, binding interactions, and catalysis, this lack of searchability is a significant oversight. Secondly, the structural search tools developed to date lack the ability to return results in real-time, severely impeding the natural search-design-repeat feedback cycle and slowing design efforts. To address these deficiencies, in chapter 6 we create a novel, all-atom, real-time

structural search algorithm called Suns. Suns uses the popular molecular-visualization package PyMOL (17) to allow the user to select structural motifs to use as a query to our database of non-redundant protein structures, and immediately returns results matching the given query to an arbitrary tolerance. We show how Suns can be used to design novel protein structures, find scaffolds which can accommodate given side-chain motifs, and quickly discover secondary structures which would be good candidate scaffolds for the hemagglutinin hot-spot residues discussed in chapter 4. With its near-instantaneous search capability, Suns promises to dramatically increase the utility of the structural data available, and open up new avenues for assessing a protein structure's designability.

1.2 References

1. **Yearly Growth of Total Structures.** ***RCSB PDB.*** **[Online] [Cited: 11 6, 2013.] http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100.**

2. ***Membrane Proteins of Known 3D Structure.*** **[Online] [Cited: 11 5, 2013.] http://blanco.biomol.uci.edu/mpstruc/.**

3. *Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin.* **Almén, M.S., et al.** 50: BMC Biology, 2009, Vol. 7. doi:10.1186/1741-7007-7-50

4. *TOXCAT: A measure of transmembrane helix association in a biological membrane.* **Russ, W.P. and Engelman, D.M.** 3: Proceedings of the National Academy of Sciences, 1999, Vol. 96. pp 863-868.

5. *Computer simulations and modeling-assisted ToxR screening in deciphering 3D structures of transmembrane α-helical dimers: ephrin receptor A1.* **Volynsky, P.E., et al.** s.l. : Physical Biology, 2010, Vol. 7. doi: 10.1088/1478-3975/7/1/016014

6. *A functionally defined model for the M2 proton channel of influenza A virus suggests a mechanism for its ion selectivity.* **Pinto, L.H., et al.** 21: Proceedings of the National Academy of Sciences, 1997, Vol. 94. pp 11301-11306.

7. *Detecting the conformational change of transmembrane signaling in a bacterial chemoreceptor by measuring effects on disulfide cross-linking in vivo.* **Hughson, A.G. and Hazelbauer, G.L.** s.l. : Proceedings of the National Academy of Sciences, 1996, Vol. 93. pp 11546-11551.

8. *Combinatorial alanine-scanning.* **Morrison, K.L. and Weiss, G.A.** 3: Current Opinion in Chemical Biology, 2001, Vol. 5. pp 302-307.

9. *The GxxxG motif: a framework for transmembrane helix-helix association.* **Russ, W.P. and Engelman, D.M.** 3: Journal of Molecular Biology, 2000, Vol. 296. pp 911-919.

10. *The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins.* **Landschulz, W.H., Johnson, P.F. and McKnight, S.L.** 4860: Science, 1988, Vol. 240. pp 1759-1764.

11. *Overview of the generation, validation, and application of phosphosite-specific antibodies.* **Brumbaugh, K., et al.** s.l. : Methods in Molecular Biology, 2011, Vol. 717. pp 3-43.

12. *A novel main-chain anion-binding site in proteins: the nest. A particular combination of φ,ψ values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions.* **Watson, J.D. and Milner-White, J.** 2: Journal of Molecular Biology, 2002, Vol. 315. pp 171-182.

13. *Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface.* **Smith, G.P.** 4705: Science, 1985, Vol. 228. pp 1315-1317.

14. *Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin.* **Fleishman, S.J., et al.** 6031: Science, 2011, Vol. 332. pp 816-821.

15. *The designability of protein structures.* **Helling, R., et al.** : Journal of Molecular Graphics and Modelling, 2001, Vol. 19. pp 157-167.

16. *Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.* **Cornish-Bowden, A.** 9: Nucleic Acids Research, 1985, Vol. 13. pp 3021-3030.

17. The PyMOL Molecular Graphics System, Schrödinger, LLC.

Chapter 2

A photon-free approach to transmembrane protein structure determination

2.1 Abstract

The structures of membrane proteins are generally solved using samples dissolved in micelles, bicelles, or occasionally phospholipid bilayers using X-ray diffraction or magnetic resonance. Because these are less than perfect mimics of true biological membranes, the structures are often confirmed by evaluating the effects of mutations on the properties of the protein in their native cellular environments. Low-resolution structures are also sometimes generated from the results of site-directed mutagenesis when other structural data are incomplete or not available. Here we describe a rapid and automated approach to determine structures from data on site-directed mutants for the special case of homo-oligomeric helical bundles. The method uses as input an experimental profile of the effects of mutations on some property of the protein. This profile is then interpreted by assuming that positions that have large effects on structure/function when mutated project towards the center of the oligomeric bundle. Model bundles are generated and correlation analysis is used to score which structures have inter-subunit Cβ distances between adjoining monomers that best correlate with the experimental profile. These structures are then clustered and refined using energy-based minimization methods. For a set of 10 homo-oligomeric TM protein structures ranging from dimers to pentamers, we show that our method predicts structures to within 1 to 2 Å backbone RMSD relative to X-ray and NMR structures. This level of agreement approaches the precision of NMR structures solved in different membrane mimetics.

This chapter has been published in the Journal of Molecular Biology (2011 December 9; 414(4):596-610). Dr. Cinque Soto is an equal-contributor to this work and primary author of the text. William F. DeGrado is the corresponding author.

2.2 Introduction

Helical transmembrane protein structure determination represents a significant challenge. Fewer than 2% of all experimentally determined structures deposited in the Protein Data Bank (1) (PDB) are membrane proteins, yet 20-25% of open reading frame (ORFs) from recently sequenced genomes encode for proteins that embed in the membrane(2)(3). Even with advances in conventional methods for protein structure determination such as X-ray crystallography and NMR spectroscopy, the fundamental problems of obtaining diffraction-quality crystals, protein expression and purification, and protein-size limitations still remain. Computational methods for modeling transmembrane protein structure are becoming increasingly more important if we hope to decrease the discrepancy in structural information between globular and membrane proteins.

Depending on the scientific question being asked, the laborious (and sometimes insurmountable) task of experimentally determining the structure of a membrane protein using conventional methods may not be necessary. For example, Zhu et al. (4) recently used disulfide crosslinking information to build models for the helical transmembrane (TM) dimers glycophorin A (GpA) and integrin $\alpha_{II}\beta_3$. The resulting models for GpA had a root mean square distance (RMSD) over the backbone atoms of 1 to 1.5 Å with the NMR structures. Metcalf et al.(5) used mutagenesis data and protein sequence variation to build models for the TM homo dimers GpA and BNIP3 apoptosis factor. The RMSD for the GpA model was 1.3 Å. We hypothesize that other forms of low-resolution experimental data can potentially provide sufficient information to accurately model other transmembrane protein structures. Experimental data from a variety of mutagenesis experiments are ideal for studying this possibility.

The earliest structural models by Brünger and coworkers for the TM region of GpA were based solely on the energetics of interaction between helices (6)(7). The resulting models were compared against mutagenesis data showing the disruptive effects that non-polar mutations had on GpA's ability to dimerize (8). The structural models agreed with the mutagenesis data and showed that key residues oriented toward the helical interface were sensitive to nonpolar mutations.

The approach of modeling helical TM regions using the energetics of interaction between helices has been extended to larger homo-oligomers. Phospholamban is a TM homo-pentamer that is important in calcium storage and release in cardiocytes. Mutagenesis studies (7)(9) showed that mutations of key hydrophobic residues disrupted pentamer oligomerization. A global search of conformational space revealed five low-energy helical bundles (7), only one of which was found to be in agreement with an extensive set of mutagenesis data (7)(9), and ultimately the experimentally determined structures (10)(11). This five-fold symmetrical structure has a left-handed twist; most critical residues lie at the helix/helix interface and show large interaction energies. Interestingly, the lowest energy conformer did not agree with the experimental results, indicating that energy is a necessary -- but insufficient -- criterion for assessing models.

Herezyk and Hubbard (12) used a different approach to model helical TM homo-olgimers. Using a combination of Monte Carlo/Simulated Annealing (MCSA) and Molecular Dynamics/Simulated Annealing (MDSA) along with a set of orientational restraints derived from published mutagenesis data (7)(9), Heryzk and Hubbard constructed models for GpA and phospholamban. Unlike the modeling approach of Brünger and coworkers, which made use of mutagenesis data after the model was constructed, Herezyk and Hubbard used restraints

derived from mutagenesis data in their modeling procedure. The resulting model for GpA had an RMSD to native of 0.9 Å over the backbone atoms. A comparison of the profile between interaction energy and mutagenesis data revealed an excellent level of agreement for phospholamban.

More recent approaches for modeling helix TM homo-oligomers fall into one of these two categories: modeling methods based purely on energetics (13)(14)(15)(16) and those that use some combination of energetics and low-resolution experimental data (4)(5)(17)(18). The incorporation of experimental data directly into the modeling process provides two obvious benefits. First, the experimental data corrects for inaccuracies in the force field and for approximations regarding the environmental conditions. Second, by using experimental data directly in the modeling process, the conformational space that needs to be sampled can be greatly reduced.

We have developed a novel approach for modeling helical TM homo-oligomers that incorporates a variety of low-resolution mutagenesis data directly into the modeling process. Our modeling approach consists of two phases. In the first phase we use a symmetric rigid-body search to generate an ensemble of models that is consistent with a given set of low-resolution data. In the second phase we cluster and then refine only the centroid models using the CHARMM22 force field. At the heart of our rigid-body search is a simple scoring function that restrains the conformational search by maximizing the correlation between inter-subunit Cβ distance and experimental data while minimizing steric clashes between helices. Our correlation term allows us to use a variety of low-resolution mutagenesis data without the need for scaling the data or converting the data into distance restraints[4] or angular restraints (12). We demonstrate the accuracy of our modeling approach by using a variety of low-resolution

experimental data such as mutagenesis, ToxR, TOXCAT, ion channel and crosslinking data to model the TM regions of GpA, phospholamban, M2, BM2, BNIP3 and the ephrin receptor tyrosine kinase (EphA1). The final models ranged in RMSD from 0.6 Å to 2.1 Å when compared to the native structures. This approach to modeling helical TM protein structure can be of enormous benefit when conventional methods of protein structure determination fall short.

2.3 Results

Overview of modeling protocol

Our modeling protocol can be broken down into two phases. The first phase involves rigid-body sampling using an ideal or experimentally-determined helix. The second phase involves side chain placement, clustering and refinement of the models with a molecular mechanics force field. We briefly describe the first phase here. A detailed description of the second phase can be found in the Methods. Rigid-body sampling (RBS) begins with a helix that is transformed to the global frame of reference so that the axis of the helix is coincident with the global Z-axis and its geometric center is at the origin. Four degrees of freedom are required to define the relationship between monomers in a structure with exact rotational symmetry. Here we apply two rotations and two translations to define the location of the helix in the unit cell. The individual steps in our modeling protocol are illustrated in **Figure 1**.

At the heart of our RBS method is the use of the correlation coefficient (r) to evaluate the degree to which experimental data correlates with the projection of the side-chains in the oligomer, as defined by the inter-subunit contact distance for each residue in the structure. The inter-subunit contact distance is defined here as the distance between Cβ atoms on identical residues of a homo-dimer, and this provides a quantitative measure that can be correlated with the extent of perturbation or crosslinking associated at the same position in the sequence. The

correlation coefficient is a measure of the linear relationship between two variables and ranges from a value of 1 for two perfectly correlated variables to a value of -1 for two perfectly anti-correlated variables. The correlation coefficient is used to restrain the RBS protocol by incorporating it directly into a scoring function that is used to optimize each pose (see Methods). In this study, we correlate inter-subunit Cβ distance with the degree of experimental perturbation associated with mutations or the extent of cross-linking in a Cys-scanning experiment to determine how well a given hypothetical model agrees with experimental data. The extent of Cys crosslinking and the perturbational effects of mutations generally increase with decreasing inter-subunit distance (negative correlation). However, for simplicity, we refer to all correlations as positive for structures that are in agreement with the expected experimental outcome.

We demonstrate the utility of our RBS protocol by using it in three tests. In the first test we use it to search for a set of idealized helical conformations using native inter-subunit Cβ distances as "experimental data." The second test is similar to the first test but uses a set of nine symmetric helical TM structures obtained from the PDB instead of idealized helical arrangements. It should be noted that the first and second tests are used to determine how well our search strategy works under the most ideal conditions (i.e., where experimental data correlates perfectly with inter-subunit Cβ distances). In the third and final test, we model these same nine structures using low resolution experimental data to restrain the search. The resulting ensembles of models from this test are clustered using a k-medoid clustering algorithm (19). Side chains are then added to each of the centroid models using SCAP (20) followed by all-atom refinement using the CHARMM22 force field implemented in the XPLOR-NIH package (21).

Idealized hide-and-seek test

To test the RBS protocol, we constructed a set of ten helix dimer conformations by randomly choosing values for the four search parameters (Tx, Tz, θ, and ϕ). Each set of four parameters is then used to position a 16 residue ideal poly-alanine helix in space. The symmetry mate is generated by rotating a copy of the helix 180° about the global Z axis. After construction of the ten dimers, we determined the inter-subunit Cβ distances along the length of the helices. These distances were used as simulated experimental data to restrain the rigid-body search with the goal of recapitulating the original dimer conformation. In all ten cases, the simple scoring function selects a model with an RMSD of 0.6 Å or less to the starting conformation (see **Supplementary Table 1**).

Hide-and-seek test using TM structures from the PDB

The RBS protocol can generate the native pose with high accuracy for idealized cases. A more challenging test would entail modeling actual helical structures from the PDB which may not contain idealized geometry. We repeated the hide-and-seek test on a set of nine symmetric helical TM structures from the PDB. Three of these structures are dimers, five are tetramers and one is a pentamer. For each test case, we determined the inter-subunit Cβ distances from the first two chains of the native structure. If a glycine is present along the protein sequence we computed the distance between Cα atoms. For structures solved using NMR, we use the average structure (see Methods) to obtain the native distances.

We use three separate measures of RMSD in assessing the performance of the RBS protocol on experimentally determined structures. The first measure, $RMSD_{Score}$, denotes the RMSD between the best scoring model in the ensemble and the native structure. The second measure, $RMSD_{Min}$, denotes the smallest RMSD in the ensemble. The third measure, $RMSD_{Native}$,

denotes the RMSD of the best scoring model when the native helix is used in place of the ideal helix in the rigid-body search. As shown in Table 1, all nine cases have a $RMSD_{Score}$ of 2.9 Å or less. The dimer BNIP3 gives the best results with a $RMSD_{Score}$ of 0.9 Å. The worst performing case, phospholamban, gives a $RMSD_{Score}$ of 2.9 Å. The remaining cases yield $RMSD_{Score}$ values between 1.2 and 2.0 Å. While our scoring function does not select the lowest RMSD model in the ensemble, it does perform reasonably well at generating low RMSD models (**Figure 2**). With the exception of the BM2 case, a sizable population of models with RMSDs below 1.5 Å is always generated. Producing an ensemble of models with relatively low RMSD to native is critical for two reasons. First, models that are near-native will generally yield more favorable scores in the refinement stage. Second, clustering will be more effective at assigning near-native models as centroids.

We suspected that our sampling algorithm could generate a larger population of near-native models if we introduced natural curvature into the starting helix. Superimposing an ideal helix onto the corresponding native helix gives an RMSD that is larger than 1.0 Å for GpA, BM2, and all of the M2 structures. To better assess how this deviation from ideality influences the final result, we carried out the same search using the native helix in place of the ideal helix. The resulting $RMSD_{Native}$ values are 0.6 Å or less for all cases with the exception of the M2(xtal) case (**Table 1**). However, we note the existence of models with $RMSD_{Native}$ values of 0.6 Å or less for all of the ensembles generated using a native helix (**Figure 2**).

Restrained sampling using low resolution experimental data

The first two tests show that when sufficient information between monomers is given in the form of native distances, our RBS protocol can generate models with $RMSD_{Min}$ values between 0.9-1.6 Å. However, in a practical situation, exact distance information will likely be

unavailable. Therefore, to assess the ability of the sampling protocol to perform similarly in a practical situation, we used low-resolution experimental data to restrain the search. Besides being the most stringent test thus far, given the inherent noise present in low-resolution experimental data, this test will provide a meaningful benchmark in terms of the practicality of our method. A description of the low-resolution experimental data is provided in the Methods section.

Before carrying out the search, we wanted to test our hypothesis that inter-subunit Cβ distance correlates with low-resolution experimental data. To do this we determined the correlation coefficient and the associated p values between the inter-subunit Cβ distance data obtained from each native structure and the corresponding set of experimental data (see **Supplementary Tables 2-8** for the experimental data). Phospholamban has the strongest correlation with |r|=0.91 (p=4.6E-7). The dimer GpA has roughly the same |r| value of 0.78 (with an approximate p value of 5.0-E-6) for both the crosslinking and mutagenesis data. The dimer EphA1 has |r|= 0.76 (p=3.2E-3). The M2 cases have roughly the same |r|=0.72 (with a p value of about 3.7E-4). BNIP3 and BM2 have the weakest correlations with |r| =0.44 (p=5.7E-2) and |r|=0.58 (p=4.7E-3) respectively. For all but one of the cases, the p-value for the correlation between experimental data and inter-subunit Cβ distance is less than 0.05, indicating that the correlation is unlikely due to chance. Based on the |r| values and associated p-values obtained for the native structures, it would seem that correlating inter-subunit Cβ distance with low-resolution experimental data can provide a useful filter when modeling TM homo oligomers (**Figure 3**).

Using the low-resolution experimental data to generate TM bundles, we obtained an $RMSD_{Score}$ of 2.1 Å or less for eight out ten cases (**Table 2**). BNIP3 and phospholamban are the

largest outliers with $RMSD_{Score}$ values of 3.0 Å. For BNIP3 it is not surprising that the $RMSD_{Score}$ is so large given the weak correlation between inter-subunit Cβ distance and the experimental data. For phospholamban, we noticed that the bundle radius for the top scoring model is about 1.0 Å smaller than in the native structure. A more important measure of performance of the sampling protocol is how close to native conformation our sampling can reach. Clearly, if the sampling protocol cannot generate a sufficient number of models that are close to native, it is likely that all-atom refinement will be of little value in generating good models. The $RMSD_{Min}$ value is 1.6 Å or less for 9 out of 10 cases (**Table 2**). With the exception of BM2, the sampling protocol generates ensembles with a significant fraction of models less than 2.0 Å RMSD to native (**Figure 4**). Based on these results it appears that when inter-subunit Cβ distance data correlates strongly with mutational data, rigid-body sampling alone can be used to generate reasonable starting conformations that can be further refined. However, since our scoring function is designed as a filter it may not select the most energetically favorable conformation in the ensemble of models. For this, we use a more detailed all-atom scoring function.

Refinement using XPLOR-NIH

The resolution of our simple scoring function does not capture detailed energetic interactions such as van der Waals packing and Coulombic interactions. These interactions are important for obtaining optimal packing between helices. To capture these important interactions, we first cluster the ensemble of models generated using our RBS protocol, add side chains to all centroids and then subject them to all-atom refinement using the CHARMM22 force field in XPLOR-NIH (21). The most favorable scoring model according to XPLOR is deemed our best prediction.

Refinement of the centroids gives an $RMSD_{Score}$ of 2.1 Å or better for all ten cases (**Figure 5**). For the dimers GpA, GpA(Crosslinking), EphA1 and BNIP3, the $RMSD_{Score}$ is 1.4 Å or less. Results for larger homo oligomeric states are equally as impressive with $RMSD_{Score}$ ranging in value from 1.1-2.1 Å. Given the spread in RMSD values between individual models in the native NMR ensembles, which can be as large as 0.9 Å for some of the structures considered here, our results would indicate that the RBS protocol coupled to clustering and refinement with XPLOR-NIH has the potential to generate models comparable in accuracy to those obtained using medium-resolution NMR. The importance of using a detailed all-atom scoring function is clearly illustrated for the case of BNIP3. Using our simple scoring function to select a model from the ensemble will give an RMSD to native of 3.0 Å. If we refine all of the models in the ensemble and then select the most favorable scoring model according to CHARMM22, we obtain an RMSD to native of 0.6 Å. Clearly, refining the entire ensemble of 1,000 models would be a time consuming task and so we cluster the ensemble of models first and then refine only the centroids. Using this approach, we also obtain a model with an RMSD to native of 0.6 Å but do so in a fraction of the time it would take to refine the entire ensemble of models. The r value between the experimental data and the inter-subunit Cβ distance for the refined models either remained the same or improved when compared with the corresponding value for native.

As a control, we applied the same XPLOR-NIH refinement protocol to all the native structures. This involved refinement of all the individual models in each NMR ensemble and not the average model. We expect the experimentally determined structures after refinement to have scores that are similar to or more favorable than the scores of our centroid models. We observe this trend for all cases with exception of BM2 (**Figure 6**). We find that the refined native models for BM2 are about 100 XPLOR energy units less favorable than our best scoring model.

This seems to imply that the native BM2 bundle may not be tightly packed which ultimately leads to a less favorable van der Waals score. For most cases, refinement with XPLOR-NIH does not significantly perturb the native structure. The RMSD between the unrefined and refined native models is on average less than 1.0 Å (represented as blue circles **Figure 6**). For phospholamban and BM2, refinement perturbs the native conformation to a larger extent. In particular, the RMSD after refinement of the native BM2 ensemble resulted in two models having RMSDs larger than 1.7 Å.

2.4 Discussion

We have presented a method for modeling helical TM homo-oligomers that uses a rotationally symmetric rigid-body search followed by clustering and energy refinement using the CHARMM22 force field in XPLOR-NIH. At the heart of our modeling procedure is a simple scoring function composed of a VDW clash term and a correlation coefficient between mutational data and inter-subunit Cβ distance. The simple scoring function is optimized to obtain maximal agreement with experimental data while avoiding clashes between helices. The novelty of our method is in its ability to directly restrain the search using low-resolution experimental data. This prevents the search from needlessly meandering through space and focuses the sampling to give the best agreement with experimental data.

Our method performs best when the experimental data correlate with $|r| > 0.5$ with the native inter-subunit Cβ distance. In these cases, the rigid-body search does a reasonable job at generating near-native backbone conformations. As the correlation becomes weaker, so does the structural similarity between the native structure and the best scoring model. The combination of clustering, all-atom refinement and ranking with the XPLOR-NIH scoring function

improves the RMSD value to native. In seven of the ten cases, the RMSD to native is 1.6 Å or less.

Modeling TM homo-dimers

As a prerequisite for addressing the general TM homo-oligomer problem, we first applied our modeling approach to the homo-dimer GpA using two sets of mutational data. One set of data is from a fairly recent study and is comprised of crosslinking efficiency (4). Another data set consists of dimer disruption data and has been used extensively by others to propose different methods for modeling the TM region of GpA (5)(12)(13)(14). Using either a combination of energetics and restraints derived from mutational data or using energetics alone, all of these methods generate models for the TM region of GpA with RMSDs to native in the range of 0.7-1.5 Å. Using either set of low-resolution experimental data, our modeling approach achieves a similar level of accuracy for GpA.

Earlier work in our group made use of a Monte Carlo-simulated annealing (MCSA) protocol to propose a model for the TM region of BNIP3 (5). The MCSA method used two energy terms that would penalize both neutral and disruptive mutations. The method we propose here is different in two ways. First, we do not use a stochastic approach for sampling conformational space. Second, the present method does not rely solely on the energy to decide on the plausibility of a model, but instead also relies on how well the inter-subunit Cβ distance correlates with mutational data. While both methods manage to accurately model the backbone of BNIP3, only the MCSA protocol correctly models the hydrogen bond between Nε2 of HIS 173 and Oγ from SER 172 reported by Sulistijo and Mackenzie (22). Since our refinement protocol in XPLOR-NIH does not incorporate side chain rotamer sampling, we could not optimize detailed hydrogen bond interactions between side chains. This prompted us to see if we could

model this hydrogen bond by simply changing the rotameric state of HIS 173 and SER 172 in our best scoring model. Changing the rotameric states results in a new model that scores better than our original model. This suggests that the hydrogen bond may not be absolutely necessary for dimerization of the helices (our best refined model did not have this hydrogen bond), but if formed produces a slightly more stable complex conferring specificity to the dimer as pointed out in the recent work of Lawrie et al (23).

Modeling larger TM homo-oligomeric complexes

Our modeling protocol performed well on larger TM homo-oligomeric complexes. The largest complex we considered is the pentamer phospholamban. Similar to the case for GpA, the mutagenesis data for phospholamban has been used extensively in proposing a model for the TM region (7)(12)(14). It is difficult to compare our results directly with earlier studies since they were carried out before publication of the NMR structure for phospholamban. However, a plot of the interhelical van der Waals energy per residue for phospholamban reveals a similar periodic pattern observed in plots from earlier studies (**Supplementary Figure 1**). A salient feature of using inter-subunit Cβ distance over interaction energy when constructing a profile is that the former descriptor is less sensitive to force field effects. We note that our model for phospholamban has a smaller radius than what is seen in the NMR structure. However, since our modeling protocol does not account for the membrane environment or make use of experimentally derived distance information (i.e., inter-monomer NOEs), the effect of the non-bonded forces from the molecular mechanics force field dominate resulting in tightly packed helices.

We also applied our modeling approach to the influenza proton transporters M2 and BM2. Our modeling protocol generates models for M2 with an RMSD of 1.7 Å to the high-

resolution X-ray structure. When compared to the NMR model of M2, our protocol achieves an RMSD of about 1.0 Å. Our automated method provides predictions for M2 that are better than earlier predictions that relied heavily on the expertise and intuition of the investigators (24). We also applied our modeling protocol to the recent solid state NMR structures of Sharma et al. (25) and Cady et al. (26). Our best scoring models have an RMSD of 1.8 Å and 1.6 Å respectively to these solid state structures. As a point of comparison, the NMR (solution and solid state) and the high-resolution X-ray structures show a spread in RMSD between 0.8-1.6 Å.

In an earlier study, we also made use of correlation analysis in modeling the BM2 proton transporter (18). In our previous approach we adopted a less efficient method that included the generation of a large ensemble of sterically feasible helical bundles (both ideal and coiled helices). The ensemble was scored using the correlation coefficient between the pertubility index (PI) and an estimate of the phase angle for the helix. The surviving models were subjected to refinement and then clustered. Two out of eight proposed models from our earlier study are within 1.0 Å of our current best scoring model. It should be noted that all of the models from our previous study exhibit a weaker correlation with the experimental data than the model we propose here. The current study along with our earlier study show the generality of the use of the correlation approach in modeling TM homo-oligomers; different geometric descriptors between helices can be used in modeling TM homo-oligomers.

Two clear strengths with our modeling protocol are speed (~8 minutes on a single 2.40 GHz processor) and the ability to use data directly from experiments conducted in native cellular membranes. This is in contrast to previous methods which often require the conversion of experimental data into distance restraints (4)(27), angular restraints (12), or pseudo-energy terms (5). A potential downside of these approaches is their reliance on setting thresholds *a*

priori. In contrast, we use experimental data directly to correlate against geometric descriptors between helices. We feel this makes for a simpler protocol and allows us to avoid choosing "optimal" values through an intermediate training step. Moreover, since our method relies heavily on the correlation value, data from a variety of different experimental contexts can be used without the need for scaling. Other approaches need to determine different thresholds and penalty functions for different sets of experimental data, which can make them difficult to apply.

One potential drawback of our modeling protocol is the requirement of both neutral and destabilizing mutations. If a mutagenesis experiment is carried out only on residues at the dimer interface, our correlation approach will fail due to its reliance upon detectable differences between residues close to the dimer interface and those farther away. Put another way, if all the values for a particular mutagenesis experiment are identical the correlation value would be undefined since the difference between each experimental value would be identical to the mean value. Alternative approaches do not have the same constraint. However, we anticipate that most mutagenesis experiments would involve mutations at a number of consecutive residues to determine which residues are located at the interface.

A second potential drawback of our method is the use of an "ideal" helix during the rigid-body search. This drawback has been discussed by Bowie and coworkers (14) who note that experimentally determined helices can contain large deviations from ideal geometry that result in significant kinks or curvature. The importance of accounting for curvature as seen in experimentally determined helices was demonstrated by performing a hide-and-seek test using the low-resolution data; carrying out the search using the native helix yielded an $RMSD_{Min}$ in the range 0.3-1.1 Å while using an ideal helix yields an $RMSD_{Min}$ in the range 1.1-1.8 Å. One way of

incorporating experimentally determined helices into our modeling protocol is through the use of helical protein structures deposited in the PDB. Initial tests using experimentally-derived helices extracted from the PDB reveal significant improvements in $RMSD_{Min}$ when compared to the case of using an ideal helix. Using the native helix from each case in **Table 2**, we searched the PDB using a rapid distance-matrix structure search method[28] and extracted all helices below an RMSD of 0.5 Å to the native helix. The helix with the smallest RMSD to the native helix was used to carry out a rigid-body search using the low-resolution experimental data. The $RMSD_{Min}$ value when using the PDB-derived helices range from 0.5-1.3 Å which is not significantly different from the case of using the native helix. However, these PDB-derived helices were obtained using the native helix which will likely be unavailable in a practical modeling situation. It is clear that devising a way to incorporate structural information from the PDB into our modeling protocol would provide substantial enrichment of near-native conformations.

The ultimate utility of our method to experimental biologists would be to avoid performing exhaustive mutagenesis experiments when attempting to model homo-oligomeric helical TM structure. Earlier work in our group used phylogeny information along with lattice models to determine how much information experimental information is needed to make reasonable predictions (29). For the method developed in our current study, we find that the more experimental data points provided as input, the more accurate the final predictions will be. However, a judicious choice of sequence region to target for carrying out the mutagenesis experiments can yield accurate results with far fewer experimental data points. We find that for GpA, 8 contiguous experimental data points are sufficient to generate predictions that are within 1 Å RMSD to native (**Supplementary Table 9**). Selecting 8 contiguous experimental data points from the N-terminal, center or C-terminal regions of the TM sequence produces results

that are similar to what is obtained using the full set of 23 experimental data points. Splitting the 8 contiguous experimental data points into 4 contiguous experimental data points at both n-terminal and c-terminal ends of the TM sequence for GpA yields a prediction that is also near 1 Å RMSD to native. We find a similar result for the phospholamban case when using 4 contiguous experimental data points at both the N-terminal and C-terminal ends.

Future improvements to our method will include adding additional terms to our simple two-term scoring function. One possibility would be to include knowledge-based terms to improve the packing between helices. Work by Harrington and Ben-Tal (30) show that five types of chemical interactions common to TM helices could be used to essentially generate sub angstrom predictions. It would be interesting to see if these five types of chemical interactions could be used to complement our current scoring function to filter out conformations that do not exhibit structural determinants common to TM helices. Such an approach would incur minimal computational cost while enriching the ensemble with more native-like models.

While the manuscript was in review, a refined structure for phospholamban was published by Verardi et al. (11). Comparing our prediction for phospholamban to this new structure (PDB ID: 2KYV) gives a final prediction of 0.8 Å. Using 4 contiguous experimental data points at both the N-terminal and C-terminal ends also gives a prediction around 0.8 Å.

2.5 Conclusion

In summary, we have developed a tool for rapidly modeling helical TM homo-oligomers that uses low-resolution experimental data directly in the modeling process. At the heart of our modeling protocol is the use of a correlation term that restrains the rigid-body sampling and avoids costly searches in regions of conformational space that do not correlate with experimental information. We show that correlating mutagenesis, crosslinking and ion channel

data with inter-subunit Cβ distance data followed by refinement provides accurate models for helical TM proteins exhibiting exact rotational symmetry. One area where our modeling approach is likely to have a significant impact is in situations where it is either too difficult or time consuming to obtain a complete set of NMR data.

2.6 Materials and Methods

Details regarding the low-resolution experimental data

At the core of our sampling methodology is the use of experimental information in the form of mutagenesis data, crosslinking data, TOXCAT and TOXR data. We provide a short description of the data below. All of the experimental data used in this study can be found in **Supplementary Tables 2-8**.

Phospholamban—Phospholamban is homo-pentameric bundle located in the sarcoplasmic reticulum of cardiocytes and is responsible for calcium transport. The mutagenesis data for phospholamban was taken from Table 2 of Simmerman et al. (9). The data from this table shows the extent of pentamer formation following mutation to either an alanine or a phenylalanine along the transmembrane region. For this study, we used the alanine mutational data only.

M2 and BM2—Both M2 (A/M2) and BM2 are homo-tetrameric TM proton transporters belonging to different types of influenza viruses. These proton transporters are responsible for acidifying the interior of the virus which ultimately leads to virion uncoating in the endosomes. For both M2 and BM2 we used the pertubational index (PI) which is a combination of reversal potential, current and specific activity data (see Pinto et al. (24) for details). PI data for M2 from was obtained from Figure 1 of Pinto et al. (24). PI data for BM2 was obtained from Figure 3 of Ma et al. (18).

GpA—GpA is homo dimeric sialoglycoprotein from erythrocyte cells. Two sets of data for the TM region of of GpA were used. The first set of data was obtained from Figure 5 of Lemmon et al., (8) and shows the relative degree of disruption of the GpA dimer by mutation of the native sequence with a nonpolar residue. The degree of disruption of the dimer interface uses a scale of 0 (no effect on dimer formation) to 3 (no dimer formation). The second set of data was taken from Supplementary Tables of Zhu et al (4) and shows the percentage of crosslinking between residues in the transmembrane region of the αIIbβ3/GpA chimera. For this study, we considered only symmetric crosslinking data between residues.

BNIP3— is a homo-dimer and a member of the Bcl-2 homology domain-3 subfamily of proapoptotic Bcl-2 proteins. BNIP3 is associated with apoptotic response in the myocardium. Mutational data for the TM region of BNIP3 dimer comes from Figure 7 of Lawrie et al. (23) and represents a combination of TOXCAT and SDS-PAGE page phenotype scores based on percentage dimer disruption. The "unified score" gives the average disruptive effect of different amino acid substitution along the protein sequence of the transmembrane helix. The unified scale ranges in value from 0 (no dimer formation) to 10 (strong dimerization). We used all the phenotype scores from the alanine mutations.

EphA1—ephrin receptor A1 is part of a receptor tyrosine kinase and is involved in animal development and certain cancers. ToxR data was taken from Table 3 of the Volynsky et al. (15). It should be noted that the ToxR data from Volynsky et al. do not consider every possible residue along the TM region. For our purposes, the mutations to glutamine and serine were not as informative, since mutation to a polar residue in the membrane could cause the dimer to be disrupted even when the residue is not along the dimer interface. Similarly, glycine mutations may cause a structural change in the helices despite not residing at the oligomer interface.

Therefore, we concentrated on only the hydrophobic mutations (i.e., the isoleucine, alanine and valine mutations).

Generation of the homo-oligomeric models

Generation of the oligomer begins with the construction of an ideal helix using CHARMM22 internal geometry with ϕ and ψ dihedral angles set to -60 and -40 degrees respectively. Modeling was carried out only for the TM region that had experimental data. As such, the sequence length of the helix was dependent on the available experimental data.

The sampling procedure begins with an ideal helix containing the native sequence. Side chains were not considered at this stage but all residues (with the exception of glycine) contained a Cβ atom. The individual steps used to position the helix in space are depicted in cartoon form in **Figure 1**. Each search for the best dimer configuration begins with a set of initial parameters for the four variables Tx, Tz, θ, and ϕ that were applied to an ideal helix centered at the origin of the global frame of reference. To maximize the correlation of experimental data with inter-subunit Cβ distance while maintaining a sterically feasible distance between helices, the scoring function below was optimized using the Nelder-Mead simplex algorithm from the Gnu Scientific Library (GSL) (31).

$$Score = \sum_{i=0}^{M}\sum_{j=0}^{M} C_1 \cdot \left(\max\left(0, \left(S \cdot R_{ij}^{Min}\right)^2 - R\left(T_X, T_Z, \theta, \phi\right)_{ij}^{\ 2}\right)\right) + \sum_{i=j=0}^{N} C_2 \cdot \begin{cases} 1 - \mathrm{corr}(R\left(T_X, T_Z, \theta, \phi\right)_{ij}, \mathrm{y_{ij}}) & \textit{if correlated} \\ 1 + \mathrm{corr}(R\left(T_X, T_Z, \theta, \phi\right)_{ij}, \mathrm{y_{ij}}) & \textit{if anti-correlated} \end{cases}$$

The weighting parameters C_1 (kcal/mol·Å^2) and C_2 (kcal/mol) were set to 75 and 100 respectively. The scale variable, S, is used to soften the van der Waal's radii. For the study carried out here, S was set to 0.80. R_{ij}^{Min} is the sum of the van der Waal's radii of two atoms ij (the radii for different atoms were taken directly from the XPLOR manual (32)). The distance between two atoms is denoted as '$R(T_x,T_z,\theta,\phi)_{ij}$' and is a function of the four search parameters

Tx, Tz, θ, and ϕ. The 'corr' term is the correlation between the inter-subunit Cβ distances and the corresponding experimental data. The experimental data (denoted 'y' in the above equation) denotes the same residue on symmetric helices and is correlated with the distance between these residues. The variable M represents all the atoms from each helix while the variable N represents only the Cβ atoms.

Initial values for the sampling parameters were obtained by coarse enumeration between a suitable set of numerical boundaries and were subject to the three following conditions: 1) the bundle radius, T_x, for a dimer must lie between 2 and 4 Å. For oligomeric states larger than 2, the bundle radius should be restrained between 6-9 Å. 2) the tilt angle, θ, must lie in the range of -30 to 30 degrees; and 3) the translation along the Z axis, T_z, measured from the geometric center of the helix lie in the range of -10 to 10 Å. To avoid a combinatorial explosion of values for the parameters, we capped the maximum number of initial values to 1000. Only models with a Score less than 50 were retained for the Refinement phase.

Clustering

Models were clustered using a k-medoid algorithm from the C clustering library19. The number of initial clustering attempts was set to 100. The model with smallest RMSD to all other models in the cluster was selected as the centroid model.

Side chain placement

Side chains were added to all the centroid models using the side chain prediction program SCAP (20). Default options were used with SCAP.

Refinement with XPLOR-NIH

After side chain addition, all of the centroid models from the first phase were subjected to 100 steps of rigid-body minimization (RBM) using XPLOR-NIH with the CHARMM22 force field.

The goal of this step in the refinement procedure is to enforce proper packing between helices while removing any steric clashes that arise as a result of having used a reduced representation for the side chain during the generation of the oligomer in the first phase. The RBM step is then followed by two thousand steps of Powell minimization. The dielectric constant was set to 4 and the non-bonded cutoff distance was set to 12.0 Å. All other options were left at their default values.

Correlation versus anti-correlation

It should be noted that while some experimental data will correlate positively with inter-subunit Cβ distance, some data will correlate negatively with inter-subunit Cβ distance. This can be understood by considering the case for cross-linked residues. If two corresponding residues in a homo-dimer are close in space the distance between the Cβ atoms will be small. In this scenario, the Cβ atoms should cross link strongly. The data would then be anti-correlated with a maximal value of -1 since a small distance yields a stronger (large magnitude) crosslinking signal. The statistical significance attributed to r is the probability of arriving at the current value if the correlation coefficient were in fact actually zero (the null hypothesis). For the purposes of this study r is considered statistically significant if the associated p value is less than 5% ($p<0.05$). Correlation coefficients and their respective p-values were calculated in Matlab. The p-values are one-sided, and represent the probability that two uncorrelated sequences of the given length would have a correlation value as good as the calculated correlation by chance.

Root mean square distance (RMSD) calculations

The root mean square distance (RMSD) is computed by optimally superimposing N, Cα, C and O atoms from a model onto the native structure. For the purposes of comparing our best prediction with a NMR models, we used the average model computed from all the individual

models in the NMR ensemble. For the results involving refinement of the native models, we superimposed the refined native model onto the unrefined native model to determine the RMSD.

2.7 Acknowledgments

We acknowledge support from NIH grants 54616 and 60610 from the NIH to WFD and support from the MRSEC program of NSF. BTH would like to thank the DOD for support through an NDSEG Fellowship. We would like to acknowledge Dr. Gevorg Grigoryan for carefully reading the manuscript and for allowing us to use his rapid distance-matrix based method for scanning the PDB. We also would like to thank Gerald Cadena for help with **Figure 1**.

A web interface is available at http://www.degradolab.org/OSTRICh

2.8 Figures

Name	Oligomer	PDB ID	Residues	$RMSD_{Score}$	$RMSD_{Min}$	$RMSD_{Native}$
GpA	Dimer	1AFO	73-96	**1.7**	**1.4**	**0.1**
EphA1	Dimer	2K1L	549-560	**1.3**	**1.0**	**0.2**
BNIP3	Dimer	2KA2	170-184	**0.9**	**0.9**	**0.3**
M2(xtal)	Tetramer	3LBW	26-43	**1.4**	**1.4**	**1.4**
M2(NMR)	Tetramer	2RLF	26-43	**1.5**	**1.2**	**0.1**
M2(ssNMR1)	Tetramer	2KQT	26-43	**1.3**	**1.2**	**0.0**
M2(ssNMR2)	Tetramer	2LOJ	26-43	**1.5**	**1.2**	**0.2**
BM2	Tetramer	2KIX	7-25	**2.0**	**1.6**	**0.6**
Phospholamban	Pentamer	1ZLL	37-52	**2.9**	**0.9**	**0.3**

Table 1

Application of the sampling method using inter-helical Cβ distances from experimentally determined helical TM structures.

Name	Type of Data	$RMSD_{Score}$	$RMSD_{Min}$
GpA	Mutagenesis	**1.4**	**1.4**
GpA(Crosslinking)	Crosslinking	**1.6**	**1.4**
EphA1	TOXCAT	**1.6**	**1.4**
BNIP3	TOXCAT/Mutagenesis	**3.0**	**1.6**
M2(xtal)	Ion conductance	**1.9**	**1.5**
M2(NMR)	Ion conductance	**1.4**	**1.3**
M2(ssNMR1)	Ion conductance	**1.6**	**1.5**
M2(ssNMR2)	Ion conductance	**1.7**	**1.3**
BM2	Ion conductance	**2.1**	**1.8**
Phospholamban	Mutagenesis	**3.0**	**1.0**

Table 2

Application of the sampling method using low-resolution experimental data.

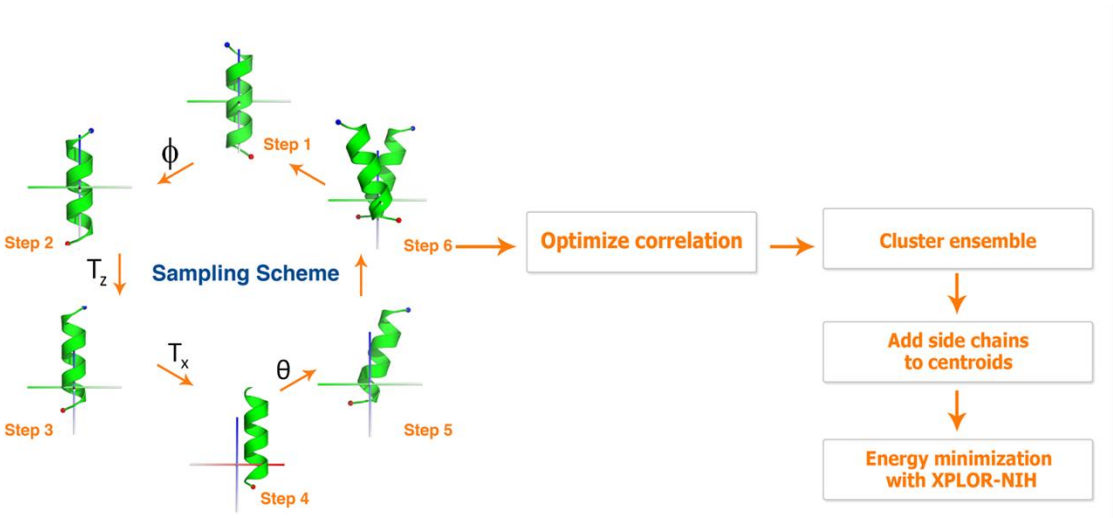


Figure 1

An illustration of the steps used to model a TM oligomer with exact rotational symmetry. **Step 1** starts with an ideal helix transformed to the global frame of reference such that the geometric center of the helix is positioned at the origin. **Step 2** involves rotation about the global Z axis and determines which residues will form the interface of the dimer (denoted by the variable ϕ). **Step 3** is a translation along the global Z axis and will affect the point of closest approach (denoted by the variable T_z). **Step 4** is a translation along the global X axis and will affect the radius of the bundle (denoted by the variable T_x). **Step 5** is a rotation about the global X axis and will affect the tilt of the bundle with respect to the global Z axis (denoted by the variable θ). **Step 6** is a rotation about the global Z axis used to generate the symmetry mate followed by optimization between experimental data and inter-subunit Cβ distance (see Methods for a description of the two-term scoring function used in the optimization step). Once an ensemble of 1000 poses has been generated the ensemble is clustered and side chains are added to the centroid models. The centroid models are then refined using XPLOR-NIH. The spheres on the

end of the helices denote the n-terminus (blue) and c-terminus (red). The individual axes on the global frame are color coded as follows: red denotes the positive X axis, green denotes the positive Y axis and blue denotes the positive Z axis.

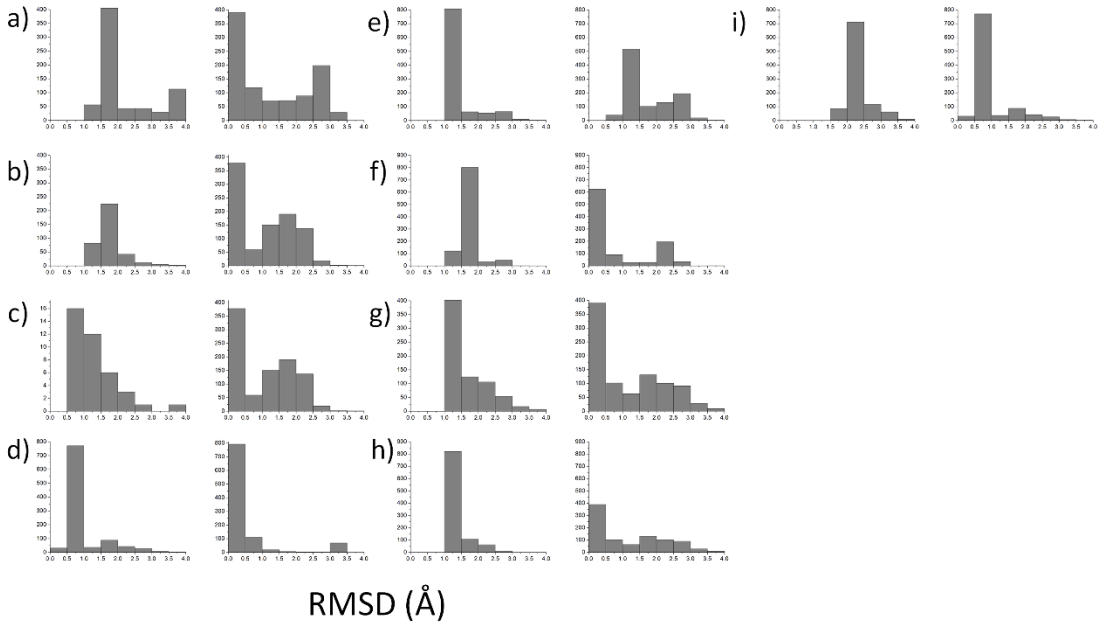


Figure 2

RMSD distributions for models generated using native inter-subunit Cβ distances. Each panel consists of two distributions. The distribution to the left was generated using an ideal helix. The distribution to the right was generated using the native helix. The RMSD value is between each model in the generated ensemble and the native structure. The distributions are: a) GpA b)EphA1 c) BNIP3 d) phospholamban e) M2(xtal) f) M2(NMR) g) M2(ssNMR1) h)M2(ssNMR2) and i) BM2.

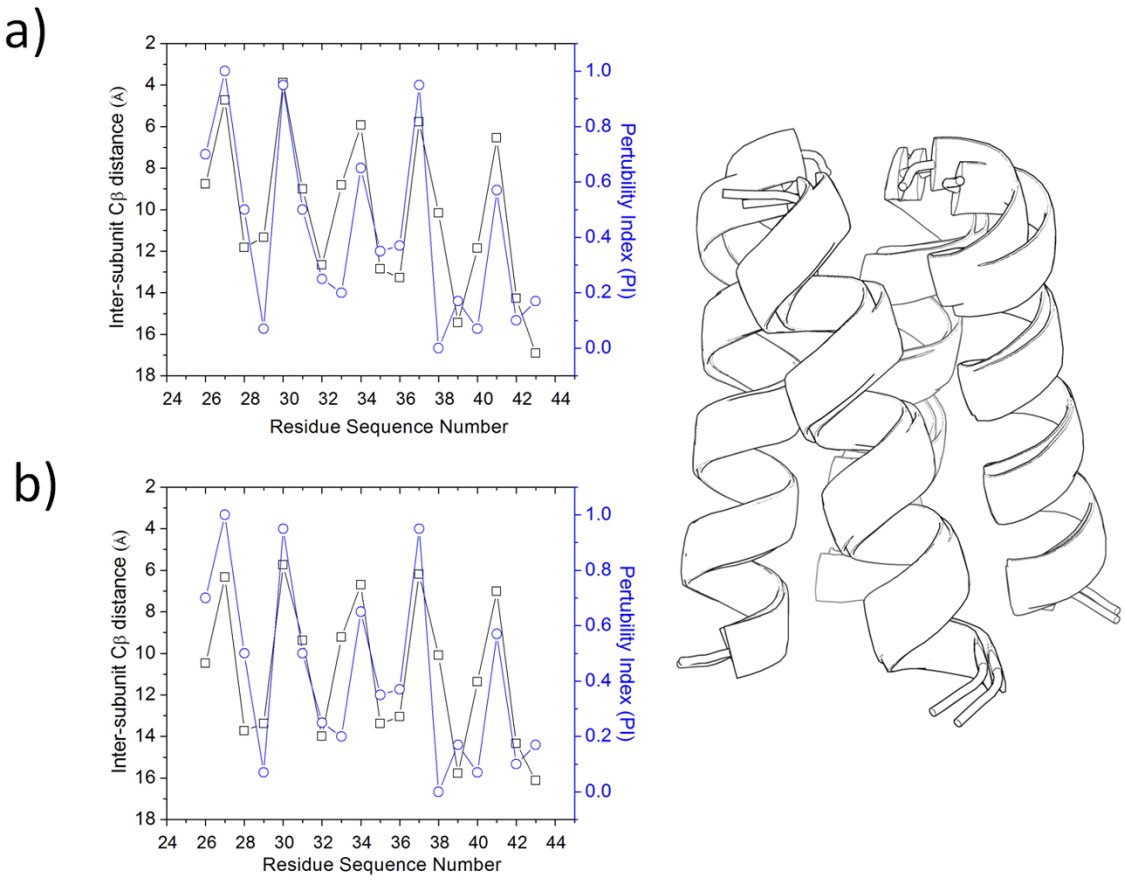


Figure 3

PI versus inter-subunit Cβ distance profiles for a) native M2(NMR) structure and b) our best scoring model after refinement with XPLOR-NIH. A superimposition of our best scoring model and the native M2(NMR) structure is shown on the right.

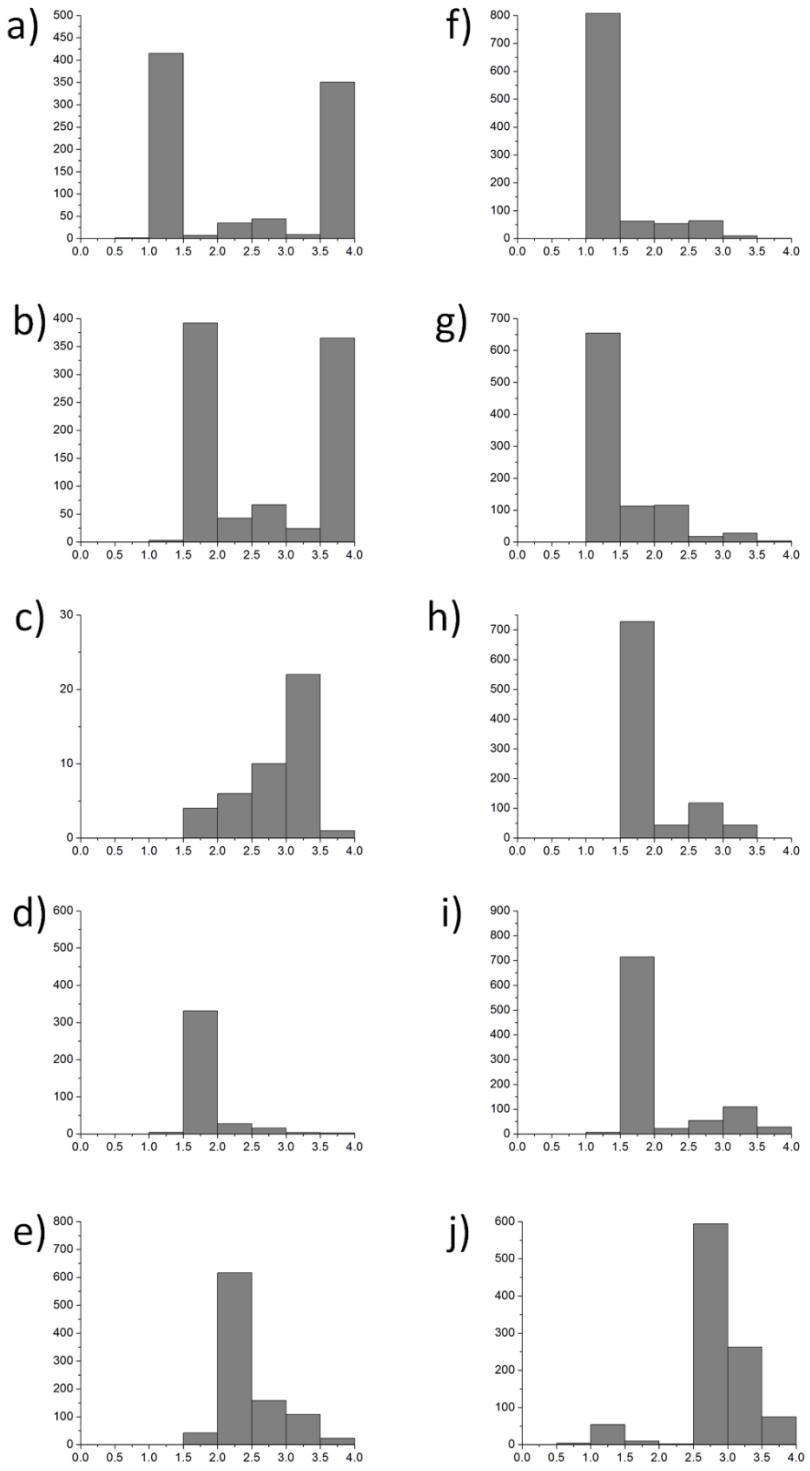

a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
RMSD (Å)

Figure 4

RMSD distributions for models generated using low-resolution experimental data. Each panel consists of a single distribution that was generated using an ideal helix. The RMSD value is obtained between each model in the ensemble and the native structure. The distributions are as follows: a) GpA b)GpA(Crosslinking) c) BNIP3 d) EphA1 e) BM2 f) M2(xtal) g) M2(NMR) h)M2(ssNMR1) i) M2 (ssNMR2) and j) phospholamban.

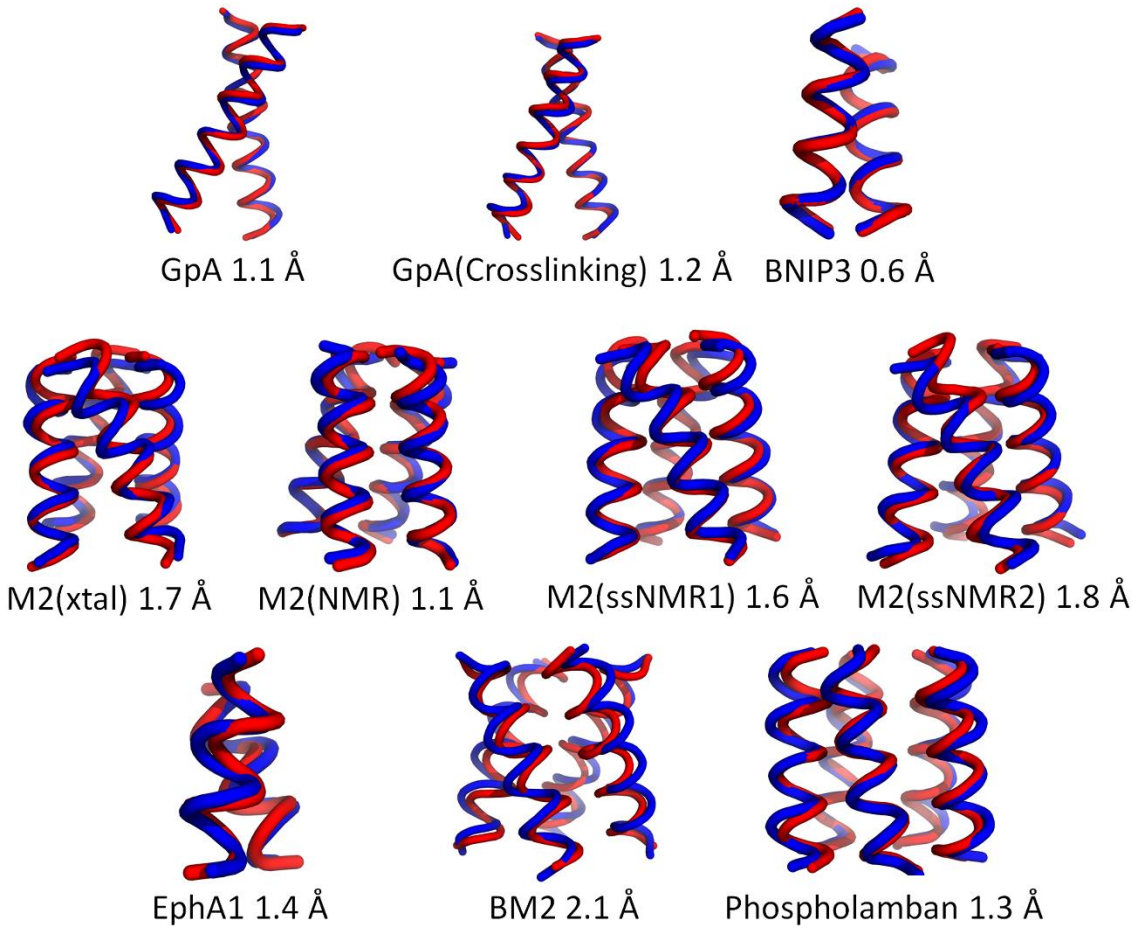


Figure 5

A comparison between the backbone of the native structure (shown in blue) and best scoring model (shown in red) after refinement

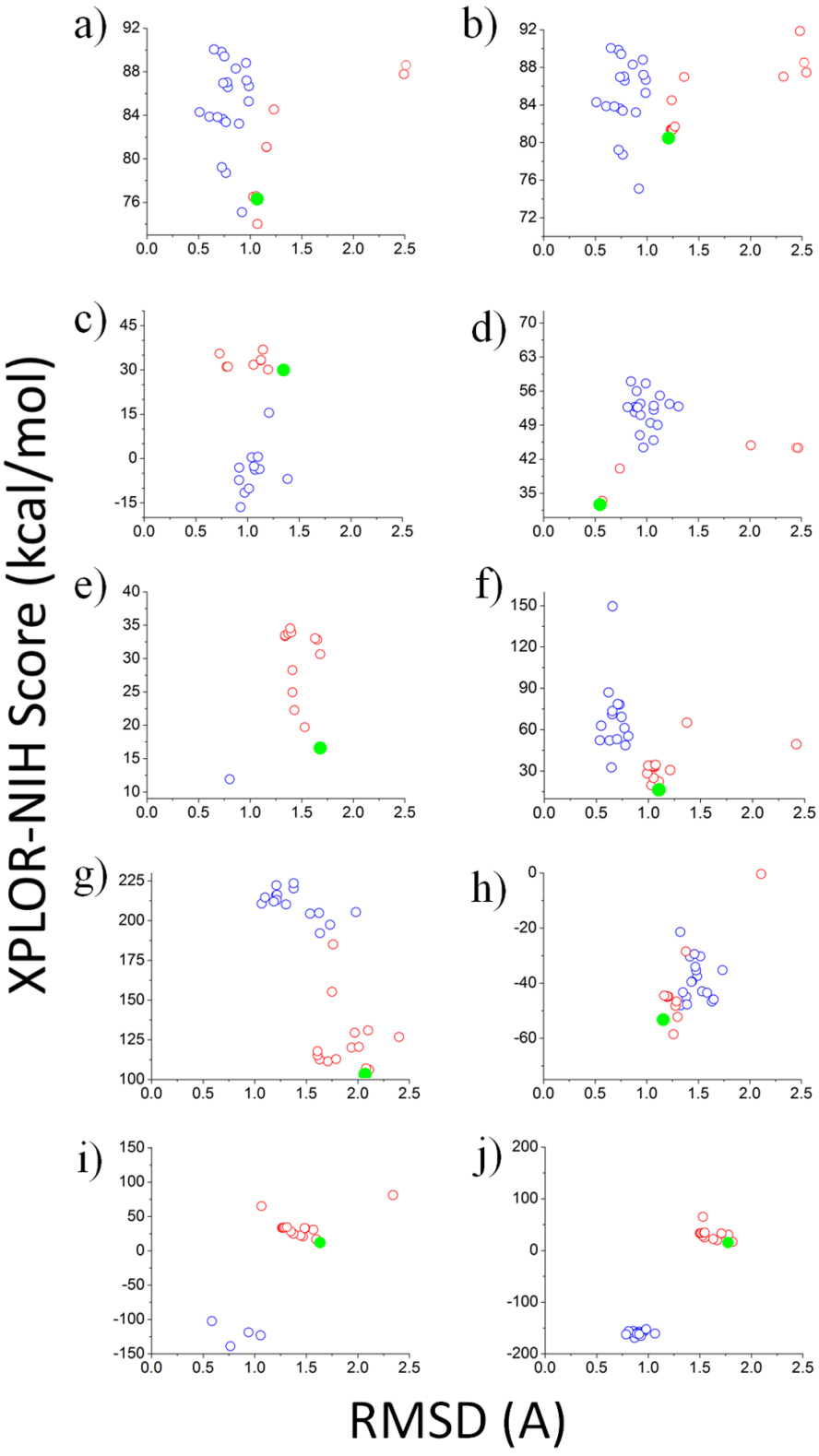

a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
XPLOR-NIH Score (kcal/mol)
RMSD (A)

Figure 6

Energy profiles versus RMSD to native after refinement with XPLOR-NIH. Blue circles show the RMSD between the starting native model (from the NMR ensemble) and the native model after refinement with XPLOR-NIH. Red circles show the RMSD between each of the centroids and the corresponding native structure. The filled green circles represent the most favorable scoring model among the 20 centroid models. Only models with RMSD values below 2.5 Å are displayed on the graph. The panels are labeled as follows (a) GpA (b) GpA(Crosslinking) (c) BNIP3 (d) EphA1 (e) M2(xtal) (f) M2(NMR) (g) BM2 (h) phospholamban (i) M2(ssNMR1) and (j) M2(ssNMR2).

2.9 Supplementary Figures

Model	RMSD(Å)	ΔTx (Å)	Δθ (deg)	Δϕ (deg)	ΔTz (Å)
1	0.3	0.3	0.6	1.6	0.1
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.9	0.0
7	0.0	0.0	0.0	0.8	0.0
8	0.0	0.0	0.0	0.0	0.0
9	0.6	0.6	1.8	2.8	0.2
10	0.5	0.0	1.4	1.1	0.2

"Δ" indicates the difference between the values of the actual parameters and those obtained after the rigid-body search

Supplementary Table 1

Hide-and-seek test using inter-subunit Cβ distance as simulated experimental data.

Residue	SeqNumber	Chain	Atom	Dimer Disruption	Flag
ILE	73	A	CB	0.30	1.0
THR	74	A	CB	0.00	1.0
LEU	75	A	CB	1.80	1.0
ILE	76	A	CB	2.00	1.0
ILE	77	A	CB	0.25	1.0
PHE	78	A	CB	0.00	1.0
GLY	79	A	CA	2.30	1.0
VAL	80	A	CB	1.00	1.0
MET	81	A	CB	0.00	1.0
ALA	82	A	CB	0.00	1.0
GLY	83	A	CA	3.00	1.0
VAL	84	A	CB	2.50	1.0
ILE	85	A	CB	0.00	1.0
GLY	86	A	CA	0.20	1.0
THR	87	A	CB	2.00	1.0
ILE	88	A	CB	0.75	1.0
LEU	89	A	CB	0.65	1.0
LEU	90	A	CB	0.40	1.0
ILE	91	A	CB	0.65	1.0
SER	92	A	CB	0.00	1.0
TYR	93	A	CB	0.30	1.0
GLY	94	A	CA	0.00	1.0
ILE	95	A	CB	0.40	1.0

A Flag value of '1.0' indicates that the data value was used

Supplementary Table 2

GpA disruption data obtained from Figure 5 of Lemmon *et al.* (8)

Residue	SeqNumber	Chain	Atom	%Crosslinking	Flag
ILE	73	A	CB	0.00	1.0
THR	74	A	CB	1.00	1.0
LEU	75	A	CB	44.50	1.0
ILE	76	A	CB	29.25	1.0
ILE	77	A	CB	6.75	1.0
PHE	78	A	CB	1.25	1.0
GLY	79	A	CA	87.25	1.0
VAL	80	A	CB	66.67	1.0
MET	81	A	CB	1.67	1.0
ALA	82	A	CB	0.00	1.0
GLY	83	A	CA	83.00	1.0
VAL	84	A	CB	38.00	1.0
ILE	85	A	CB	1.17	1.0
GLY	86	A	CA	28.00	1.0
THR	87	A	CB	59.67	1.0
ILE	88	A	CB	38.00	1.0
LEU	89	A	CB	6.83	1.0
LEU	90	A	CB	38.00	1.0
ILE	91	A	CB	53.00	1.0
SER	92	A	CB	1.50	1.0
TYR	93	A	CB	5.00	1.0
GLY	94	A	CA	15.50	1.0
ILE	95	A	CB	3.00	1.0

A Flag value of '1.0' indicates that the data value was used

Supplementary Table 3

GpA disulfide cross-linking data obtained from Supplementary Information of Zhu *et al.* (4)

Residue	SeqNumber	Chain	Atom	Pentamer Disruption	Flag
LEU	37	A	CB	0.7	1.0
ILE	38	A	CB	43.8	1.0
LEU	39	A	CB	27.3	1.0
ILE	40	A	CB	1.6	1.0
CYS	41	A	CB	28.3	1.0
LEU	42	A	CB	44.6	1.0
LEU	43	A	CB	34.8	1.0
LEU	44	A	CB	2.5	1.0
ILE	45	A	CB	43.2	1.0
CYS	46	A	CB	42.1	1.0
ILE	47	A	CB	2.0	1.0
ILE	48	A	CB	38.1	1.0
VAL	49	A	CB	44.3	1.0
MET	50	A	CB	35.6	1.0
LEU	51	A	CB	14.4	1.0
LEU	52	A	CB	49.1	1.0

A Flag value of '1.0' indicates that the data value was used

Supplementary Table 4

Pentamer disruption data obtained from Table 1 of Simmermann *et al.* (9)

Residue	SeqNumber	Chain	Atom	TOXR	Flag
VAL	549	A	CB	102.0	0.0
ALA	550	A	CB	50.0	1.0
VAL	551	A	CB	94.0	0.0
ILE	552	A	CB	113.0	1.0
PHE	553	A	CB	82.0	1.0
GLY	554	A	CA	4.0	1.0
LEU	555	A	CB	74.0	1.0
LEU	556	A	CB	116.0	1.0
LEU	557	A	CB	116.0	1.0
GLY	558	A	CA	17.0	1.0
ALA	559	A	CB	95.0	1.0
ALA	560	A	CB	103.0	1.0

A Flag value of '1.0' indicates that the data value was used. Only mutations to hydrophobic residues were considered.

Supplementary Table 5

EphA1 TOXR data taken from Table 2 of Volynsky *et al.* (15)

Residue	SeqNumber	Chain	Atom	UnifiedScore	Flag
LEU	170	A	CB	3.0	1.0
LEU	171	A	CB	2.0	1.0
SER	172	A	CB	4.0	1.0
HIS	173	A	CB	9.0	1.0
LEU	174	A	CB	2.0	1.0
LEU	175	A	CB	ND	ND
ALA	176	A	CB	0.0	1.0
ILE	177	A	CB	4.0	1.0
GLY	178	A	CA	1.0	1.0
LEU	179	A	CB	3.0	1.0
GLY	180	A	CA	10.0	1.0
ILE	181	A	CB	5.0	1.0
TYR	182	A	CB	1.0	1.0
ILE	183	A	CB	5.0	1.0
GLY	184	A	CA	7.0	1.0

'ND' indicates that there was no data for this residue. A Flag value of '1.0' indicates that the data value was used.

Supplementary Table 6

BNIP3 unified mutagenesis score values for alanine were taken from Figure 7 of Lawrie *et al.* (23)

Residue	SeqNumber	Chain	Atom	PI (Avg)	Flag
LEU	26	A	CB	0.70	1.0
VAL	27	A	CB	1.00	1.0
VAL	28	A	CB	0.50	1.0
ALA	29	A	CB	0.07	1.0
ALA	30	A	CB	0.95	1.0
SER	31	A	CB	0.50	1.0
ILE	32	A	CB	0.25	1.0
ILE	33	A	CB	0.20	1.0
GLY	34	A	CA	0.65	1.0
ILE	35	A	CB	0.35	1.0
LEU	36	A	CB	0.37	1.0
HIS	37	A	CB	0.95	1.0
LEU	38	A	CB	0.00	1.0
ILE	39	A	CB	0.17	1.0
LEU	40	A	CB	0.07	1.0
TRP	41	A	CB	0.57	1.0
ILE	42	A	CB	0.10	1.0
LEU	43	A	CB	0.17	1.0

A Flag value of '1.0' indicates that the data value was used

Supplementary Table 7

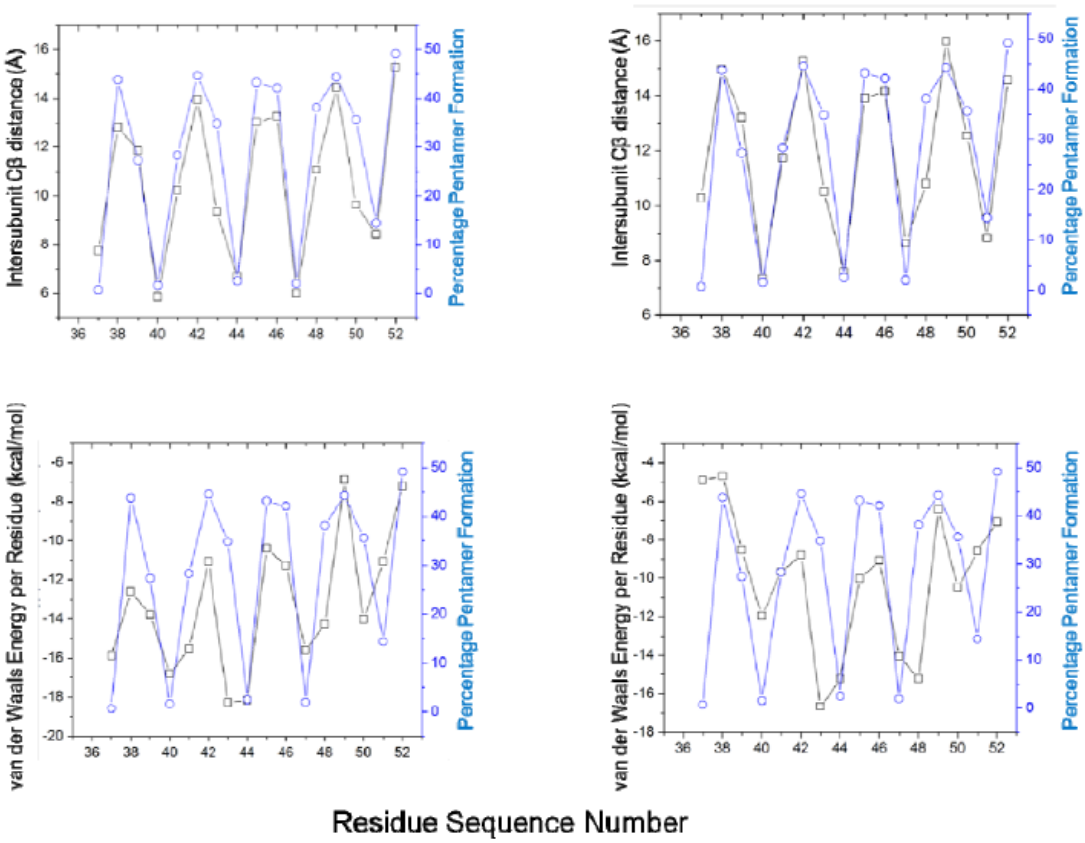
M2 perturbility index (PI) data taken from Figure 2 of Pinto *et al.* (24)

Residue	SeqNumber	Chain	Atom	PI (Avg)	Flag
ILE	7	A	CB	0.30	1.0
LEU	8	A	CB	0.23	1.0
SER	9	A	CB	0.52	1.0
ILE	10	A	CB	0.45	1.0
CYS	11	A	CB	0.09	1.0
SER	12	A	CB	0.34	1.0
PHE	13	A	CB	0.65	1.0
ILE	14	A	CB	0.29	1.0
LEU	15	A	CB	0.48	1.0
SER	16	A	CB	0.33	1.0
ALA	17	A	CB	0.23	1.0
LEU	18	A	CB	0.14	1.0
HIS	19	A	CB	0.99	1.0
PHE	20	A	CB	0.39	1.0
MET	21	A	CB	0.06	1.0
ALA	22	A	CB	0.17	1.0
TRP	23	A	CB	0.77	1.0
THR	24	A	CB	0.06	1.0
ILE	25	A	CB	0.00	1.0

A Flag value of '1.0' indicates that the data value was used

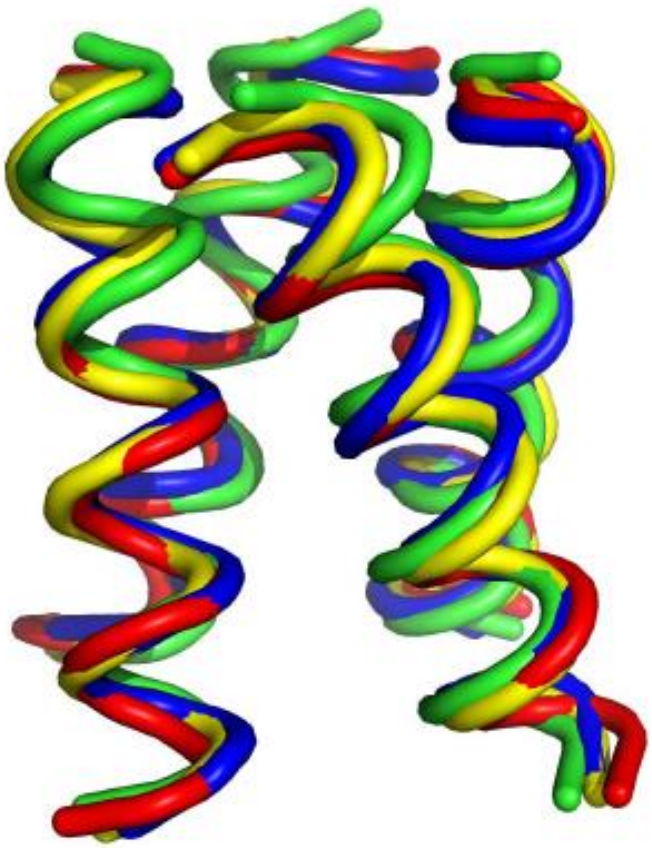
Supplementary Table 8

BM2 perturbility index (PI) data taken from Figure 3 of Ma *et al.* (18)



Supplementary Figure 1

Inter-subunit Cβ distance versus percentage pentamer formation for native phospholamban (upper left panel) and the best scoring model (upper right panel). A similar profile is shown using van der Waals interaction energy per residue versus percentage pentamer formation for native phospholamban (lower left panel) and the best scoring model (lower right panel). The van der Waals interaction energy per residue was obtained using the CHARMM22 force field implementation in XPLOR-NIH.



Only the TM portion of the structures are shown.

Supplementary Figure 2

Structural heterogeneity between different structures of M2 that include the solution NMR structure (green, PDB ID 2RLF), a high-resolution X-ray structure (blue, PDB ID 3LBW), a solid-state NMR structure from Mei Hong's group (yellow, PDB ID 2KQT) and a solid-state NMR structure from Tim Cross' group (red, PDB ID 2L0J).

2.10 References

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
2. Kyogoku, Y., Fujiyoshi, Y., Shimada, I., Nakamura, H., Tsukihara, T., Akutsu, H., Odahara, T., Okada, T. & Nomura, N. (2003). Structural genomics of membrane proteins. *Acc Chem Res* **36**, 199-206.
3. Elofsson, A. & von Heijne, G. (2007). Membrane protein structure: prediction versus reality. *Annu Rev Biochem* **76**, 125-40.
4. Zhu, J., Luo, B. H., Barth, P., Schonbrun, J., Baker, D. & Springer, T. A. (2009). The structure of a receptor with two associating transmembrane domains on the cell surface: integrin alphaIIbbeta3. *Mol Cell* **34**, 234-49.
5. Metcalf, D. G., Law, P. B. & DeGrado, W. F. (2007). Mutagenesis data in the automated prediction of transmembrane helix dimers. *Proteins* **67**, 375-84.
6. Treutlein, H. R., Lemmon, M. A., Engelman, D. M. & Brunger, A. T. (1992). The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry* **31**, 12726-32.
7. Adams, P. D., Arkin, I. T., Engelman, D. M. & Brunger, A. T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat Struct Biol* **2**, 154-62.
8. Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry* **31**, 12719-25.
9. Simmerman, H. K., Kobayashi, Y. M., Autry, J. M. & Jones, L. R. (1996). A leucine zipper stabilizes the pentameric membrane domain of phospholamban and forms a coiled-coil pore structure. *J Biol Chem* **271**, 5941-6.
10. Oxenoid, K. & Chou, J. J. (2005). The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc Natl Acad Sci U S A* **102**, 10870-5.
11. Verardi, R., Shi, L., Traaseth, N. J., Walsh, N. & Veglia, G. Structural topology of phospholamban pentamer in lipid bilayers by a hybrid solution and solid-state NMR method. *Proc Natl Acad Sci U S A* **108**, 9101-6.
12. Herzyk, P. & Hubbard, R. E. (1998). Using experimental information to produce a model of the transmembrane domain of the ion channel phospholamban. *Biophys J* **74**, 1203-14.
13. Fleishman, S. J. & Ben-Tal, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J Mol Biol* **321**, 363-78.
14. Kim, S., Chamberlain, A. K. & Bowie, J. U. (2003). A simple method for modeling transmembrane helix oligomers. *J Mol Biol* **329**, 831-40.
15. Volynsky, P. E., Mineeva, E.A., Goncharuk, M. V., Ermolyuk, Y. S., Arseniev, A. S., and R. G. Efremov. (2010). Computer simulations and modeling-assisted ToxR screening in deciphering 3D structures of transmembrane alpha-helical dimers: ephrin receptor A1. *Phys. Biol.* **7**, 15.

16. Rouse, S. L., Carpenter, T., Stansfeld, P. J. & Sansom, M. S. (2009). Simulations of the BM2 proton channel transmembrane domain from influenza virus B. *Biochemistry* **48**, 9949-51.
17. Dieckmann, G. R. & DeGrado, W. F. (1997). Modeling transmembrane helical oligomers. *Curr Opin Struct Biol* **7**, 486-94.
18. Ma, C., Soto, C. S., Ohigashi, Y., Taylor, A., Bournas, V., Glawe, B., Udo, M. K., Degrado, W. F., Lamb, R. A. & Pinto, L. H. (2008). Identification of the pore-lining residues of the BM2 ion channel protein of influenza B virus. *J Biol Chem* **283**, 15921-31.
19. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. (2004). Open source clustering software. *Bioinformatics* **20**, 1453-4.
20. Xiang, Z., Steinbach, P. J., Jacobson, M. P., Friesner, R. A. & Honig, B. (2007). Prediction of side-chain conformations on protein surfaces. *Proteins* **66**, 814-23.
21. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* **160**, 65-73.
22. Sulistijo, E. S. & Mackenzie, K. R. (2009). Structural basis for dimerization of the BNIP3 transmembrane domain. *Biochemistry* **48**, 5106-20.
23. Lawrie, C. M., Sulistijo, E. S. & MacKenzie, K. R. Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: roles for sequence context in helix-helix association in membranes. *J Mol Biol* **396**, 924-36.
24. Pinto, L. H., Dieckmann, G. R., Gandhi, C. S., Papworth, C. G., Braman, J., Shaughnessy, M. A., Lear, J. D., Lamb, R. A. & DeGrado, W. F. (1997). A functionally defined model for the M2 proton channel of influenza A virus suggests a mechanism for its ion selectivity. *Proc Natl Acad Sci U S A* **94**, 11301-6.
25. Sharma, M., Yi, M., Dong, H., Qin, H., Peterson, E., Busath, D. D., Zhou, H. X. & Cross, T. A. Insight into the mechanism of the influenza A proton channel from a structure in a lipid bilayer. *Science* **330**, 509-12.
26. Cady, S. D., Schmidt-Rohr, K., Wang, J., Soto, C. S., Degrado, W. F. & Hong, M. Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers. *Nature* **463**, 689-92.
27. Sale, K., Faulon, J. L., Gray, G. A., Schoeniger, J. S. & Young, M. M. (2004). Optimal bundling of transmembrane helices using sparse distance constraints. *Protein Sci* **13**, 2613-27.
28. Grigoryan, G., Kim, Y. H., Acharya, R., Axelrod, K., Jain, R. M., Willis, L., Drndic, M., Kikkawa, J. M. & DeGrado, W. F. Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science* **332**, 1071-6.
29. Nanda, V. & DeGrado, W. F. (2005). Automated use of mutagenesis data in structure prediction. *Proteins* **59**, 454-66.
30. Harrington, S. E. & Ben-Tal, N. (2009). Structural determinants of transmembrane helical proteins. *Structure* **17**, 1092-103.
31. GNU Scientific Library *http://www.gnu.org/software/gsl/*.
32. Brunger, A. (1991). X-PLOR Version 3.1 A System for X-ray Crystallography and NMR. *Yale University Press*.

Chapter 3

Nature-inspired design of motif-specific antibody scaffolds

3.1 Abstract

Aberrant changes in post-translational modifications (PTMs) such as phosphorylation underlie a majority of human diseases. However, detection and quantification of PTMs for diagnostic or biomarker applications often requires monoclonal PTM-specific antibodies, which are challenging to generate using traditional antibody generation platforms. Here we outline a renewable synthetic antibody strategy by installing a novel motif-specific hot spot into an antibody scaffold. Inspired by a natural phosphate-binding motif, we designed antibody scaffolds with hot spots specific for phosphoserine, phosphothreonine, or phosphotyrosine. Crystal structures of the phospho-specific antibodies revealed two distinct modes of phosphoresidue recognition. These hot spots function independently of the surrounding scaffold as phage display antibody libraries based upon these scaffolds successfully yielded over fifty phospho- and target-specific antibodies against 70% of target peptides. Ultimately, our motif-specific scaffold strategy may provide a general solution for the rapid, robust development of monoclonal anti-PTM or anti-peptide antibodies for signaling, diagnostic, and therapeutic applications.

This chapter has been published in Nature Biotechnology (2013 August 18; 31:916-921). Dr. James T. Koerber is the first author of the work. Nathan D. Thomsen, Brett T. Hannigan, and William F. DeGrado are co-authors. James A. Wells is the corresponding author. I performed the computational analysis of antibody structures, identified the antibody structure with a "nest" motif, and contributed to designing the gene library used in phage-display selections.

3.2 Introduction

Post-translational modifications (PTMs), such as phosphorylation, acetylation, and ubiquitination, play essential roles in modulating protein function throughout biology. In

particular, phosphorylation is one of the most common regulatory mechanisms in eukaryotes where roughly 20-30% of all proteins can be phosphorylated by over 500 kinases[1]. Given the ubiquitous role of phosphorylation in signal transduction, it is not surprising that aberrant phosphorylation either directly causes or is a consequence of many human diseases, such as cancer and neurodegenerative disorders[2]. Recent advances in phosphoproteomic methods have greatly expanded the number of known phosphorylation sites (>170,000) and identified global phosphorylation changes that occur during disease[3-6]. Ultimately, the validation of key phosphorylation events is best conducted at the single-cell level where recent studies, utilizing phospho-specific (PS) monoclonal antibodies (Abs), have elucidated how stochastic fluctuations and signaling cross-talk contribute to the overall cellular state[7, 8]. Unfortunately, very few commercially available Abs are suitable for this purpose[7] and since the number of functionally important phosphorylation sites steadily increases, there exists the need for a rapid, robust method to generate high-quality, renewable, monoclonal PS detection reagents. Furthermore, we seek to make renewable, recombinant Abs to provide genetically encoded functional tools for cell biology.

The state of the art in PS detection reagents is the generation of Abs by the immunization of animals[9]. However, the generation of a polyclonal PS Ab is often imprecise, low-throughput, expensive, time-consuming and not renewable. Furthermore, the development of monoclonal PS antibodies requires additional screening of numerous hybridomas, which is made more challenging by the rarity of PS Ab clones, estimated to be 0.1-5%[10, 11]. Finally, disproportionately more phosphotyrosine (pTyr)-specific Abs exist than phosphoserine (pSer)- or phosphothreonine (pThr)-specific Abs. This fact has hindered the study of serine and threonine phosphorylation, which account for 90% and 10% of all phosphorylation sites,

respectively, compared to <0.05% for tyrosine[12]. Unfortunately, attempts to generate recombinant PS Abs using *in vitro* selection methods, such as phage display[13-17], yeast display[18], and ribosome display[19], have been even less efficient than immunization methods[18, 20-22]. Engineered endogenous phosphopeptide-binding domains such as Src-homology-2 (SH2) or forkhead-associated (FHA) domains may provide an alternative to Abs, but the general utility of these scaffolds remains to be demonstrated[23-25].

Recently, the combination of immunization and phage display was utilized to isolate a high affinity PS Ab from chickens[21]. While this approach was successful and led to the first PS Ab structure, this approach relies upon a low-throughput and time-consuming immunization step. We hypothesize that both immunization and *in vitro* methods for generating PS Abs fail to routinely yield high quality Abs because most naïve Abs do not possess any initial affinity for the small peptide antigens. In light of these difficulties, we envisioned a novel structure-guided Ab generation strategy that employs Ab scaffolds with engineered pockets tailored to a particular sequence motif. This motif-specific anchoring pocket would provide initial antigen-binding affinity and guide the selection of Abs targeted to epitopes containing the motif (e.g. a pSer- or pTyr-containing peptide). Investigators have termed these motif residues "hot spots" that contribute a substantial fraction of the binding energy to a protein-protein interaction[26, 27]. Here we engineer Ab scaffolds with designed binding pockets for pSer, pThr, or pTyr residues and thus, make these residues hot spots in the antigen-Ab interaction. Guided by a natural phosphate-binding motif and knowledge of Ab structure-function, we first identified a parent Ab scaffold in which to install the designed pocket in the complementarity-determining regions (CDRs). We then mutated the scaffold to specifically bind pSer, pThr, or pTyr and solved the X-ray crystal structures of PS Ab:peptide complexes. In the second step, we constructed two large

diverse single-chain Fv (scFv) Ab phage display libraries based upon these scaffolds and successfully selected 51 PS Abs against seven different pSer- or pThr-containing peptides. These results suggest that the phosphoresidue-binding pocket functions independently of additional structural and functional changes in other CDRs of the Ab.

3.3 Results

Design of PS Ab scaffolds

To design a phosphate-binding motif into an Ab scaffold, we drew upon structural knowledge of how protein domains recognize anions, such as phosphate. The most common anion-binding motif, termed a nest, occurs within many different protein super-families, such as ATPases and kinases, and consists of three consecutive residues where multiple main-chain amides form hydrogen bonds with the anion (**Supplementary Fig. 1a**)[28]. Inspired by this ubiquitous motif, we sought to find an existing Ab scaffold into which we could build a similar short, localized loop. We focused our search on sixty anti-peptide Ab structures and manually inspected the CDRs for the desired nest conformation. We identified a region of CDR H2 within a mouse Fab (PDB ID 1i8i)[29] that adopts the desired conformation due to a hallmark α_L glycine at 54_H (Fig. 1a). Interestingly, this Ab utilized the H2 loop to bind an acidic residue via six loop residues that anchor the peptide (52_H and $52A_H$), stabilize the conformation (54_H), or confer side chain specificity (53_H, 55_H, and 56_H) (**Fig. 1a** and **Supplementary Table 1**). Strikingly, a larger search of all Ab-antigen structures identified eight Abs that utilize this loop to bind an aspartate or glutamate in the antigen (**Supplementary Fig. 1b**).

To characterize this class of Ab-antigen interactions, we synthesized the gene encoding a humanized version of the 1i8i Fab and cloned this construct into both a phage display and protein expression vector (**Supplementary Table 2**). This humanized scaffold, which expressed

at yields > 3mg/L in bacteria, bound the peptide with similar affinity as reported for the mouse Fab[29]. To understand the importance of the Asp-loop (residues 52_H-56_H) interaction in peptide binding, we performed competition phage ELISAs to analyze Fab binding to a panel of peptides. ELISA data confirmed that the Asp8 residue of the antigen is a hot spot for binding as mutation to Ala, Ser, Thr, or Tyr substantially reduced Fab binding (>100-fold less) to the peptide (**Fig. 1b**). We reasoned that the carboxylate group of Asp8 residue might mimic a phosphorylated residue and thus, the Ab may bind peptides with pSer, pThr, or possibly pTyr in place of Asp8. ELISA data confirmed the ability of this Fab to bind pSer- or pThr-containing peptides, albeit with weak affinities (>2000 nM) (Fig. 1b and Table 1). No Ab binding was observed to the pTyr peptide probably due to its large size. Structural analysis of the peptide:Fab complex suggested that steric clashes with several side chains and the main chain of the CDR were likely responsible for the weak affinities.

Therefore, we constructed three Ab phage display libraries to optimize the CDR region for each phosphorylated residue. The six-residue CDR region (52_H-56_H) was replaced with six random residues (H2 library) or seven random residues (H2+1 library) to relieve steric clashes with the Ab backbone. The third library design was similar to the H2 library, but fixed Gly or Ser at 53_H and 54_H (GS library). These strategies allowed us to assess the importance of the anchor (52_H and $52A_H$) and conformation (55_H) residues as well as alter the specificity residues (53_H, 55_H, and 56_H). Using standard phage display methods, we then performed four rounds of selection against pSer, pThr, and pTyr peptides. Impressively, we observed strong enrichment against each of the pSer, pThr, and pTyr peptide targets using all three libraries, except for selections with the H2+1 library against pTyr (**Fig. 1c** and data not shown).

Characterization of PS Ab scaffolds

For each phosphopeptide antigen, we isolated single phage clones and sequenced the CDR H2 region for clones that bound to the phosphopeptide by single-point ELISA (data not shown). Selections against the pSer and pThr peptides gave similar sequences and thus were combined into one sequence logo. Sequence logos from the H2- and GS-library selections against pSer/pThr highlighted the conservation of the key anchoring residue $T52A_H$ and conformation residue $G54_H$ in the loop, whereas more diversity was observed in the specificity residues (55_H and 56_H) (**Fig. 2a**). Interestingly, in the H2+1 libraries, we observed a strong enrichment for a Pro-Arg insertion in place of $G53_H$ and conservation of $G54_H$ (**Fig. 2b**). The $G54_H$ residue occupies a region of the Ramachandran plot in which only glycine is allowed, thus suggesting that this glycine is critical for the conformation[29-31]. The pTyr Abs contained a different binding motif from the pSer/pThr Abs suggesting that the mode of pTyr recognition differs from that of pSer/pThr recognition (**Fig. 2c**).

Next, we analyzed the phage clones by competition ELISA to identify the best scaffold for each target (pSer, pThr, or pTyr) (data not shown). We identified a pSer-specific scaffold (pSAb with sequence ATGGHT), a pSer/pThr-specific scaffold (pSTAb with sequence STPRGST), and a pTyr-specific scaffold (pYAb with sequence VTGGRK). Interestingly, we were unable to isolate a pThr scaffold that did not cross-react with the pSer peptide. To determine the phospho-selectivity of these scaffolds, we analyzed binding to the phosphorylated and unphosphorylated peptides by ELISA and Biacore. Strikingly, we observed high affinity and selectivity for the phosphorylated peptide in all cases (**Fig. 2** and **Table 1**).

Structural analysis of phosphopeptide recognition

To explore the mode of phosphoresidue recognition, we determined the X-ray structure of four Fab:peptide complexes (pSAb:pSer, pSTAb:pSer, pSTAb:pThr, and pYAb:pTyr) as well as the unbound pYAb Fab (**Supplementary Table 4 and 5**). We observed strong electron density for the bound peptide in all pSer and pThr structures (**Supplementary Fig. 3**). For the pYAb Fab, only one of the two Fab copies in the asymmetric unit was fully occupied by the peptide, likely due to the packing arrangement of the Fabs (**Supplementary Fig. 3**). No changes in the positions of the CDRs were observed between the mouse[29] and humanized Abs (cα RMSD of 0.78 Å). Furthermore, binding of the peptide to the Ab did not induce any major CDR movements (cα RMSD of 1.3 Å) (**Supplementary Fig. 4**). For all phosphopeptides, the recognition is achieved through two sectors: the phosphoresidue-binding pocket and a neighboring peptide sequence "reader" region, which consists primarily of CDRs L3 and H3 (**Fig. 3e**). Additionally, all peptide:Ab contacts outside of the phosphoresidue also occur in the parent Fab (**Supplementary Fig. 4c**)[29].

Structures of the peptide:Fab complexes illustrate how CDR H2 specifically recognizes each phosphoresidue (**Fig. 3**). For all three scaffolds, mutations found in the parent H2 loop make the main chain more accessible, creating a large electropositive binding pocket (indicated by arrow in **Supplementary Fig. 5**). The phosphoresidue side chain is almost fully engulfed by the Ab in pSAb (80% buried) and pSTAb (92% buried) and anchored by multiple hydrogen bonds (**Fig. 3a-c**, and **Supplementary Table 6**). In pSAb, the pSer residue makes key contacts with specificity residues $G53_H$, $R55_H$, and $T56_H$, whereas in pSTAb, the pSer and pThr residues make key contacts with $R53_H$, $G54_H$, and $S55_H$. In pSTAb, the insertion of $P52B_H$ allows the $T52A_H$ anchor to flip out and still contribute a hydrogen bond from the main-chain carbonyl. In stark contrast, pYAb does not utilize the original designed loop conformation to bind pTyr (**Fig. 3d**). A

key ionic interaction with $K56_H$ and a hydrophobic interaction with $V52_H$ contribute to the recognition mode. Interestingly, the H2 nest pocket is occupied by a water molecule that is stabilized by the free C-terminus of the peptide, indicating that pYAb may bind differently to the pTyr residue in longer peptides without this neighboring free carboxylate (**Fig. 3d**). Combined, our *in vitro* characterization and X-ray crystal structures confirmed that we successfully designed novel Ab scaffolds that utilize pSer, pThr, or pTyr as hot-spot residues.

Generation of novel PS Abs using the pSer and pSer/pThr scaffolds

We hypothesized that an Ab library in which the phosphoresidue-binding pocket was conserved and "reader" regions were mutated would enable rapid generation of new PS Abs. Since every member of the initial library contains a phosphoresidue-binding pocket, each Ab should have a weak initial affinity for the phosphorylated antigen, dramatically enhancing the selection of new Abs. As a proof of principle, we targeted pSer- and pThr-containing antigens, as reagents capable of detecting these modifications are significantly lacking. We diversified surface-exposed positions in CDR H2 (50_H, 56_H, and 58_H) outside of the phosphate-binding pocket, CDR H3 (95_H-101_H), and CDR L3 (91_L-94_L, 96_L) (**Supplementary Table 7**).

We chose a set of ten biologically relevant pSer- or pThr-containing epitopes as target antigens (**Table 2**). As a stringent test, we did not perform counter-selections against the unphosphorylated antigens, since we reasoned that the binding pocket could be sufficient for selection of Abs that required the phosphorylated residue. We performed three rounds of selection and analyzed single phage clones from the third round of selection by single-point ELISA. Impressively, for seven targets, we isolated at least one scFv that bound only to the phosphorylated antigen (**Table 2** and **Fig. 4a**). To demonstrate the specificity of the isolated clones, we performed a panel of ELISAs to assay binding of each scFv to each of the ten

phosphorylated peptides (**Fig. 4b**). The data demonstrated the exquisite target selectivity of most scFv clones, indicating the absence of promiscuous pSer-/pThr-peptide binding scFvs. Western blot analysis confirmed that a sample set of Abs specifically recognized the corresponding phosphoprotein (**Fig. 4c**). Finally, the scFv-Fc fusions exhibited affinities ranging from 42 to 2430 nM (**Table 2**), which matches or exceeds previous reports of PS Ab affinities[18, 21].

3.4 Discussion

Here we described a novel, recombinant Ab generation method that entails the design of a motif-specific (e.g. pSer, pThr, or pTyr) Ab scaffold followed by structure-informed mutagenesis of the scaffold to generate monoclonal Abs against a panel of phosphopeptide antigens. The high success rate of our strategy (PS Abs against 7 of 10 targets), which does not employ counter-selections against the unphosphorylated epitope, demonstrates how the motif-specific pocket greatly improves the selection process, as even past Ab libraries generated from immunized animals required stringent counter-selections to enrich for PS Abs[21, 22]. In the case of pSAb and pSTAb, the pocket contains a hallmark α_L glycine at 54_H that contributes to the main-chain conformation of CDR H2. There is a remarkably high frequency of occurrence for this H2 conformation in Abs (~12% of all H2 conformations[31]) and multiple Ab structures with anionic molecules (e.g. aspartate, glutamate, or sulfate) bound at this site (Supplementary Fig. 1).

While our studies were in progress, the structure of a chicken scFv, which was generated from an immunized phage display library, was reported that utilized a similar H2 conformation to bind pThr-containing phosphopeptide[21]. Interestingly, a structural comparison of this chicken scFv with our Abs reveals that the phosphoresidue binds to the same H2 loop conformation albeit with a different hydrogen bonding pattern (**Supplementary Fig. 7**). This

strikingly similarity suggests there may be a germline-encoded anion-binding pocket capable of binding phosphate or sulfate groups. In fact, previous work on Abs that bind phospholipids suggested a "phosphate-binding subsite" that conferred recognition of only the phosphorylated or sulfated forms of multiple lipids and haptens[32]. Furthermore, anion-binding pocket-containing Abs may provide a protective role in the recognition of phosphorylated or sulfated antigens, such as lipid A in Gram-negative bacteria[32], or conversely, a more sinister role in autoimmune diseases, such as antiphospolipid syndrome[33]. Future crystallographic studies of these Ab:antigen complexes will better elucidate this possibility.

Interestingly, the main-chain dominated mode of pSer/pThr recognition is completely different from most endogenous pSer/pThr-binding domains such as SH2, 14-3-3, and FHA, that predominantly utilize side chains to bind the phosphoresidue[34] (**Fig. 3** and **Supplementary Fig. 6**). Only the WW domain sometimes utilizes two main-chain amides to bind a phosphate. In fact, our pSer/pThr scaffolds bind more efficiently to the phosphoresidue than naturally occurring domains by burying a larger surface area and contributing more hydrogen bonds (Supplementary Table 6). Others have recently suggested that these endogenous phosphoresidue-binding and other PTM-binding domains have evolved to bind shorter epitopes with moderate affinities to support the dynamic nature of signal transduction pathways, which potentially limits the range of epitopes they can bind[34-36]. Additionally, our designed PS pockets appear to function independently of the other CDRs as we could diversify those CDRs to target highly diverse phosphopeptides (**Fig. 4**).

Surprisingly, pYAb utilizes a completely different motif to recognize pTyr. It is notable that we achieved highly specific recognition of pTyr, despite not burying most of the pTyr phenyl ring (**Fig. 2c** and **3d**). However, we have yet to determine how the presence of the free

carboxylate, which stabilizes a water molecule in the nest, contributes to the binding affinity. We are currently developing new scaffolds in which most of the pTyr residue is buried and bound in a more nest-like region to boost the ligand efficiency and affinity.

Our bacteriophage-derived PS Ab platform, which can be automated, rapidly generates Abs within two weeks as opposed to the several months required for hybridoma methods. In stark contrast to traditional monoclonal or polyclonal PS Abs, our recombinant PS Abs utilize a single framework that permits high-level bacterial expression (> 3mg/L) and mammalian expression (~0.5-5 μg/mL media) in a renewable format. The use of a single framework greatly simplifies mutagenesis protocols (e.g. affinity maturation), sequence-function analysis, and conversion to other Ab formats (e.g. IgG).[17] Finally, we hypothesize that this motif-specific scaffold method should be generalizable to targeting virtually any antigen with a defined motif. Since many other PTM-binding motifs exist in nature, these motifs may be similarly designed into Abs to generate high-affinity monoclonal reagents capable of detecting other PTMs. Ultimately, the rapid *in vitro* generation of monoclonal anti-PTM antibodies will greatly enhance the study of PTMs throughout biology.

3.5 Acknowledgements

We thank members of the Wells lab for helpful discussions regarding this manuscript and Sam Pfaff for assistance with Biacore experiments. We thank Chris Waddling at the UCSF X-ray facility for assistance with generating protein crystals and James Holton, George Meigs, and Jane Tanamachi at the Advanced Light Source beam line 8.3.1 at the Lawrence Berkeley National Laboratory for help with collection of diffraction data. We thank the Court lab at the National Institutes of Health for generously providing the recombineering vectors. James Koerber is a Fellow of the Life Sciences Research Foundation and Nathan Thomsen is the Suzanne and Bob

Wright Fellow of the Damon Runyon Cancer Research Foundation. This work was supported by grants from the National Institutes of Heath (R01 CA154802 to JAW and GM54616 to WFD).

3.6 Methods

Vector construction

We constructed a series of p3 phage display vectors along with compatible protein expression vectors (**Supplementary Table 2**). We modified the human Fab template by Kunkel mutagenesis, according to standard protocols[37]. All restriction enzymes and DNA polymerases were purchased from NEB (Ipswich, MA). Oligonucleotides were purchased from IDT and all constructs were verified by DNA sequencing (Quintara Biosciences).

Generation of Phage Libraries

A humanized Fab in pJK1 with two stop codons within the CDR H2 was used as a template for Kunkel mutagenesis with oligonucleotides designed to correct the stop codons and introduce the designed mutations at each site[17, 37]. To make the H2-targeted libraries, we generated three libraries in which the codons encoding for the parent H2 sequence (STGGYN) was replaced with either i) six random amino acids encoded by NNK (H2 library), ii) seven random amino acids encoded by NNK (H2+1 library), or iii) a core set of two or three amino acids, which were allowed to be only Gly or Ser, and were flanked on both sides by two random amino acids encoded by NNK (GS library). Mutagenic oligonucleotides are listed in **Supplemental Table 3**. The resulting mutagenesis reactions were electroporated and phage were produced as previously described[17]. The final diversities of the H2, H2+1, and GS libraries were 6.5×10^9, 1.6×10^{10}, and 5.3×10^9, respectively.

To make the PS Ab libraries, we constructed two scFv templates, which consisted of either the pSAb or pSTAb variable light chain linked to the corresponding variable heavy chain

by a $(Gly_4Ser)_3$ linker and contained two stop codons in the CDR H3. These plasmids were then used as templates for Kunkel mutagenesis. The light chain CDR L3 (91_L-94_L, 96_L) and the heavy chain CDR H2 (50_H, 56_H, and 58_H) were diversified using degenerate codons designed to mimic the natural sequence diversity found at these positions (**Supplemental Table 7**)[17, 38]. CDR H3 was diversified using three to nine random amino acids (DVK) followed by three terminal residues (F/M, A/D, and Y) commonly observed in anti-peptide Abs. For the mutagenesis reactions, L3 oligonucleotides (P1 and P2) were mixed at a 1:1 molar ratio, H2 oligonucleotides (1, 2, and 3) were mixed at a 0.1:1:2 ratio and H3 oligonucleotides (PX.1 and PX.2, where X = CDR length) were mixed at a 2:1 ratio. The resulting libraries were produced using Hyperphage[39] to enhance recovery of rare binders and the final diversities of the pSAb and pSTAb libraries were 3.4×10^{10} and 2.7×10^{10}, respectively.

Phage Display Selections, ELISAs, and Western blots

All phage preparations, selections, and ELISAs were performed according to standard protocols (Supplemental Methods)[17]. Western blots with biotinylated scFvs were performed as described in Supplemental Methods.

Protein expression and purification

Selected Fabs were expressed in a protease-deficient C43 strain[40]. Expressed Fabs were purified from total cell lysates by Protein A, ion exchange, and gel filtration chromatography as previously described[17, 38]. Fabs were stored at 4°C for short-term analysis or flash frozen in 10% glycerol for storage at -80°C. ScFv-rFc constructs were transiently transfected into 293T cells and purified from the media using Protein A chromatography. Biotinylated scFvs contained a C-terminal biotin acceptor peptide and were co-expressed with BirA to enzymatically biotinylate each protein (pJK5). Nonphosphorylated versions of all peptides were fused to the C-terminus of

NusA, which contained an N-terminal His_6 tag and biotin acceptor peptide. Recombinant proteins were purified on a His GraviTrap column (GE Healthcare, Piscataway, NJ) followed by monomeric Avidin resin (Thermo Scientific, Rockford, IL) to a final purity of >95%. All biotinylated peptides were purchased from Elim Biopharmaceuticals (Hayward, CA) or Peptibody, Inc. (Charlotte, NC).

Biacore analysis

Surface plasmon resonance data was measured on a Biacore model 4000 (Biacore, Uppsala, Sweden). All proteins were in TBS containing 0.1mg/mL BSA and 0.01% Tween-20. A Biacore CM5 chip was coated with NeutrAvidin at ~3000 RU and biotinylated antigens were captured at <100 RU. Serial dilutions of the Fabs were flowed over the immobilized antigens and 1:1 Langmuir binding models were used to calculate the k_{on}, k_{off}, and K_D for each Fab:antigen pair.

Crystalization of peptide:Fab complexes

Fabs were expressed as described above and concentrated to 10-15 mg/mL in 10 mM Tris pH 7.5, 50 mM NaCl. Complexes of the Fab with the corresponding peptide were formed at a 1:2 molar ratio of Fab:peptide. Crystals were grown in hanging drop format by mixing 100 nL protein solution and 100 nL crystallization solution using a Mosquito nanoliter pipetting system (TTP Labtech). Crystals formed within one to two weeks at either 18°C or 4°C. Initially, the crystals we obtained for the Fabs bound to the pSer peptides diffracted very weakly. We therefore employed a microseeding strategy with a seed stock generated from finely ground pSTAb:pThr crystals in 50 uL cryoprotectant solution[41]. Crystals for the pSAb:pSer and pSTAb:pSer complexes were generated by hanging drop vapor diffusion with 300 nL drops consisting of 150 nL protein solution, 120 nL reservoir solution, and 30 nL 1:100 dilution of seed

stock. All crystals were soaked in cryoprotectant solution and flash frozen in liquid nitrogen. Crystallization conditions and cryoprotectant solutions are listed in **Supplementary Table 4**.

Diffraction data were collected using the Advanced Light Source beam line 8.3.1 at the Lawrence Berkeley National Laboratory (Berkeley, California) with a wavelength of 1.1 Å. The data were indexed, integrated, and scaled using ELVES[42] or HKL2000[43]. The structure of the pSTAb:pThr complex was solved by molecular replacement using Phenix[44]. The initial search model consisted of the variable heavy domain from 3n9g and the variable light domain, constant heavy domain, and constant light domain from 2gcy[45]. The pSTAb Fab structure was used as the search model for all other structures. Iterative rounds of model building and refinement were carried out with Phenix and Coot[46]. For isomorphous crystals, the same refinement test sets for calculating Rfree were used. Simulated annealing composite omit maps calculated using Phenix were used to remove model bias. After two rounds of refinement, peptides were built into each model using Coot. Riding hydrogens as implemented in Phenix were used in the final stages of refinement for the pSAb:pSer, pSTAb:pSer, and pSTAb:pThr complexes. Final refinement statistics can be found in **Supplementary Table 5**. The final coordinates were validated using MolProbity[47]. The final Ramachandran statistics (% Favored:% Outlier) were 98:0.2, 98:0.2, 98:0.2, 98:0, and 97:0.2 for pSAb:pSer, pSTAb:pSer, pSTAb:pThr, pYAb:pTyr, and pYAb, respectively. MacPyMol (DeLano Scientific) was used to generate structure figures. Electrostatic surfaces were calculated using APBS[48] and buried surface areas were calculated using CCP4[49].

Accession codes

The X-ray coordinates have been deposited in the Protein Data Bank for pSAb:pSer, pSTAb:pSer, pSTAb:pThr, pYAb:pTyr, and pYAb with accession IDs 4JFZ, 4JG0, 4JG1, 4JFX, and 4JFY, respectively.

3.7 Figures

Fab	Peptide	k_{on} (M^{-1} s^{-1})	k_{off} (s^{-1})	K_D (nM)
Parent	**WT Asp**	3.38×10^5	0.0032	9.6
	pSer	n.d.	n.d.	>2000[a]
	pThr	n.d.	n.d.	>2000[a]
	Ser/Thr	n.d.	n.d.	>2000[a]
pSAb	**pSer**	1.0×10^5	0.0075	71
	pThr	4.7×10^4	0.041	866
	Ser/Thr	n.d.	n.d.	>2000[a]
pSTAb	**pSer**	4.8×10^4	0.0082	172
	pThr	2.8×10^4	0.0064	232
	Ser/Thr	n.d.	n.d.	>2000[a]
pYAb	**pTyr**	1.9×10^5	0.070	360
	Tyr	2.84×10^4	0.249	8700

Table 1

Affinity measurements of Ab scaffolds as determined by Biacore. [a]No binding seen by competition ELISAs. Peptide sequences for WT, pSer, pThr, and pTyr are GEKKGNYVVTDH, GEKKGNYVVTpSH, GEKKGNYVVTpTH, and GEKKGNYVVTpYA, respectively.

Peptide	Sequence	Number of unique scFvs	Number of phosphospecific scFvs[a]	K_D (nM)[b]
P1: Caspase 3 (S12)	NTENSVDSK**pS**IKNLEPKII	5	0	n.d.
P2: RIPK3 (S227)	REVELPTEP**pS**LVYEAV	6	2	102 ± 15 (P2.A11)
P3: RIPK3 (S199)	LFVNVNRKA**pS**T ASDVYSF	23	17	250 ± 13 (P3.28)
P4: Smad2 (T8)	MSSILPF**pT**PPVVKRLL	3	2	78 ± 14 (P4.B9)
P5: CREB (S133)	RREILSRRP**pS**YRKILNDL	4	4	151 ± 8 (P5.G10)
P6: HtrA2 (S212)	RRRVRVRLL**pS**GDTYEAVV	21	21	2430 ± 150 (P6.C12)
P7: Akt1 (T308)	KEGIKDGATMK**pT**F	0	0	n.d.
P8: Akt1 (S473)	ERRPHFPQF**pS**YSASGTA	1	1	>5000[c] (P8.H9)
P9: PKC Θ (S695)	DQNMFRNF**pS**FMNPGMER	1	0	n.d.
P10: Sgk1 (S422)	EAAEAFLGF**pS**YAPPTDSF	4	4	42.2 ± 2.8 (P10.D6)

Table 2

Summary of scFv hits versus ten new phosphopeptide targets. [a]scFv clones that exhibited >5-fold higher ELISA signal against phosphorylated peptide compared to unphosphorylated peptide (Fig. 4). [b]As determined by competition ELISA with scFv-Fc protein (n = 2-3, error values represent standard deviation). Clone ID is shown in parentheses. [c]Only partial competition was observed at the concentrations of peptide used.

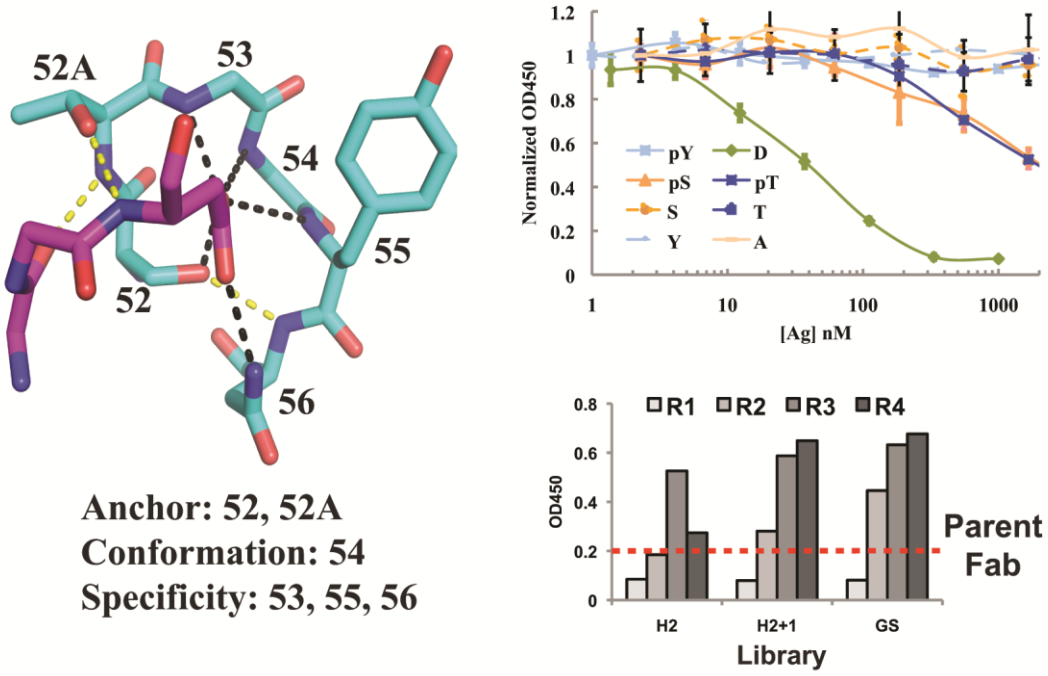


Figure 1

Design of phospho-specific Ab scaffold. a) Structure of CDR H2 loop from Ab (PDB ID 1i8i) bound to aspartate in peptide antigen[31]. Each H2 residue contributes to anchoring the peptide (52_H and $52A_H$), specificity (53_H, 55_H, and 56_H), or conformation (53_H). Hydrogen bonds that confer specificity are shown in black and anchoring hydrogen bonds are shown in yellow. The peptide is shown in magenta and Ab heavy chain is shown in cyan. b) Competition phage ELISAs with humanized Fab. Eight different mutant peptides containing D, A, S, T, Y, pS, pT, or pY at position 8 of the peptide were used as soluble competitors to inhibit Fab-phage binding to the immobilized wild-type peptide (KGNYVVTDH) (n=3, error bars represent standard deviation). Strong competition was observed for the wild-type peptide (green line), whereas no competition was observed for the S, T, A, or Y peptides (dashed lines) indicating that D is a hot-spot residue. Strikingly, the Fab binds to phosphorylated species as weak competition was

observed for the pSer and pThr peptides (orange and blue solid lines, respectively). c) Representative pooled phage ELISAs from selection of H2-targeted library against pSer peptide. After three rounds of selection, all library pools exhibited higher binding signal to the pSer peptide than the parent Fab (dashed line).

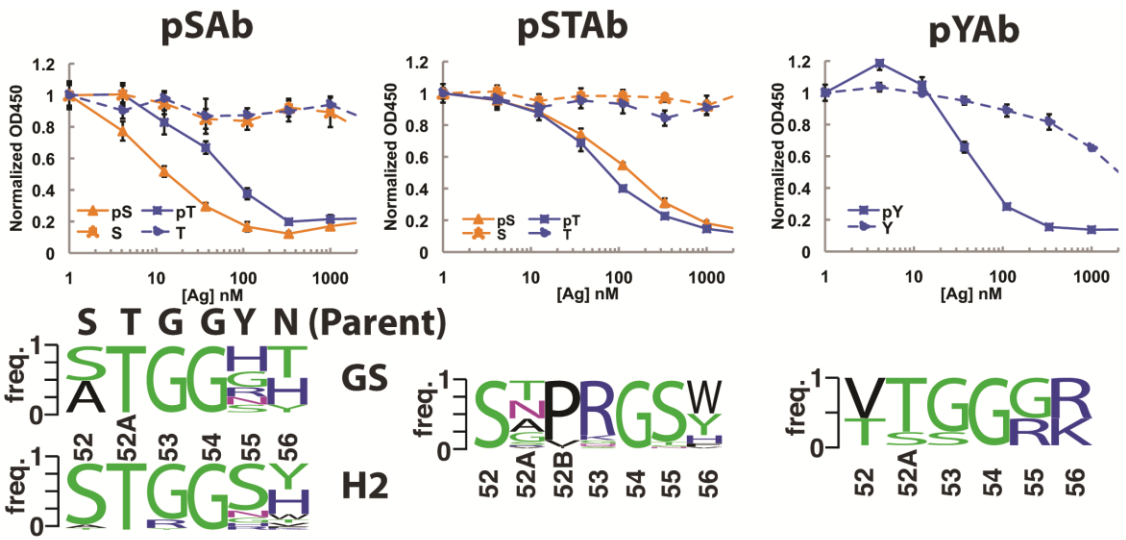


Figure 2

Selection and characterization of pSer-, pSer/pThr-, and pTyr-specific scaffolds. Competition ELISAs were used to determine the specificity of each Ab scaffold (n=3, error bars represent standard deviation). For both pSAb (a) and pSTAb (b), no binding inhibition was observed for the unphosphorylated peptides up to 2 μM, whereas strong inhibition was observed for the phosphorylated peptides. For pYAb (c), weak inhibition was observed at high concentrations of the unphosphorylated Tyr peptide, but ~20-fold less pTyr peptide was required to observe the same level of inhibition. The sequence frequency logos of the Ab pools from which each lead clone was derived are depicted in the bottom panels. GS and H2 indicate the sequence logos from GS and H2 libraries selected against pSer and pThr. For the six-residue loops selected for pSer or pThr binding, clear enrichment for the $G53_H$ and $G54_H$ is seen. For the seven-residue loops selected for pSer or pThr binding, we observed a replacement of $G53_H$ with Pro-Arg, likely opening up the binding pocket to better accommodate pThr. All clones that bound pTyr came from the six-residue libraries and contain two positively charged amino acids at H55 and H56. The H2 sequences of pSAb, pSTAb, and pYAb are ATGGHT, STPRGST, and VTGGRK, respectively.

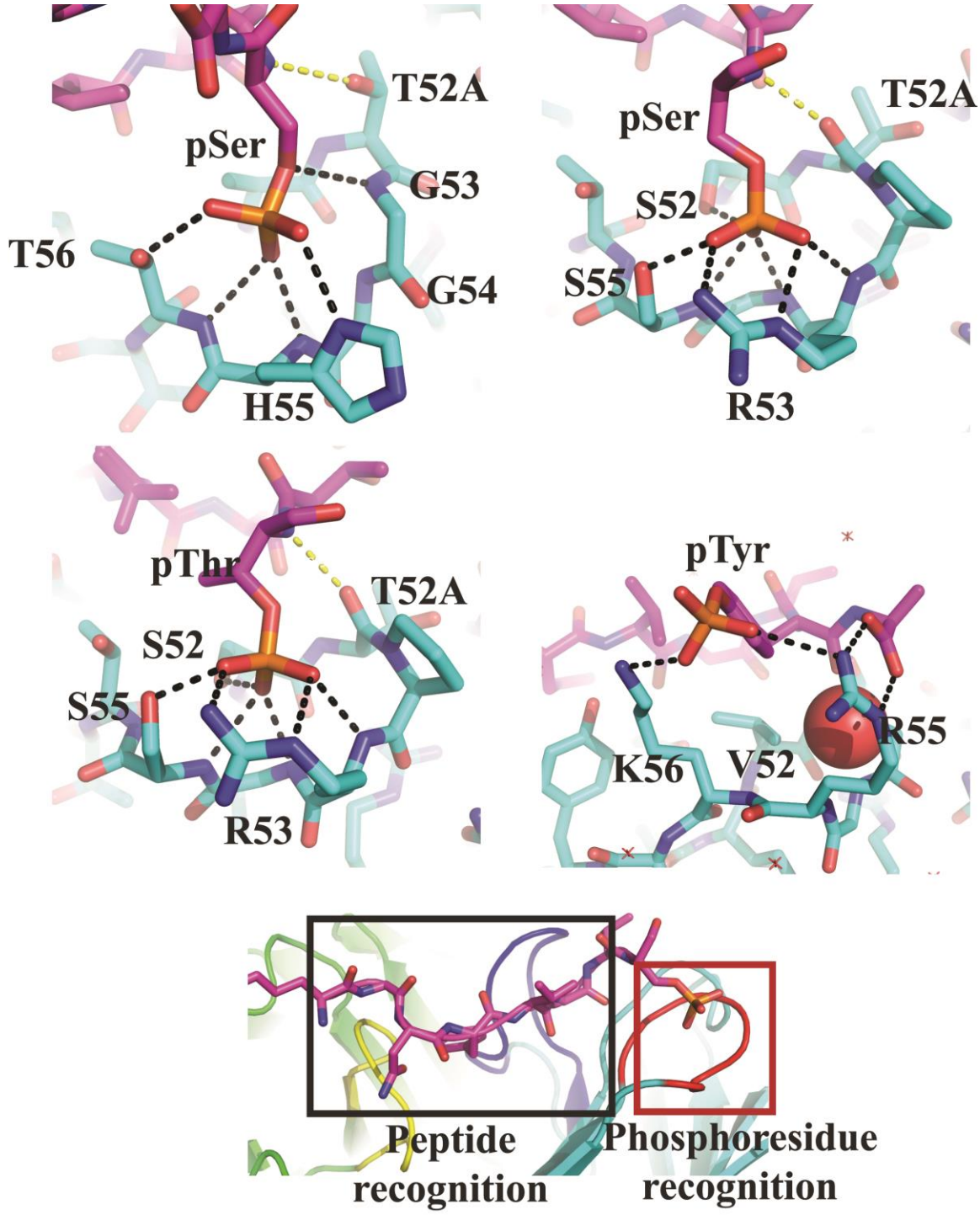

T52A
pSer
G53
T56
G54
H55
pSer
T52A
S52
S55
R53
pThr
T52A
S52
S55
R53
pTyr
R55
K56
V52
Peptide
recognition
Phosphoresidue
recognition

Figure 3

X-ray crystal structures of phosphoresidue-binding pocket from pSAb (a), pSTAb (b and c), and pYAb (d). a) In pSAb, pSer makes hydrogen bonds with all three specificity residues ($G53_H$, $H55_H$ and $T56_H$). The anchoring hydrogen bond (yellow) to $T52A_H$ is conserved. b and c) In pSTAb, the pSer/pThr makes hydrogen bonds with two specificity residues ($R53_H$ and $S55_H$), one anchor residue ($S52_H$), and the conformation residue ($G54_H$). In both pSTAb structures bound to pSer and pThr, $R53_H$ forms a bidentate interaction with the phosphate. The anchor residue $T52A_H$ is flipped compared to pSAb, which allows the backbone carbonyl to make a new anchoring hydrogen bond (yellow). d) The pTyr is recognized by a salt bridge with $K56_H$ and a hydrophobic interaction between $V52_H$ and the phenyl ring of the pTyr. However, the phosphate group of pTyr does not occupy the phosphate-binding pocket, which is instead occupied by a water molecule (shown as red sphere). e) The structures demonstrate two distinct recognition sectors: a phosphoresidue-binding pocket (red box) and the peptide-binding "reader" region (black box). Key CDRs L3, H2, and H3 are colored yellow, dark blue, and red. Phosphopeptides are shaded magenta and the Ab light and heavy chains are shaded green and cyan, respectively. Yellow and black dashed lines indicate hydrogen bonds between the phosphoresidue and Ab scaffold.

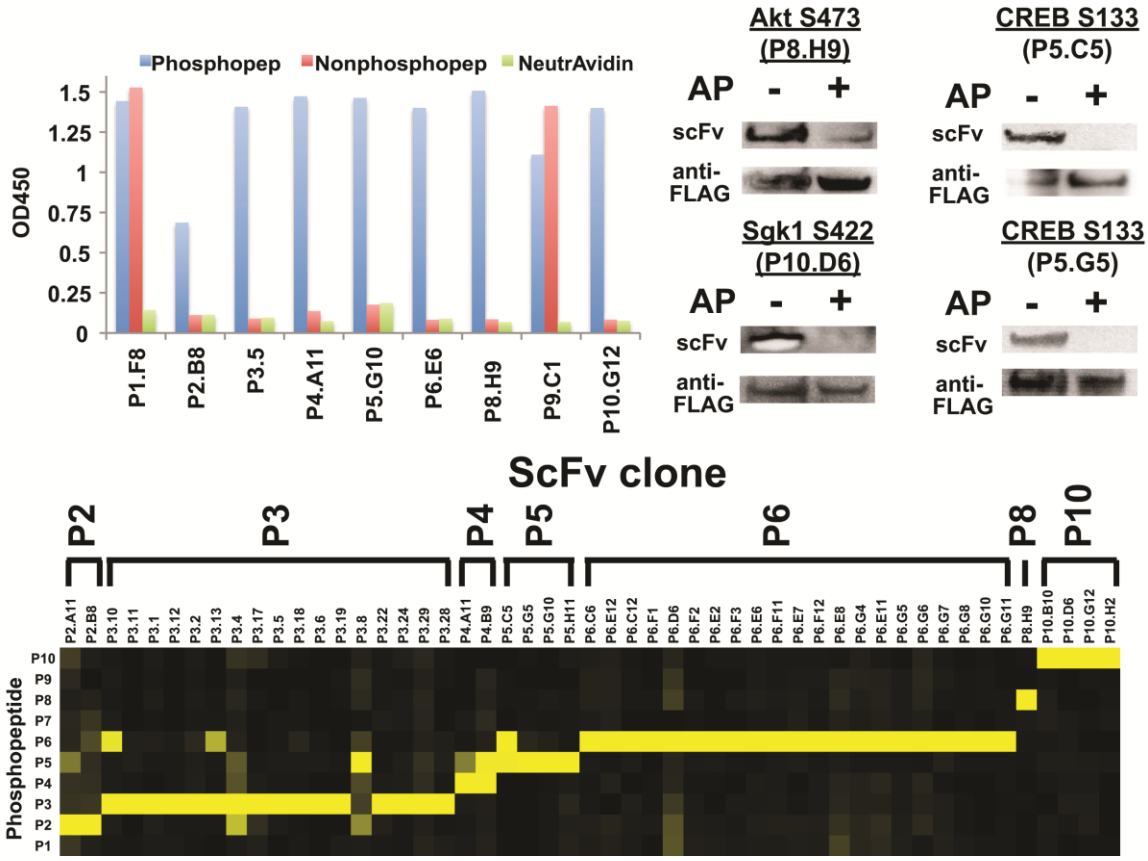


Figure 4

Generation of novel recombinant phospho-specific (PS) Abs using the pSAb and pSTAb scaffolds. a) Representative phage ELISAs of one scFv clone selected against each of the nine phosphopeptide targets demonstrates that we selected PS Abs to seven out of the ten targets. No hits were observed against P7. To analyze target specificity, we characterized the binding of each scFv-phage to ten different phosphopeptides by phage ELISA (n = 2 - 3) (b). Heatmap representation of the phage ELISA binding signals for each scFv-phage (vertical axis) against each of the ten phosphopeptides (horizontal axis). Strikingly, most of these scFvs bind only to the phosphopeptide against which they were selected. For each scFv, signals were normalized to the highest overall ELISA signal observed against the ten peptides. The scale goes from zero (black)

to one (yellow). c) ScFvs also recognize the phosphorylated protein in Western blots. FLAG-tagged target proteins were immunoprecipitated from transiently transfected HEK293T. To verify PS binding, samples were either dephosphorylated using alkaline phosphatase (AP) or treated with buffer only. Membranes were probed with biotinylated scFv (20 μg/mL) overnight and bound scFv was detected using NeutrAvidin-HRP. Total levels of target protein were monitored using anti-FLAG-HRP (Supplemental Methods).

3.8 Supplementary Figures

Heavy chain residue	Label	Function
52	Anchor	Accepts hydrogen bond from main-chain amide of 56; Donates hydrogen bond to carboxylate of Asp
52A	Anchor	Hydrogen bonds to main-chain amide of Asp; Potential hydrogen bond to phosphate
53	Specificity	Lack of side chain prevents steric clashes with Asp
54	Conformation	Critical α_L glycine
55	Specificity	Side chain can confer specificity and enhance binding
56	Specificity	Side chain can confer specificity and enhance binding

Supplementary Table 1

Functional description of H2 loop residues.

Vector	Type	Promoter	Description
pJK1	Phagemid with truncated g3	phoA	Displayed protein is fused to C-terminal domain of g3
pJK2	Phagemid with full-length g3	phoA	Displayed protein is fused to full-length domain of g3
pJK3	Protein expression in bacteria	T7	Expression under control of T7 promoter
pJK4	Protein expression in bacteria	pTac	Expression under control of pTac promoter
pJK5	Protein expression in bacteria	T7	Expression under control of T7; co-expression of BirA
pJK6	Protein expression in mammalian cells	hEFI-HTLV	Mammalian cell expression of protein fused to rabbit Fc

Supplementary Table 2

List of vectors utilized in this study.

Protein	Condition	Cryoprotectant solution	Temperature (ºC)
pSAb:pSer	23% PEG1500, 0.1M PCB pH 6.8	Mother liquor with 10% PEG200 and 25% PEG1500	4
pSTAb:pSer	22% PEG1500, 0.1M PCB pH 6.4	Mother liquor with 10% PEG200 and 25% PEG1500	4
pSTAb:pThr	25% PEG1500, 0.1M PCB pH 6	Mother liquor with 10% PEG200 and 25% PEG1500	4
pYAb:pTyr	20% PEG3350, 0.2M KCl	Mother liquor with 10% PEG200 and 25% PEG3350	4
pYAb	25% PEG1500, 0.1M MMT pH 4	Mother liquor with 10% PEG200 and 25% PEG1500	18

Supplementary Table 3

Crystallization and cryoprotection conditions for Fab complexes.

	pSTAb:pThr	pSTAb:pSer	pSAb:pSer	pYAb:pTyr	pYAb
Data collection					
Space group	P 2_1 2_1 2_1	P 2_1 2_1 2_1	P 2_1 2_1 2_1	P 3_2 2 1	P 3_2 2 1
Cell dimensions					
a, *b*, *c* (Å)	43.81, 95.59, 119.82	43.95, 95.89, 119.92	43.5, 94.87, 120.58	152.85, 152.85, 85.29	152.26, 152.26, 83.55
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 120	90, 90, 120
Resolution (Å)	50 – 1.55 (1.604 – 1.55)	74.89 - 1.81 (1.875 - 1.81)	74.56 – 1.75 (1.813 – 1.75)	50 – 1.95 (2.02 - 1.95)	76.13 – 2.63 (2.724 – 2.63)
R_{sym} or R_{merge}	0.065 (0.51)	0.119 (0.67)	0.113 (0.71)	0.097 (0.67)	0.115 (0.97)
I / σI	15.22 (2.39)	6.14 (1.90)	7.22 (1.90)	9.32 (2.11)	11.61 (1.86)
Completeness (%)	97.85 (86.93)	99.70 (99.48)	99.49 (99.33)	99.92 (99.81)	99.49 (95.97)
Redundancy	5.6 (2.9)	3.8 (3.8)	3.9 (3.9)	4.1 (4.1)	7.8 (5.4)
Refinement					
Resolution (Å)	50 – 1.55	74.89 - 1.81	74.56 – 1.75	50 – 1.95	76.13 – 2.63
No. reflections	72503 (3659)	46969 (2418)	50977 (2622)	83503 (4166)	33257 (1719)
R_{work} / R_{free} (%)	15.1 / 17.4	16.1 / 20.2	15.4 / 19.9	16.3 / 20.2	18.8 / 23.6
No. atoms					
Protein	3607	3458	3470	6764	6543
Ligand	5	5	5	26	52
Water	621	607	675	977	56
Wilson B-value (Å)	13.33	18.36	15.42	23.38	59.75
B-factors					
Protein	18	23	19	32	88.6
Water	30.6	33.7	30.7	38.9	68.7

R.m.s. deviations					
Bond lengths (Å)	0.01	0.003	0.01	0.009	0.007
Bond angles (°)	1.35	0.9	1.32	1.2	0.86
Ramachandran statistics (%)					
Favored	98	98	98	98	97
Outliers	0.2[3]	0.2[3]	0.2[4]	0	0.2[4]

[1]Values in parentheses are for highest-resolution shell.

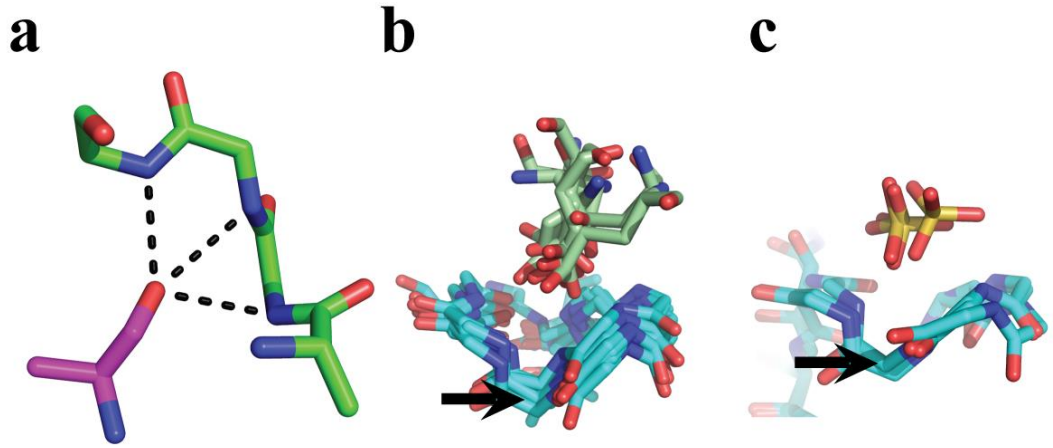
[2]Data was collected from a single crystal for each structure.

[3]Outlier residue ($Pro52B_H$) is the same in both structures with excellent density.

[4]Outlier residue ($Pro149_H$) is the same in both structures with excellent density in the high resolution structure

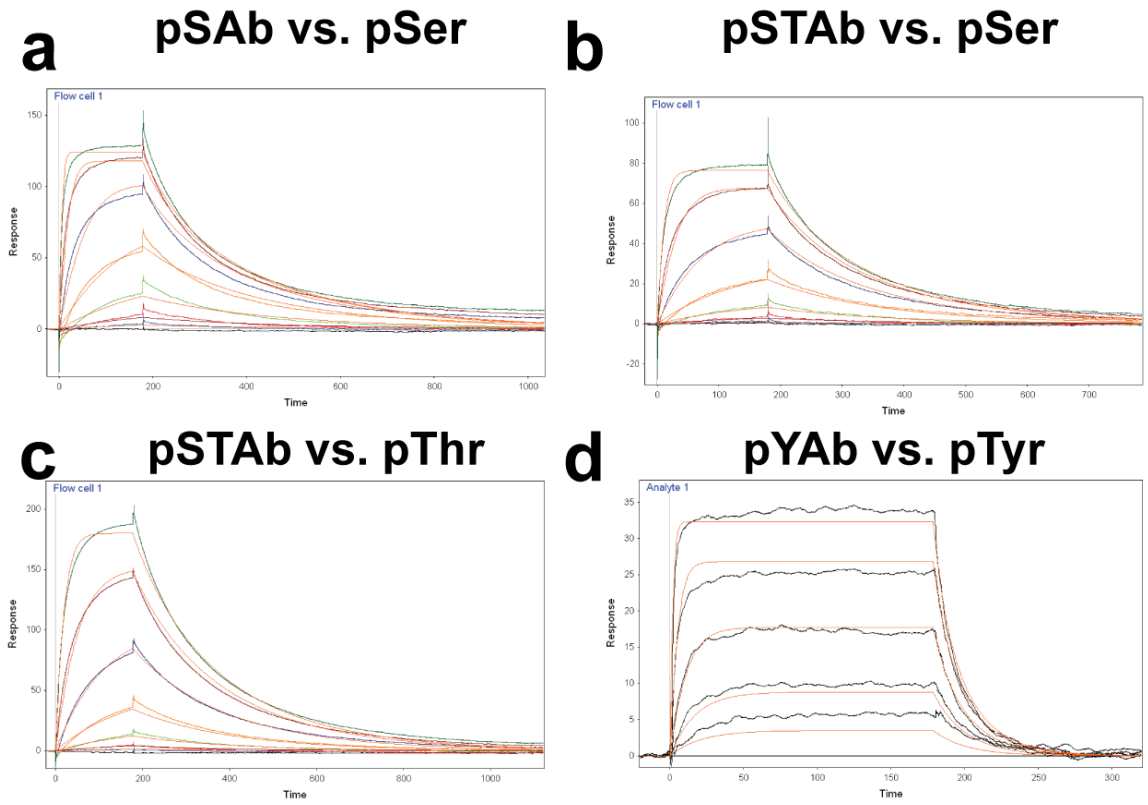
Supplementary Table 4

Data collection and refinement statistics (molecular replacement)



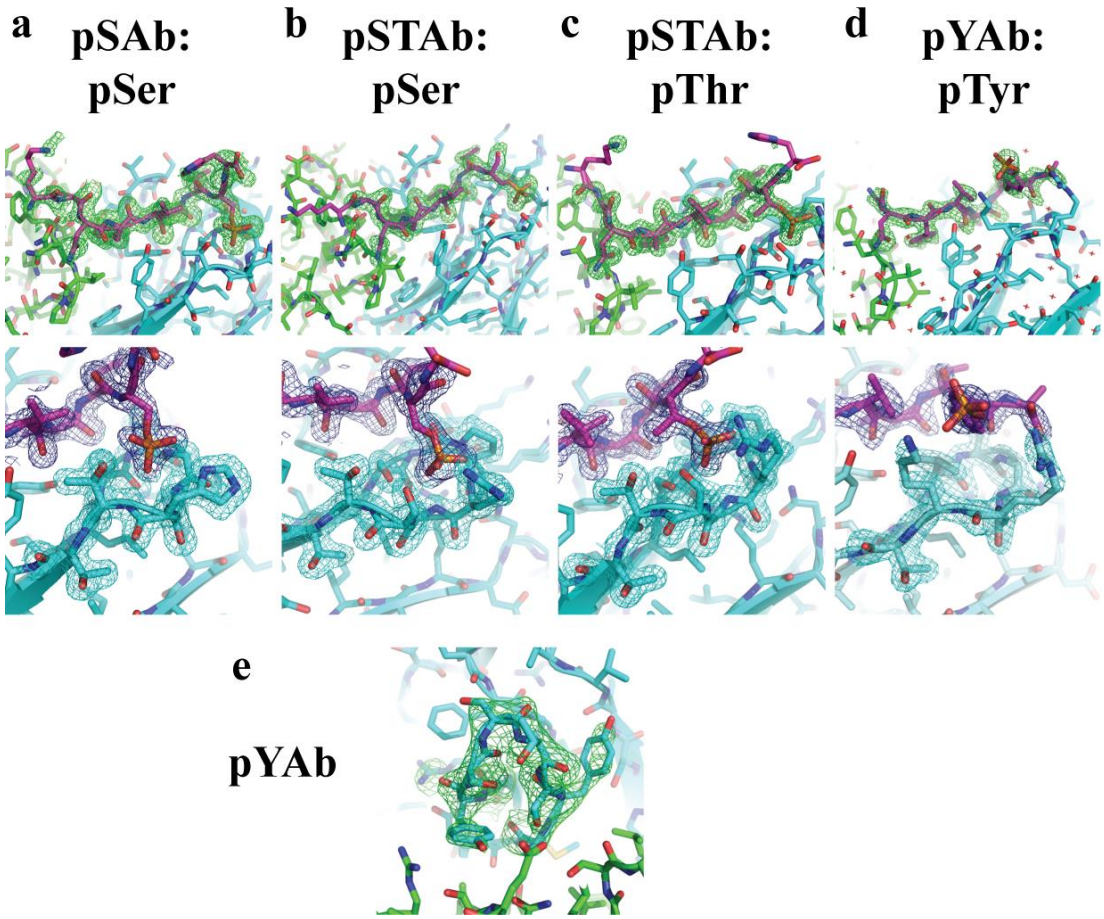
Supplementary Figure 1

Structure of nest motif in non-antibody and antibody scaffolds. a) Nest motif present in barnase[50], in which three consecutive main-chain amides contact the carbonyl group from a different residue (magenta). b) Structural alignment of CDR H2 bound to Asp/Glu. Alignment of CDR H2 region (50_H-56_H) from PDB ID 1i8i[29], 1frg[51], 2igf[52], 2qhr[53], 1dqj[54], 2nyy[55], 3bn9[56], and 3ffd[57]. All the antibodies contain $G54_H$ (indicated by arrow) and make at least two hydrogen bonds between the Asp/Glu antigen residue and main-chain NH groups. Asp and Glu residues are colored light green and CDR H2 is colored cyan. c) Alignment of the same CDR H2 region bound to sulfate ions from PDB ID 1seq[58], 2gsg[59], and 3vg0[60].



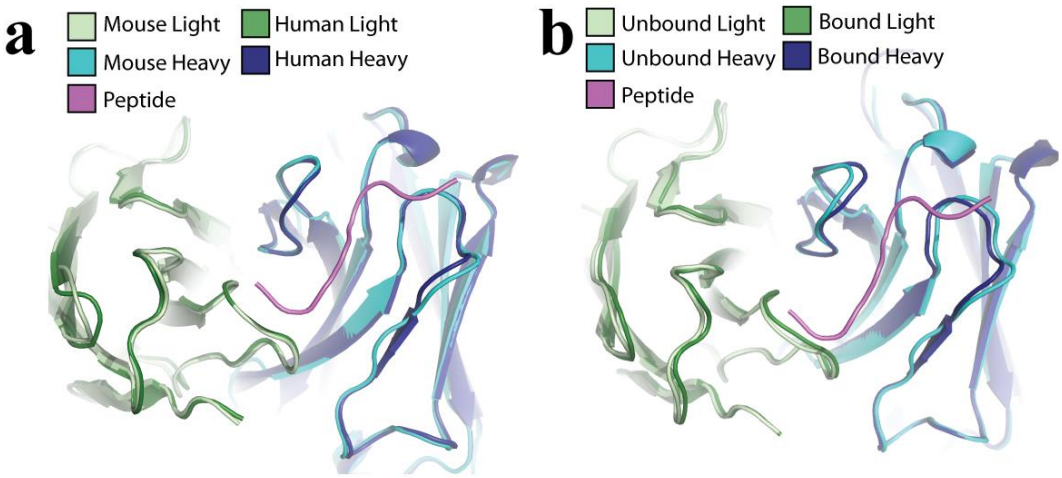
Supplementary Figure 2

Biacore traces of phospho-specific Fabs binding to phosphorylated peptides. a) pSAb binding to pSer peptide. b) pSTAb binding to the pSer peptide. c) pSTAb binding to the pThr peptide. d) pYAb binding to the pTyr peptide. Black lines represent the raw data and orange lines represent the best fit curves obtain from global fitting.



Supplementary Figure 3

Density maps of Fab structures. Strong electron density in pSAb:pSer (a), pSTAb:pSer (b), pSTAb:pThr (c), and pYAb:pTyr (d) complexes was observed for the peptide (top panels Fo-Fc maps) and for the phosphoresidue and CDR H2 loop (50_H -56_H) (bottom panels 2Fo-Fc maps). We observed weak density for the N-terminal lysine and C-terminal histidine in each peptide. Additionally, we observed good density for the unbound pYAb (e). Fo-Fc maps were contoured to 3σ and 2Fo-Fc maps were contoured to 1.25σ. The heavy chains are shaded cyan and the light chains are shaded green. Fo-Fc mesh for the peptide is shaded green. 2Fo-Fc mesh for the peptide is shaded dark blue and 2Fo-Fc mesh for CDR H2 is shaded light blue. The peptides are shaded magenta.

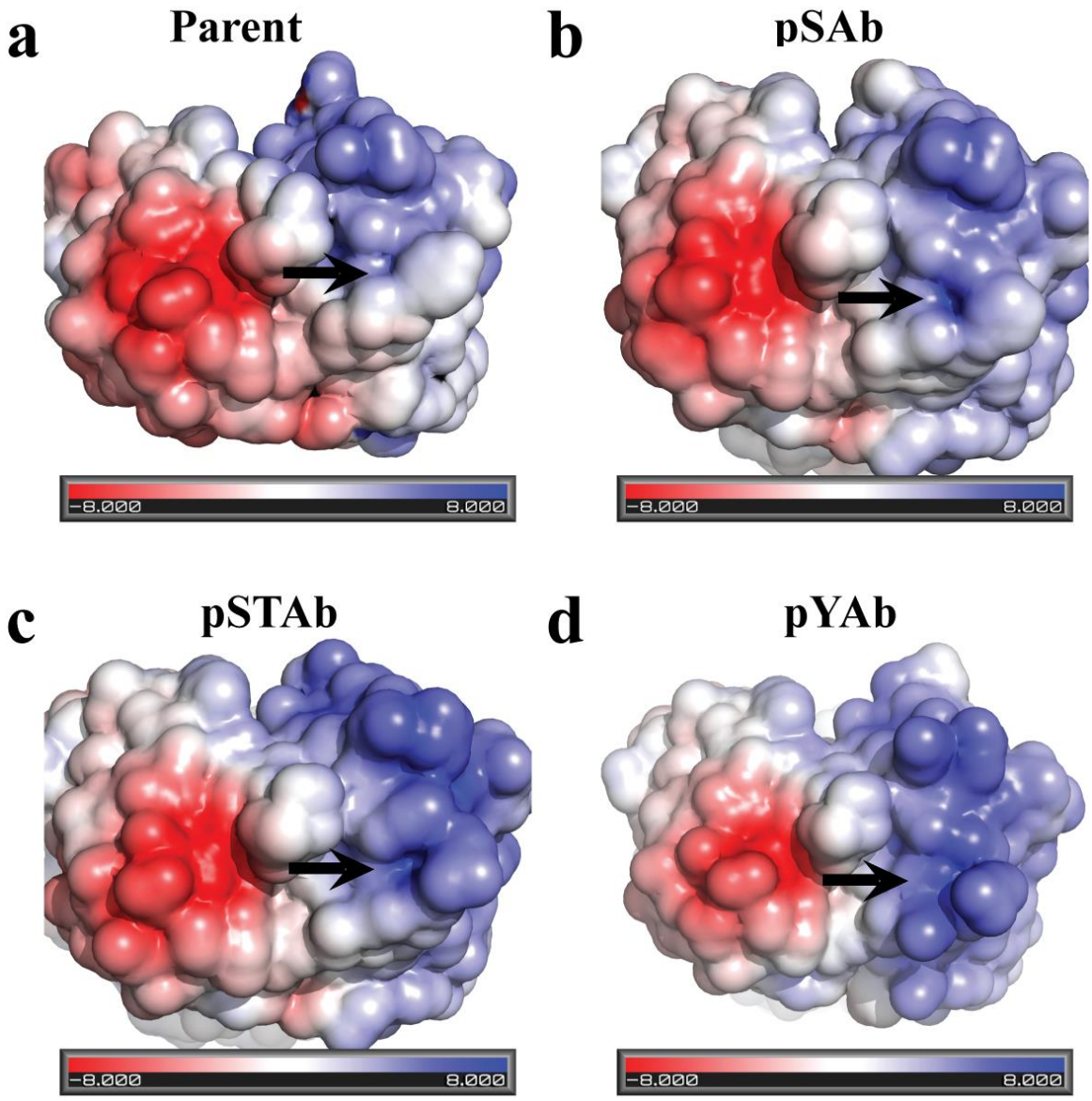


c

Peptide Residue	Main Chain H Bonds	Side Chain H Bonds	Hydrophobic
G2	$F92_L$	$N93_L$	
N3	$N93_L$, $V94_L$		
Y4	$S91_L$, $Y100A_H$		$V94_L$, $L96_L$, $Y100A_H$
V5		$Y58_H$, $Y100A_H$	
V6	$T52A_H$		$Y100A_H$
T7	$S98_H$		
pS/pT/pY8	See Figure 3		

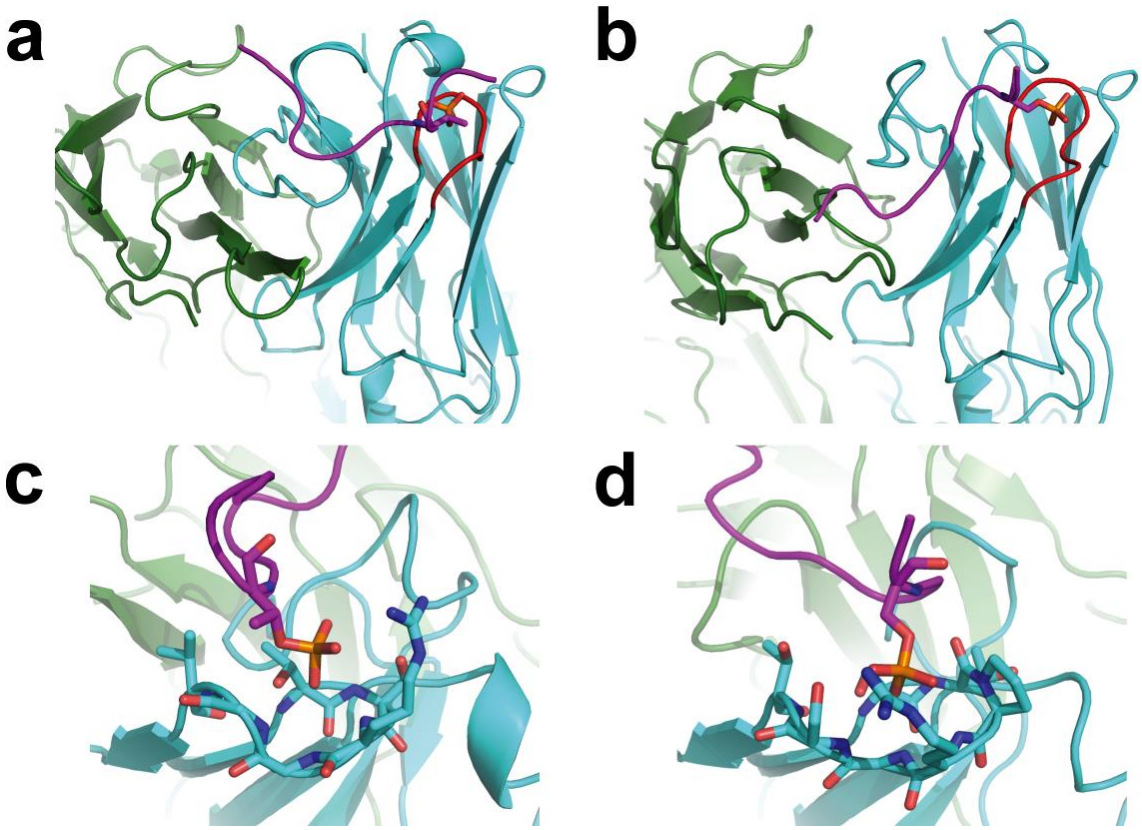
Supplementary Figure 4

Structural comparison between the mouse and humanized Fab (a) and the bound and unbound Fab (b). a) Alignment of the mouse Fab with the humanized Fab reveals no major deviations in the position of CDRs between the Fabs (cα RMSD of 0.78 Å). The light and heavy chains of the humanized Fab are colored dark green and dark blue, respectively. The light and heavy chains of the mouse Fab are colored pale green and cyan, respectively. The peptide is shown in pink. b) Comparison of the unbound and bound forms of pYAb reveals no major shifts in CDR position upon binding to the peptide (cα RMSD of 1.3 Å). The light and heavy chains of the bound Fab are colored dark green and dark blue, respectively. The light and heavy chains of the unbound Fab are colored pale green and cyan, respectively. c) List of all contacts between peptide residues, not including the phosphoresidue, and antibody. All of these contacts are conserved among pSAb, pSTAb, pYAb, and the parent 1i8i Fab[29].



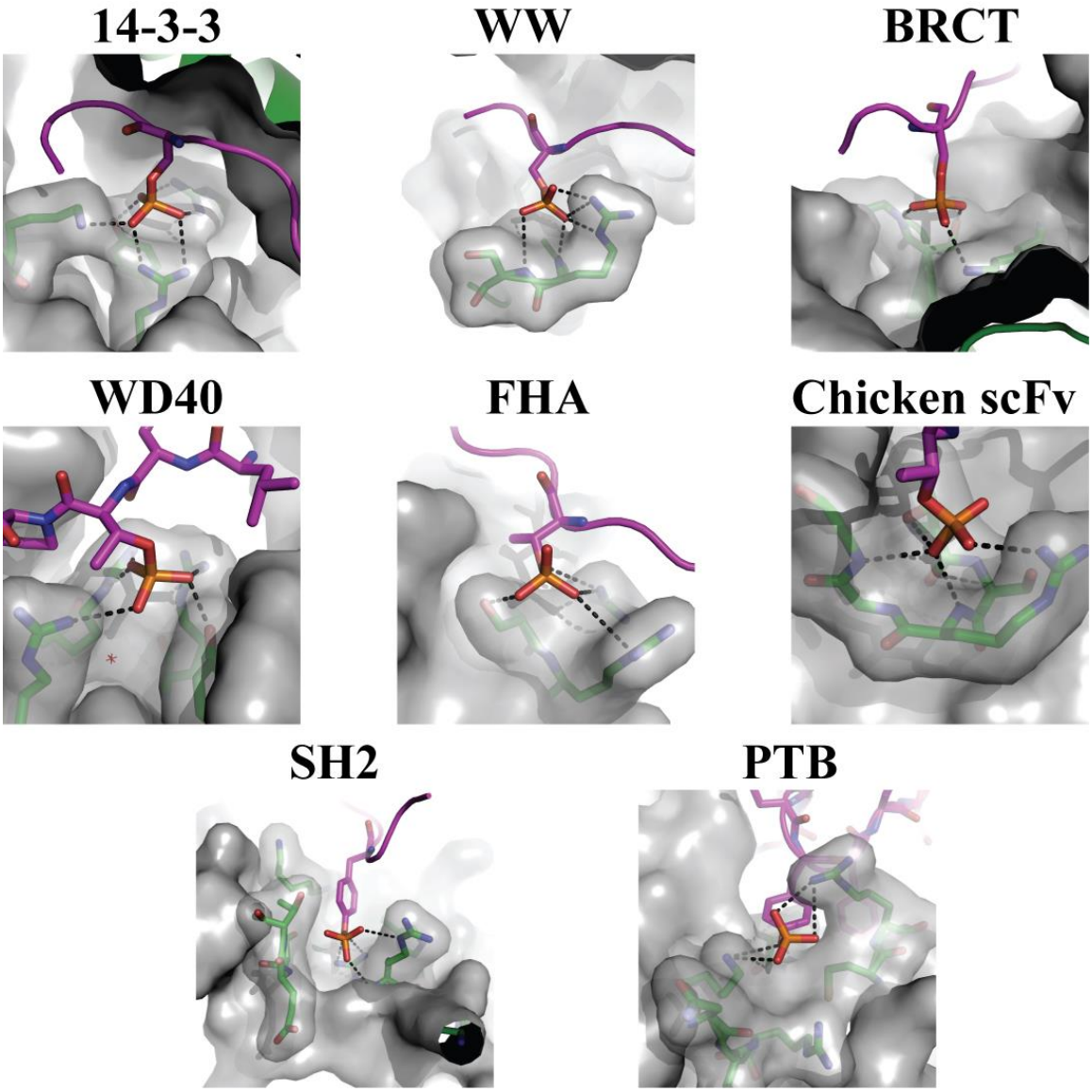
Supplementary Figure 5

Electrostatic surface representations of parent Fab (a), pSAb (b), pSTAb (c), and pYAb (d). Analysis of phosphate-binding pocket (indicated by arrow) in CDR H2 reveals larger electropositive pocket for all the phospho-specific scaffolds compared to the parent Fab. Surfaces were calculated with APBS and generated with MacPymol.



Supplementary Figure 6

Comparison between the natural PS chicken scFv and designed pSTAb structures. For the chicken scFv, the heavy chain contributes a majority of the contacts with the phosphopeptide (a), whereas for the pSTAb Fab the light and heavy chains contribute a similar number of contacts with the phosphopeptide(b). (c) The chicken H2 loop (sequence = TSRGG) binds the side of the pThr residue using hydrogen bonds with the $T52_H$ side chain and several main chain amides. $R53_H$ contributes an electrostatic component. (d) Our H2 loop (sequence = STPRGS) engulfs more of the pSer residue using multiple hydrogen bonds and a bidentate electrostatic interaction with $R53_H$. Phosphopeptides are shown in magenta and the light and heavy chains are shown in green and light blue, respectively.



Supplementary Figure 7

Phosphoresidue-binding pocket from natural phosphopeptide-binding domains. The structures highlight several distinct motifs used to bind the phosphoresidues. In all structures, at least one Lys or Arg makes a salt bridge with the phosphoresidue. The CDR H2 pocket from a scFv isolated from an immunized chicken that binds to a pThr-containing peptide is also shown[21].

Phosphopeptides are shown in magenta and key protein domain side chains and main chains are shown in green. Representative structures for the 14-3-3, WW, BRCA1 C-terminus (BRCT), WD40, forkhead-associated (FHA), Src Homology 2 (SH2), and phosphotyrosine-binding (PTB) domains are from PDB ID 1ywt[61], 1f8a[62], 1t15[63], 1nex[64], 1j4l[65], 1a0t[66], and 1shc[67], respectively.

3.9 References

1. Cohen, P. The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem Sci* **25**, 596-601 (2000).

2. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).

3. Blagoev, B., Ong, S.E., Kratchmarova, I. & Mann, M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat Biotechnol* **22**, 1139-1145 (2004).

4. Zhou, H., Watts, J.D. & Aebersold, R. A systematic approach to the analysis of protein phosphorylation. *Nat Biotechnol* **19**, 375-378 (2001).

5. Hornbeck, P.V., Chabra, I., Kornhauser, J.M., Skrzypek, E. & Zhang, B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4**, 1551-1561 (2004).

6. Beausoleil, S.A. et al. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* **101**, 12130-12135 (2004).

7. Bendall, S.C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687-696 (2011).

8. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. & Nolan, G.P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523-529 (2005).

9. Brumbaugh, K. et al. Overview of the generation, validation, and application of phosphosite-specific antibodies. *Methods Mol Biol* **717**, 3-43 (2011).

10. Dopfer, E.P. et al. Analysis of novel phospho-ITAM specific antibodies in a S2 reconstitution system for TCR-CD3 signalling. *Immunol Lett* **130**, 43-50 (2010).

11. DiGiovanna, M.P. & Stern, D.F. Activation state-specific monoclonal antibody detects tyrosine phosphorylated p185neu/erbB-2 in a subset of human breast tumors overexpressing this receptor. *Cancer Res* **55**, 1946-1955 (1995).

12. Nita-Lazar, A., Saito-Benz, H. & White, F.M. Quantitative phosphoproteomics by mass spectrometry: past, present, and future. *Proteomics* **8**, 4433-4443 (2008).

13. Marks, J.D. et al. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J Mol Biol* **222**, 581-597 (1991).

14. McCafferty, J., Griffiths, A.D., Winter, G. & Chiswell, D.J. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552-554 (1990).

15. Kang, A.S., Barbas, C.F., Janda, K.D., Benkovic, S.J. & Lerner, R.A. Linkage of recognition and replication functions by assembling combinatorial antibody Fab libraries along phage surfaces. *Proc Natl Acad Sci U S A* **88**, 4363-4366 (1991).

16. Mersmann, M. et al. Towards proteome scale antibody selections using phage display. *N Biotechnol* **27**, 118-128 (2010).

17. Sidhu, S.S. et al. Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J Mol Biol* **338**, 299-310 (2004).

18. Feldhaus, M.J. et al. Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nat Biotechnol* **21**, 163-170 (2003).

19. Hanes, J., Schaffitzel, C., Knappik, A. & Pluckthun, A. Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. *Nat Biotechnol* **18**, 1287-1292 (2000).

20. Cobaugh, C.W., Almagro, J.C., Pogson, M., Iverson, B. & Georgiou, G. Synthetic antibody libraries focused towards peptide ligands. *J Mol Biol* **378**, 622-633 (2008).

21. Shih, H.H. et al. An ultra-specific avian antibody to phosphorylated tau protein reveals a unique mechanism for phosphoepitope recognition. *J Biol Chem* **287**, 44425-44434 (2012).

22. Vielemeyer, O. et al. Direct selection of monoclonal phosphospecific antibodies without prior phosphoamino acid mapping. *J Biol Chem* **284**, 20791-20795 (2009).

23. Kaneko, T. et al. Superbinder SH2 domains act as antagonists of cell signaling. *Sci Signal* **5**, ra68 (2012).

24. Pershad, K., Wypisniak, K. & Kay, B.K. Directed evolution of the forkhead-associated domain to generate anti-phosphospecific reagents by phage display. *J Mol Biol* **424**, 88-103 (2012).

25. Malabarba, M.G. et al. A repertoire library that allows the selection of synthetic SH2s with altered binding specificities. *Oncogene* **20**, 5186-5194 (2001).

26. Clackson, T. & Wells, J.A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383-386 (1995).

27. Bogan, A.A. & Thorn, K.S. Anatomy of hot spots in protein interfaces. *J Mol Biol* **280**, 1-9 (1998).

28. Watson, J.D. & Milner-White, E.J. A novel main-chain anion-binding site in proteins: the nest. A particular combination of phi,psi values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions. *J Mol Biol* **315**, 171-182 (2002).

29. Landry, R.C. et al. Antibody recognition of a conformational epitope in a peptide antigen: Fv-peptide complex of an antibody fragment specific for the mutant EGF receptor, EGFRvIII. *J Mol Biol* **308**, 883-893 (2001).

30. Hollingsworth, S.A. & Karplus, P.A. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts* **1**, 271-283 (2010).

31. North, B., Lehmann, A. & Dunbrack, R.L., Jr. A new clustering of antibody CDR loop conformations. *J Mol Biol* **406**, 228-256 (2011).

32. Alving, C.R. Antibodies to liposomes, phospholipids and phosphate esters. *Chem Phys Lipids* **40**, 303-314 (1986).

33. Levine, J.S., Branch, D.W. & Rauch, J. The antiphospholipid syndrome. *N Engl J Med* **346**, 752-763 (2002).

34. Yaffe, M.B. & Smerdon, S.J. PhosphoSerine/threonine binding domains: you can't pSERious? *Structure* **9**, R33-38 (2001).

35. Kaneko, T., Joshi, R., Feller, S.M. & Li, S.S. Phosphotyrosine recognition domains: the typical, the atypical and the versatile. *Cell Commun Signal* **10**, 32 (2012).

36. Seet, B.T., Dikic, I., Zhou, M.M. & Pawson, T. Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* **7**, 473-483 (2006).

37. Kunkel, T.A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc Natl Acad Sci U S A* **82**, 488-492 (1985).

38. Bostrom, J. et al. Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. *Science* **323**, 1610-1614 (2009).

39. Rondot, S., Koch, J., Breitling, F. & Dubel, S. A helper phage to improve single-chain antibody presentation in phage display. *Nat Biotechnol* **19**, 75-78 (2001).

40. Thomsen, N.D., Koerber, J.T. & Wells, J.A. Structural snapshots reveal distinct mechanisms of procaspase-3 and -7 activation. *Proc Natl Acad Sci U S A* (2013).

41. Luft, J.R. & DeTitta, G.T. A method to produce microseed stock for use in the crystallization of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* **55**, 988-993 (1999).

42. Holton, J. & Alber, T. Automated protein crystal structure determination using ELVES. *Proc Natl Acad Sci U S A* **101**, 1537-1542 (2004).

43. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Method Enzymol* **276**, 307-326 (1997).

44. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).

45. Kaufmann, B. et al. Neutralization of West Nile virus by cross-linking of its surface proteins with Fab fragments of the human monoclonal antibody CR4354. *Proc Natl Acad Sci U S A* **107**, 18950-18955 (2010).

46. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132 (2004).

47. Chen, V.B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010).

48. Baker, N.A., Sept, D., Joseph, S., Holst, M.J. & McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-10041 (2001).

49. Winn, M.D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235-242 (2011).

50. Martin, C., Richard, V., Salem, M., Hartley, R. & Mauguen, Y. Refinement and structural analysis of barnase at 1.5 A resolution. *Acta Crystallogr D Biol Crystallogr* **55**, 386-398 (1999).

51. Churchill, M.E. et al. Crystal structure of a peptide complex of anti-influenza peptide antibody Fab 26/9. Comparison of two different antibodies bound to the same peptide antigen. *J Mol Biol* **241**, 534-556 (1994).

52. Stanfield, R.L., Fieser, T.M., Lerner, R.A. & Wilson, I.A. Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 A. *Science* **248**, 712-719 (1990).

53. Lee, J.E. et al. Complex of a protective antibody with its Ebola virus GP peptide epitope: unusual features of a V lambda x light chain. *J Mol Biol* **375**, 202-216 (2008).

54. Li, Y., Li, H., Smith-Gill, S.J. & Mariuzza, R.A. Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63(,). *Biochemistry* **39**, 6296-6309 (2000).

55. Garcia-Rodriguez, C. et al. Molecular evolution of antibody cross-reactivity for two subtypes of type A botulinum neurotoxin. *Nat Biotechnol* **25**, 107-116 (2007).

56. Farady, C.J., Egea, P.F., Schneider, E.L., Darragh, M.R. & Craik, C.S. Structure of an Fab-protease complex reveals a highly specific non-canonical mechanism of inhibition. *J Mol Biol* **380**, 351-360 (2008).

57. McKinstry, W.J. et al. Structural basis for antibody discrimination between two hormones that recognize the parathyroid hormone receptor. *J Biol Chem* **284**, 15557-15563 (2009).

58. Covaceuszach, S., Cattaneo, A. & Lamba, D. Neutralization of NGF-TrkA receptor interaction by the novel antagonistic anti-TrkA monoclonal antibody MNAC13: a structural insight. *Proteins* **58**, 717-727 (2005).

59. Li, P. et al. The structure of a polyQ-anti-polyQ complex reveals binding according to a linear lattice model. *Nat Struct Mol Biol* **14**, 381-387 (2007).

60. Carpenter, B. et al. Structure of the human obesity receptor leptin-binding domain reveals the mechanism of leptin antagonism by a monoclonal antibody. *Structure* **20**, 487-497 (2012).

61. Wilker, E.W., Grant, R.A., Artim, S.C. & Yaffe, M.B. A structural basis for 14-3-3sigma functional specificity. *J Biol Chem* **280**, 18891-18898 (2005).

62. Verdecia, M.A., Bowman, M.E., Lu, K.P., Hunter, T. & Noel, J.P. Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat Struct Biol* **7**, 639-643 (2000).

63. Clapperton, J.A. et al. Structure and mechanism of BRCA1 BRCT domain recognition of phosphorylated BACH1 with implications for cancer. *Nat Struct Mol Biol* **11**, 512-518 (2004).

64. Orlicky, S., Tang, X., Willems, A., Tyers, M. & Sicheri, F. Structural basis for phosphodependent substrate selection and orientation by the SCFCdc4 ubiquitin ligase. *Cell* **112**, 243-256 (2003).

65. Byeon, I.J., Yongkiettrakul, S. & Tsai, M.D. Solution structure of the yeast Rad53 FHA2 complexed with a phosphothreonine peptide pTXXL: comparison with the structures of FHA2-pYXL and FHA1-pTXXD complexes. *J Mol Biol* **314**, 577-588 (2001).

66. Mulhern, T.D., Shaw, G.L., Morton, C.J., Day, A.J. & Campbell, I.D. The SH2 domain from the tyrosine kinase Fyn in complex with a phosphotyrosyl peptide reveals insights into domain stability and binding specificity. *Structure* **5**, 1313-1323 (1997).

67. Zhou, M.M. et al. Structure and ligand recognition of the phosphotyrosine binding domain of Shc. *Nature* **378**, 584-592 (1995).

Chapter 4

Using designability to design a protein binder to hemagglutinin

4.1 Introduction

The design of protein binders to target specified epitopes is both highly desirable and incredibly difficult. It is highly desirable because so much of the life of the cell relies on protein interactions. If a protein engineer could design a protein to target an arbitrary epitope, she could disrupt or stabilize known protein-protein or protein-peptide interactions, tag and monitor a particular protein of interest, lock a protein into a particular confirmation, impede a conformational change, or even create an artificial antibody which would recruit other proteins or cells to its target. It's incredibly difficult because of the many degrees of freedom allowed in the design process. After selecting the epitope to be targeted, the designer still has to determine a protein backbone scaffold to be used, how to orient that scaffold with regard to the protein to be bound, what amino acid sequence will allow the scaffold to take on its given fold while also presenting amino acids which will drive the interaction, and finally how those amino acid side-chains will be presented.

One approach to addressing these difficulties is to choose a scaffold based upon a known interacting partner, or a homolog to an interacting partner, and then use computational repacking tools to improve or alter the binding affinity or specificity (1) (2) (3) (4). These approaches have the benefit of decreasing the search space for finding a binder by starting with a scaffold that is known to bind to a similar protein, allowing the researcher to concentrate on redesigning the amino acids at the binding interface.

In many instances, however, there may not be known interacting partners or it may be desirable to start with a novel protein framework. Recently, Fleishman et al. published a

protocol that allowed them to design a novel binder with nanomolar affinity to a conserved region of the influenza hemagglutinin protein (5). Their approach begins by designing key "hot-spot" residues that will provide a significant portion of the driving force for the interaction. They then search through a database of small, easily expressed proteins to identify members that can be oriented in such a way as to accommodate these hot-spots while also offering a degree of surface complementarity to the binding target. Hot-spot residues were then incorporated into compatible scaffolds, and additional scaffold residues were redesigned to be more compatible with the new binding interface. Of eighty eight designs selected for testing in a yeast-display assay, they were able to identify two modest binders (one with a K_d of 200 nM and the other weaker than their quantification limit), which were then strengthened to the nanomolar range via affinity maturation. A somewhat similar approach of "hot-spot" centric design was used to create a novel protein-protein interface which was matured to have a K_d of 180 pM (6).

In the search for appropriate scaffolds upon which to graft these hot-spots, we believe insufficient attention is paid to identifying scaffolds which would create interfaces that more closely mimic natural protein-protein or protein-peptide interfaces. While the possible geometry of secondary structure packing is virtually infinite, it has been shown that nature chooses a limited subset of possible geometrical arrangements in protein design (7) (8) (9) (10) (11) (12) (13). Moreover, many of the common structural motifs regularly seen in nature can accommodate a wide variety of amino acid sequences, an observation that has led to these structural motifs being called "designable." (14) (15) These observations – that some structural motifs are over-represented in nature and can accommodate a variety of amino acid sequences – can be used to focus the search for protein or peptide binder scaffolds. For instance, Yin et al.

describe the design of a transmembrane peptide which binds two closely related integrins by using structures from a database of transmembrane, helix-helix interactions as scaffolds (16). After threading the known integrin sequence onto one helix, they then redesigned the other helix partner, and were thus able to take advantage of known helix packing geometries favored by nature. Similarly, the concept of designability was used to design a peptide binder to carbon nanotubes by using knowledge of preferred coiled-coil geometries to design a scaffold which would wrap a carbon nanotube of a given radius (17).

Here we propose a general strategy of explicitly incorporating designability into the design of protein binders. As a test platform for our approach, we choose to design a peptide or protein binder to the same conserved stem region of the influenza hemagglutinin protein which Fleishman et al. targeted. We believe that we will be able to decrease the number of designs screened in order to obtain a modest binder by purposefully choosing scaffolds which would allow us to recapitulate secondary structure interaction geometries favored by nature.

4.2 Results

Overview of design approach

Our design approach begins with the identification of an epitope to target on the protein of interest. Epitopes can be chosen based upon a wide variety of criteria including: level of conservation (18), proximity to an enzymatic active site (19), known structural importance (16) (20), role in allosteric modulation (21), interactions with other binders (22), and identification by computational tools (23) (24) (25) (26). Once the epitope is chosen, a subset of key residues in that epitope are selected and used as a query to a structural search algorithm such as MaDCaT (27) or Suns (28). The number of residues in the query will vary depending upon the requirements of the search tool, but should be sufficient to identify key aspects of the

targeted epitope, such as secondary structure present and distance between structural elements. Next, the total number of similar motifs found in a non-redundant set of protein structures will give an indication as to how designable that interface is. If only a few proteins (on the order of five results from a database of 2000 chains) are identified that have similar motifs, then this interface is not very designable, and other residues within the chosen epitope should be queried. If no alternative subset of residues yields better results, then another epitope may need to be selected. If, on the other hand, dozens, if not hundreds or thousands of similar motifs are found in unrelated proteins, then this is a good indicator that the selected motif is designable, and the design process should continue.

Next, the full structures of the matches should be examined with the goal of identifying structural elements outside of the matching motif that are interacting partners with the motif. For instance, our motif may consist of a beta-strand packing against an alpha-helix. When examining matching motifs, we may observe that frequently there is a second alpha-helix which packs against this beta-strand/alpha-helix motif. In this way we expand our initial epitope motif of beta-strand/alpha-helix to include a potential binding partner – the second alpha-helix. Using the matching motifs to align the structures, these potential interacting partners are then placed in context of the original protein of interest. As the potential interacting partners are to be used as protein scaffolds upon which we will place amino acids to drive the interaction, we remove all side-chains from these structures. These potential scaffolds are then checked for steric clashes with the protein of interest, and are pared back to non-clashing, consecutive residues which are close to the desired epitope. These scaffolds, the segments which interact with our motif matches, can then be used to design peptide binders to our epitope, or alternatively we can attempt to place these scaffolds in the context of larger proteins which are known to be easily

expressible. We describe both approaches in our example binder-design targeting hemagglutinin.

Identifying peptide scaffolds for a hemagglutinin binder

We have chosen to design a binder to the same conserved stem region of hemagglutinin as Fleishman et al., facilitating comparison between the two design approaches. This region of hemagglutinin was originally selected due to its high degree of sequence conservation among all hemagglutinin subtypes, its proximity to the fusion peptide of hemagglutinin which is essential for influenza virus infectivity (29), and the fact that it is the target of a number of broadly neutralizing antibodies (30) (31) (32).

With our epitope selected, we next chose a subset of residues encompassing an alpha-helix and a neighboring parallel beta-strand to use as a query to the structural search program MaDCaT. Examining the results of the MaDCaT query, we looked for helical stretches outside of the matching segments which would be near our epitope of interest when the matching segments were superimposed. These nearby helices would be potential peptide scaffolds upon which to build our hemagglutinin binders. These potential peptide scaffolds were added to our original hemagglutinin query and these modified queries were then submitted to MaDCaT. The matches to the potential scaffold helices were extracted from this second round of MaDCaT results, and were used as an ensemble for redesign. A depiction of our procedure can be seen in **Figure 1**. From this approach, we chose two high-scoring designs to synthesize and test for binding to hemagglutinin (designs chaim_1 and chaim_2 in **Table 1**).

In an alternative approach, we modified our search to focus on solely the helical region of the hemagglutinin stem. Using an in-house database of helix dimers, we looked for helix pairs where the backbone of one helix would overlay closely on the backbone of the hemagglutinin

helix, while allowing the partner helix to rest in the hydrophobic groove of our epitope. From helix dimers which had one helix overlay exceptionally well with our hemagglutinin helix, we generated ensembles (**Figure 2**) and computationally redesigned the partner helices to interact with hemagglutinin as described in the methods. We chose three high-scoring designs to synthesize and test for binding to hemagglutinin (designs bth_1, bth_2, and bth_3 in **Table 1**).

Our five designs can be seen in **Figure 3**.

Peptide synthesis of designed binders

In our first round of synthesis, we incorporated two cysteine residues in each of our 5 designs. These cysteine residues, spaced 4 residues apart along the solvent-facing portion of each helix, allowed us to attach a chemical cross-linker to encourage helicity (33). Our cross-linker, dibromo-m-xylene, has a length approximately equal to the length between a residue at (i) and (i+4) on an alpha-helix, thus decreasing the entropic penalty the peptide must pay to form a regular secondary structure. Unfortunately, these peptides proved difficult to purify, and their spectra using circular dichroism (CD) spectroscopy lacked minima at 208 nm and 222 nm that is the characteristic signature of alpha-helices (**Figure 4**). Furthermore, when we tested for binding against hemagglutinin using bio-layer interferometry (34), no binding was detected. We hypothesized that the cross-linker increased the hydrophobicity of our peptide, making it prone to aggregation.

We therefore modified two of our designs to remove the cross-linker and used the helicity predictor Agadir (35) to help choose the solvent facing residues (**Table 1**). In addition, we made an additional peptide by solubilizing the interacting helix from one of Fleishman et al.'s successful designs in order to test whether the helix removed from its protein context would still bind. These peptides were significantly easier to purify and showed much improved helicity by

CD, especially the bth_1 design and the solubilized Fleishman-inspired helix. (**Figure 5**). Initial tests for binding using bio-layer interferometry looked promising as seen in **Figure 6a**. From 80 to 360 seconds, biotinylated hemagglutinin is loaded onto streptavidin covering the sensor of the instrument. From 540 to 720 seconds, our peptides flow across the attached hemagglutinin, and the increased intensity of the bth_1 design and the solubilized Fleishman helix indicates a change in the thickness of the biological sample, presumably due to the binding of our peptides. Note that the overall magnitude is quite small, as would be expected since the size of our peptides are miniscule compared to the immobilized hemagglutinin. Once the peptides stop flowing around 720 seconds, the intensity quickly falls back to the level seen prior, indicating that our peptides, if binding, have a fairly fast off-rate. At 900 seconds, we flow a mixture of our peptides and a known antibody which targets the same site. As can be seen, the signal is much stronger in this case as the antibody is significantly larger than our peptide. Also, after the antibody and peptides cease to flow at time 1080, the signal does dip slightly but still remains much higher than it was before. This may be indicative of our peptides falling off quickly while the antibody continues to bind.

As a negative control, we next tried flowing our peptides across the instrument without first loading hemagglutinin. Unfortunately, as can be seen in **Figure 6b**, we get a very similar signal, suggesting that the binding we saw in the presence of hemagglutinin was non-specific and not due to our designed interaction.

Phage display of peptide binders

We anticipate that binding to hemagglutinin may be weak for our starting designs, just as it was for the two Fleishman binders. Therefore, we decided to move our designs into a phage display system which allows us to test our initial design for binding and then improve

upon any weak interaction through the creation of libraries. We first modified our bth_1 design to include an additional two turns of the alpha-helix, allowing for an additional tryptophan to pack in the hydrophobic groove of hemagglutinin. We cloned DNA coding for this construct, bth_2, chaim_1, and the Fleishman-inspired helix into plasmids that contain the gene for the phage coat protein pVIII. The designs were inserted in-frame with the pVIII gene so that we could express chimeric proteins that would include our helical designs attached with a flexible linker to the pVIII coat protein.

Phage expressing our chimeric protein were then interrogated using an ELISA assay to look for possible binding to hemagglutinin. For each construct, we made two separate phage stocks. We did not induce expression of our chimeric protein in the first stock, relying on basal transcription to supply a low level of pVIII-helix design chimeras, while the majority of pVIII protein came from the wild-type sequence found in the supplied helper phage. The second stock was induced with IPTG, and thus we'd expect a higher proportion of pVIII-helix design chimeras to be present on each phage. We then presented each stock of phage to three different wells of a 96-well plate. The first well was coated with hemagglutinin (HA), the second well was coated with hemagglutinin and also contained an antibody known to target the same epitope we are targeting, and the third well did not have any hemagglutinin present.

If our peptides successfully target the proper hemagglutinin epitope, then we'd expect to see some binding signal in the first well, significantly less binding in the second well as the known antibody competes with our binders, and little to no binding in the third well which lacks hemagglutinin. Moreover, we'd expect the binding signal to increase as we move from the uninduced phage stock to the induced phage stock, as avidity effects should increase the apparent binding strength of any binder. Looking at the results shown in **Figure 7**, we can see

that both the bth_1 and bth_2 designs, and to a lesser extent the Fleishman inspired helix, show a pattern consistent with binding to our desired epitope. All three designs show a significant increase in signal when the well is coated in hemagglutinin (first row) compared to the well without hemagglutinin (third row). Moreover, the presence of an antibody which binds to the epitope we are targeting does appear to compete with the binding of our designs as seen by comparing the signal in the first and second rows. Finally, for these three constructs, the induced stocks of phage show stronger binding signals than the uninduced, which we'd expect if our peptides are involved in the binding interaction, since the induced stocks of phage should have higher effective concentrations of our peptide. Compared to the strength of binding observed for the known high-affinity antibody (ninth column), our designed binders appear to have much more modest affinity.

One puzzling observation from our initial ELISA is the observation that the binding strength of a pVIII chimera formed from a protein that was not known to bind hemagglutinin compares favorably to that of the high-affinity hemagglutinin antibody (tenth column). Interestingly, the binding strength appears to be independent of whether the high-affinity antibody is present or not (compare first and second rows), suggesting that it binds a different epitope of hemagglutinin. We had originally included this construct as a negative control so that we could observe the ELISA signal present for a phage pool without our designed binders. Since this construct fails in that regard, we performed the ELISA a second time, this time focusing on the two designs which showed the most promise (bth_1 and bth_2) and included a phage pool with a wild-type pVIII gene as a negative control. These results can be seen in **Figure 8**.

Interestingly, in this second ELISA, the differences between the induced and uninduced phage stocks disappear, with both the induced and uninduced results more consistent with the

induced stock in the previous assay. Moreover, the background binding to the well without hemagglutinin (third row) is higher than in the original assay. We are unable to explain these discrepancies. Of special note though is that the binding data for the wild-type pVIII follows a similar pattern to our two designs. The wells with hemagglutinin show a stronger binding signal than the wells without hemagglutinin (first and third rows), with the wells with hemagglutinin and the high-affinity antibody (second row) somewhere in between. One explanation for these results could be that the phage naturally sticks non-specifically to hemagglutinin. When the high-affinity antibody is present, less hemagglutinin surface area is exposed which is available for non-specific binding and consequently a drop in ELISA signal is observed.

Affinity maturation using phage libraries

As the results for the bth_1 and bth_2 designs looked especially promising in the first ELISA, we decided to attempt to affinity mature these using a phage library system. For each design, we made three libraries by introducing amino acid diversity at seven or eight positions in either the N-terminal portion, the middle portion, or the C-terminal portion, as seen in **Figure 9**. Each position was diversified by using a mixture of nucleotides that allowed us to keep the original, designed amino acid present at a frequency of between 34% and 55% and sample the other nineteen amino acids otherwise. In this way each library should be composed of at least 0.02% of the initial designed sequence (which for a modest library of size 10^9 would mean about 200,000 copies), ensuring that any new binders will need to compete favorably with the original weakly binding design to survive rounds of selection. Moreover, by dividing the constructs into thirds, every library member will keep a majority of the designed residues which drive the interaction to hemagglutinin, while sampling alternative amino acids at only one or two of the sites which make contact to hemagglutinin. The additional residues diversified in each library

are at sites designed to be solvent facing, allowing sequences that stabilize or relax the secondary structure in a way which is beneficial to binding to be sampled.

Each library was subjected to three rounds of selection against hemagglutinin. In each round, the phage library was incubated with magnetic beads with hemagglutinin attached. The beads were then washed, leaving only those phage which bound to the beads or hemagglutinin. Binders to just the beads should be rare as the libraries were subjected to a depletion round prior to selection as detailed in the methods section. Bound phage were eluted off of the beads and then amplified in a passage through bacteria. This amplified pool then became the phage library for the next round of selection.

After the final round of selection, we took the libraries obtained after each of the three rounds and subjected them to an ELISA assay. Each of the six libraries were interrogated in wells coated with hemagglutinin and wells coated with BSA in place of hemagglutinin. Ideally, what we'd like to see is the binding signal increase in the hemagglutinin coated wells as our strong binders come to dominate the library after subsequent rounds of selection. In the wells without hemagglutinin, we'd like to see a steady, low signal, indicating little to no non-specific binding by our libraries. As can be seen in **Figure 10** however, the binding signal remained low in all wells for all libraries. In fact, every library sees a drop in signal between the first and second round of selection, and most stay the same or drop again between the second and third rounds. Moreover, the signal for most libraries is remarkably similar for the wells coated with hemagglutinin and the wells coated with BSA. Rather than selecting a few specific, strong binders to hemagglutinin, it appears that our libraries consisted of weak, non-specific binders. This result unfortunately is consistent with our second ELISA performed on the initial designs (**Figure 8**) where the binding signal for our two designs was on par with the signal obtained from

phage with the wild-type pVIII gene, indicating a general trend of weak, non-specific binding for our phage.

Placing helical designs onto protein scaffolds

If our designed peptides bind hemagglutinin, we believe that the interaction is too weak for us to detect or to use as a starting point in our selection experiments. We note that the Fleishman-inspired helix, removed from its natural protein context, also fails to bind or binds too weakly for us to detect despite showing strong evidence of helicity (**Figure 5**) and maintaining the same key hot-spot residues from Fleishman's initial design and mutations found through the affinity maturation experiment. We hypothesize that the addition of a protein scaffold may facilitate the development of a binder by stabilizing the secondary structure of the binding interface and thus reducing the entropic cost of the binding event. Moreover, the protein scaffold may provide additional contacts which could be used in binding.

Therefore, we next searched for protein scaffolds upon which we could thread our helical peptide binders. We were interested in scaffolds that were small (< 150 residues), stable, and easily expressed in e. coli. Additionally, we required any protein scaffold to contain a surface exposed helix upon which we could thread our peptide designs. In the end, we chose to work with two scaffolds familiar to our lab: a designed thermo-stable version of GB1 (36) which consists of 57 residues, and the de novo designed, three-helix bundle α_3D (37), consisting of 73 residues. For our first round of experiments, we have decided to concentrate on the designed peptide bth_1. We threaded this helix onto all possible registries of the helices present in GB1 and α_3D, and selected two threadings from each scaffold to place into a phage display system. These four threadings, seen in **Figure 11,** were chosen because they avoided obvious clashes

between the scaffold and hemagglutinin and were able to accommodate the majority of the bth_1 peptide.

We inserted DNA coding for our new constructs in-frame with the gene coding for the pVIII phage coat protein as before. However, this time in addition to our construct, we also attached a FLAG-tag peptide sequence to the terminus of our design as seen in **Figure 12**. Using the FLAG-tag, we are able to monitor the expression level of our chimeric protein by performing an ELISA on our phage pool against an immobilized anti-FLAG antibody. By changing the signal peptide at the front of the construct, by altering the length and composition of the linker between our construct and the pVIII gene, or by changing whether our construct is placed at the N-terminus or C-terminus of the pVIII gene, we hope to be able to transform any low expressing constructs into high expressing constructs. Experiments are currently under way to tune these parameters to maximize expression.

4.3 Conclusions

The ability to design a de novo binder to a protein epitope of interest is a stringent test of our understanding of protein/protein and protein/peptide interactions. The two successful, designed protein binders to hemagglutinin described by Fleishman *et al.* are evidence that great progress has been made in our ability to design specific interactions which drive protein association. However, the failure to detect any binding activity for 86 of the 88 designs tested indicates that our current knowledge of protein interactions is incomplete and we need improved methods to increase the odds of a computational design being successful. We believe that the concept of “designability” promises to be such a method to increase the odds of success. By explicitly seeking to mimic interaction geometries that nature has repeatedly

selected for, we implicitly take advantage of the countless experiments carried out through evolution.

If designability offers such promise, what can account for our failure in harnessing it to design a binder to the same hemagglutinin epitope? We believe at least part of the cause can be found in our initial attempts to use a short peptide as our binder. In addition to the loss in entropy due to the packing of side chains at the binding interface, a penalty that must be paid by any binder, peptide binders also face an entropic loss due to the decrease in backbone flexibility. In one study, the loss in backbone flexibility cost peptide binders almost half of the free energy difference they would have realized had the peptide been completely rigid (38).

We initially attempted to mitigate the loss in peptide flexibility by adding a chemical staple to our peptides, constraining the distance between a pair of residues at i and i+4 to be the optimal distance for an alpha-helix. In this way, we hoped to encourage the peptide to naturally adopt a helical structure in solution, and therefore decrease the entropic penalty of forming a rigid helix upon binding. However, the chemical staple proved to significantly decrease solubility of our peptide, and we still did not measure strong helix formation in solution. We were more successful in choosing solvent-facing residues which encouraged peptide helicity in solution, but we anticipate these helices retain broad flexibility with few constraints to limit the distance between the 5' and 3' ends. It is particularly telling that the peptide we created based upon one of the Fleishman designs showed strong helicity and yet failed to provide detectable binding to hemagglutinin.

We chose a peptide as our binding scaffold because we wanted to design the "minimal" binder to our hemagglutinin epitope. From our results, we now believe the "minimal" binder would involve threading our peptide onto a protein scaffold. In this way, the loss in entropy is

paid for by the energetics involved in the folding of the protein. Strong constraints are placed on the distances between the two ends of the interacting segment by the scaffold in which it is placed. We have adopted this strategy for our designs and are currently involved in testing these constructs.

4.4 Materials and Methods

Initial scaffold search

For our computational design of hemagglutinin binders, we used the structure of hemagglutinin found in the Fleishman crystal structure PDB ID 3r2x, which comes from the 1918 H1N1 pandemic strain. For the bth_1, bth_2, and bth_3 designs, we concentrated on residues 41-56 of HA2 which form the helix in the hemagglutinin epitope recognized by a number of broadly neutralizing antibodies as well as the Fleishman binders. Using this helix as a query, we searched for similar helices in an in-house database of helices culled from a set of non-redundant PDBs specified by PISCES (38) (file: cullpdb_pc25_res1.5_R0.3_d100709_chains1486.1379.txt). Using BioPython's PDB module (39), we superimposed the backbone atoms from the hemagglutinin helix of HA2, residues 41-56, onto each 16 residue helix segment in our database. We noted each segment which superimposed with 0.75 Å or better RMSD to the backbone atoms of hemagglutinin, and looked for partner helices interacting with that matching segment in the native PDB structure. We saved helix partners that made substantial contact with the hemagglutinin-like helix and that would lie in the hydrophobic groove of our hemagglutinin epitope when superimposed. After visually inspecting the roughly 300 hits that met our criteria, we chose 14 examples for further analysis. Each of these helix/helix interactions was used as a query into a distance-map based structural search algorithm similar to MaDCaT (27), which then delivered an ensemble of

structures with similar geometries. Using the fixed backbone design program found in Rosetta (40), we then redesigned each helix in the ensemble to interact with the hemagglutinin helix rather than the native helix which aligned well to hemagglutinin. We examined the best scoring designs and chose three designs to synthesize.

For the chaim_1 and chaim_2 designs, a somewhat modified approach was used. In addition to the hemagglutinin helix formed from residues 41-56 of HA2, we also included the beta-strand residues found nearby on HA1 at positions 38-42. These two fragments were used as a query to the structural search program MaDCaT, which returned an ensemble of matches. A volume adjacent to the target was extracted from each match, and all segments of at least 10 contiguous residues within the volume are taken as potential binders. For the sake of peptide stability, we further restricted the potential binders to those that were primarily helical. These were clustered hierarchically by RMSD, and the best candidates were combined with the original target interface and used as the queries for a new round of MaDCaT searches. The results of this second search were used as starting ensembles for redesign. Candidates were chosen for redesign based on a combination of factors, including cluster size, ensemble size, and manual removal of obviously bad geometries.

In both cases, we used Rosetta to select the buried residues from a restricted alphabet that included only Alanine, Phenylalanine, Isoleucine, Leucine, Methionine, Valine, and Tryptophan. These choices allow for plenty of variation in size and shape to obtain optimal packing, while forestalling the possibility that Rosetta would insert an unfavorable polar group to satisfy a hydrogen bond. Rosetta was also allowed to optimize the rotamer of any residue on HA that was within 6 Å of the predicted position of the binder. From the top models computed by Rosetta, we chose from those predicted to bury at least one aromatic residue in the

hydrophobic groove of hemagglutinin, as we believe the burial of a large hydrophobic residue will significantly help drive our interaction.

Initially, the solvent exposed face was designed manually to create helix-stabilizing salt-bridges between lysines and glutamates, with glutamines filling in the remaining spots to maintain charge balance. In addition, two Cysteine residues were placed four residues apart on the solvent-exposed face of each helix to facilitate helix-stabilization via the addition of a dibromo-m-xylene cross-linker.

After encountering solubility issues with our first designs, we used the program Agadir to help modify the solvent facing residues of bth_1 and chaim_1 to promote helicity and solubility. We took our two peptide sequences, and at positions designed to be solvent exposed, we placed either a glutamic acid, lysine, arginine, histidine, or tyrosine. At the C-terminus we placed a glutamine, as this is a known helix-capping residue (35). These five residues are highly soluble and can be useful in stabilizing helical secondary structure. All possible sequences made using these substitutions were fed into Agadir which produces a helicity prediction score ranging from 0% to 100%. We chose the sequences with the highest predicted helicity to synthesize – 83% for bth_1 and 57% for chaim_1.

Peptide synthesis

All peptides were synthesized at 200 μM scale on ChemMatrix Rink Amide resin in a Symphony/Multiplex peptide synthesizer using FMOC protected amino acids. Peptides were cleaved using a 20 mL solution of 90% trifluoroacetic acid (TFA), 5% Thioanisole, 3% 1,2-ethanedithiol (EDT), and 2% anisole at room temperature for two and a half hours. TFA was removed by blowing nitrogen over the mixture for about half an hour, and the peptides were then precipitated out in 30 mL of diethyl ether chilled on dry ice. The peptides were pelleted

using centrifugation and the diethyl ether was removed. Peptides were resuspened in 30 mL of diethyl ether and spun down three more times. After discarding the diethyl ether from this last wash, the crude peptides were lyophilized.

Addition of cross-linkers

Dibromo-m-xylene was cross-linked to our five initial peptides. Crude peptide was dissolved in 50 mM NH_4HCO_3 to a concentration of 2 mg/mL and a 1:1 molar equivalent of tris(2-carboxyethyl)phosphine (TCEP) was added. The solution was shaken at room temperature for an hour. Three equivalents of dibromo-m-xylene dissolved in dimethylformamide (DMF) with a volume 1/10 that of the crude peptide mixture was then added and this solution was shaken for 2 additional hours. The reaction was quenched by making the solution slightly acidic through the addition of a small amount of 1 M HCl.

Peptide purification

Peptides were purified using reverse phase HPLC in a gradient between solvent A (water and 0.1% TFA) and B' (isopropanol, acetonitrile, and water at a 6:3:1 ratio with 0.1% TFA) using a C4 column. Peptide identities were confirmed using matrix-assisted laser desorption/ionization (MALDI) mass-spectroscopy.

Circular dichroism spectroscopy

Purified peptides were dissolved in 20 mM phosphate buffer with pH 7.2 at concentrations in the 50 μM to 350 μM range. Spectra were recorded on a JASCO J-810 spectropolarimeter from 190 nm to 250 nm with band width of 1 nm, 4 second response time, and 3 accumulations were averaged.

Bio-layer interferometry

Binding experiments via bio-layer interferometry were carried out in the Ian Wilson lab by Cyrille Dreyfus using an Octet instrument produced by ForteBio.

Creation of phage display constructs

Single-stranded DNA coding for our peptide designs was ordered from IDT and cloned into a plasmid containing the phage pVIII protein using Kunkel mutagenesis as described in (41). Proper insertion was verified through sequencing by Elim Biopharm.

Constructs for our peptides placed on protein scaffolds were assembled into double stranded DNA elements from overlapping primers designed by DNAWorks (42). These double stranded DNA elements were inserted into the plasmid containing the pVIII phage gene through Gibson cloning similar to the method described in (43).

A FLAG-tag sequence (DYKDDDDK) was attached to the N-terminus of each of our protein scaffolds, by a GGGGS linker. Currently we are trying a 15mer linker (GGGGSTAGSGATTSG) between our protein construct and the pVIII gene.

Phage stocks were created as described in (41).

Creation of phage libraries

Designed peptides were divided into sections of 7 or 8 residues. Single-stranded DNA was ordered to add diversity to each given section in the following manner: 1) overlapping regions 5' and 3' to the section to be diversified were selected; 2) codons in the regions to be diversified were soft-randomized using mixtures of nucleotides at each codon position so that the original, designed amino acid would be present roughly 50% of the time while the other amino acids would be sampled the remainder of the time, 3) primers including the overlapping regions and nucleotide mixtures were ordered from IDT and inserted via Kunkel mutagenesis.

The soft-mutagenesis strategy was accomplished by creating 4 mixtures of nucleotides – each mixture set one nucleotide to be present at 70% and the remaining nucleotides were present at 10%. In this way, if we originally designed a methionine, we'd replace the ATG codon with a codon compose of (70% A, 10% C,T,and G)(70% T, 10% A, C, G) (70% G, 10% A,C,T), which would thus have a methionine present 34% of the time, and some other amino acid present the other 66% of the time.

Phage selection procedures

We used streptavidin MagneSphere beads from ProMega in a Thermo KingFisher instrument for our selection experiments. First, phage libraries were pre-depleted of off-target binders to the magnetic beads by incubating 200 μL of phage at 10^{13} phage/mL with magnetic beads from 30 μL of bead slurry for 2 hours. After 2 hours, the magnetic beads were removed and the now-depleted phage library was used in the following selection procedures.

60 μL of bead slurry and 100 μL of TBS + 0.5% BSA + 0.1% Tween was added to a well for each library experiment performed. Beads were transferred to 200 μL of phage stock from the depletion + 500 nM of biotinylated hemagglutinin, and mixed for 2 hours followed by a 10 minute pause. Next, beads were transferred to wells with 200 μL PBS + 0.1% Tween and washed by mixing for 30 seconds. The washes were repeated 5 times, each time in a fresh well of 200 μL PBS + 0.1% Tween. Finally, the beads were transferred to wells with 200 μL of glycine at a concentration of 0.1 M with pH 2.2 and mixed for 10 minutes to elute bound phage. Once the beads were removed, the elution reaction was quenched by adding 25 μL of 1 M Tris at pH 11.

The eluted phage were then used to generate phage for the next round of selection by adding the quenched phage stock to 2 mL of XL1 bacteria and shaking for a half hour at 37^{o} C.

After half an hour KO7 helper phage were added and this was shaken for an addition hour at 37^{o} C at which point this culture was added to 30 mL of 2xYT media with appropriate inducers and selectors for overnight expression. The next morning phage were harvested and the selection procedure was repeated.

ELISA assays

Nunc Maxisorp 96-well plates from Thermo Scientific were coated with 100 μL of hemagglutinin at 10 μg/mL in PBS buffer. After shaking overnight at 4^{o} C, wells were emptied and any uncoated surfaces were blocked with bovine serum albumin (BSA) by adding 100 μL of 0.5% BSA in PBS and shaking at room temperature for 2 hours. Control wells without hemagglutinin were likewise blocked with BSA. The PBS + 0.5% BSA solution was removed, and wells were washed 4 times with PBS + 0.1% Tween. Phage particles were resuspended in TBS + 0.5% BSA + 0.1% Tween at a concentration of 10^{13} phage/mL, and 100 μL was added to each well. In wells that were to be co-incubated with the known, high-affinity, broadly neutralizing antibody to hemagglutinin, the antibody FI6, provided by the Ian Wilson lab, was added to the TBS + 0.5% BSA + 0.1% Tween solution to give an FI6 concentration of 10 nM. 100 μL of phage solution was added to each well, and left to shake at room temperature for 2 hours. After 2 hours we removed the phage solution, and washed each well 6 times with TBS + 0.1% Tween. HRP/Anti-M13 antibody from GE Healthcare was diluted by a factor of 1:5000 into TBS + 0.5% BSA + 0.1% Tween solution, and 100 μL was added to wells. This was allowed to shake for 30 minutes at room temperature, and was then removed. Each well was washed 2 times with PBS + 0.1% Tween followed by 2 times with PBS only. A 1:1 mixture of TMB Peroxidase Substrate and Peroxidase Substrate B supplied by KPL was made, and 100 μL was added to each well. This shook for 5 to 15 minutes at room temperature, and was then quenched by the addition of 100

μL of 1 M H_3PO_4. Endpoint values were read at 450 nm using a SpectraMax M5 Microplate reader.

4.5 Acknowledgements

I'd like to thank Chaim Schramm, Gabriel Gonzalez, and Gevorg Gregorian who contributed in the development of the design methodology. I'd also like to thank the Ian Wilson lab for contributing the hemagglutinin and FI6 antibody used in our experimental assays and especially Cyrille Dreyfus for running the bio-layer interferometry experiments. The Wells lab graciously contributed time and materials to the phage display experiments. And a big thank you to J.T. Koerber, who was instrumental in teaching me how to properly run and interpret the phage display experiments.

4.6 Figures

Design Name	Synthesis with cross-linkers	Synthesis without cross-linkers	Phage Display Constructs
bth_1	LKQWICNFNCKVQEEL	LEEWIRRFEEYWRRMQ	TEEWIRRFSEYFRRMLEEWRRN
bth_2	NKEELNCKLCELFK		NSEELIRRIEELRRL
bth_3	NEEIKCRLCEMWR		
chaim_1	AQELLCKWECQAKQL	ALDLLRRWEEEARRLQ	ALDLLRRWEEEARRL
chaim_2	QQFLCEFECWAKKM		
Fleishman		ESRFDEYMRRMWEEVFRRQ	ESRFDEYMRRMWEEVFRRN

Table 1

A list of the five hemagglutinin binder designs. Initially each design contained a pair of cysteine residues to allow for the attachment of a chemical cross-linker. Later, two designs were modified to remove this linker, and solvent-facing residues were modified to improve solubility and helicity. We also created a soluble helix based upon the structure of Fleishman et al. Residues which are designed to make contact with hemagglutinin are shown in bold, red type.

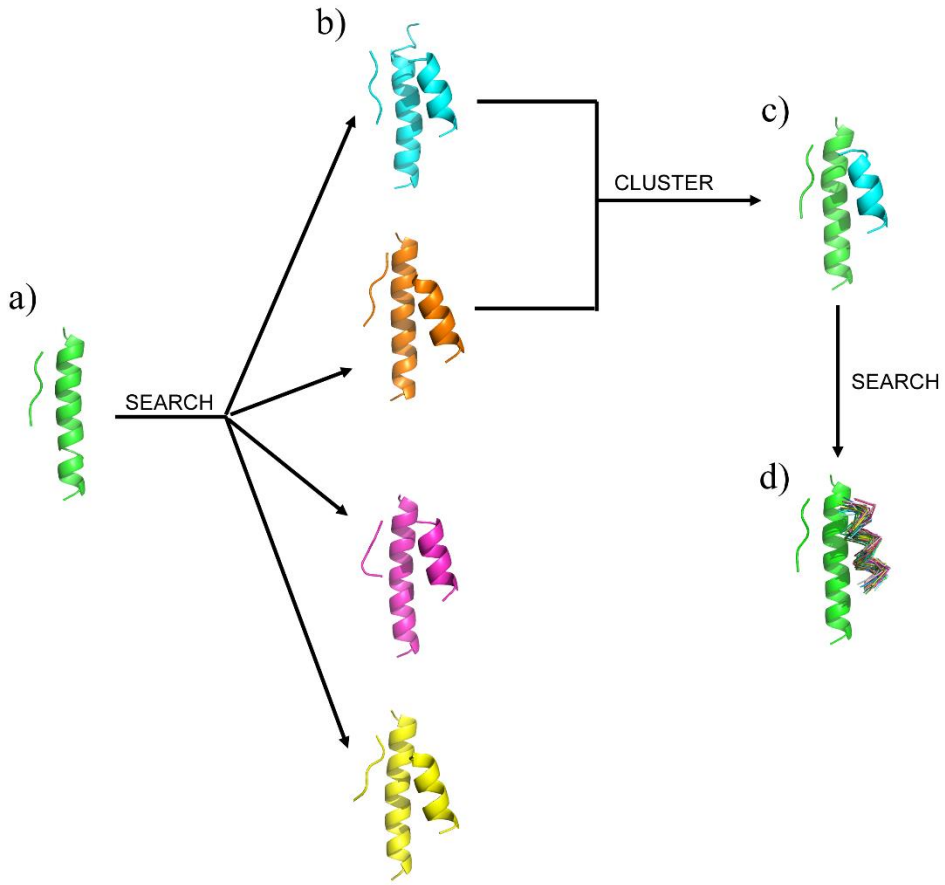


Figure 1

An overview of MaDCaT-based scaffold search.

a) Our initial MaDCaT query consists of a portion of the beta-strand and a portion of the helix in the conserved stem region of hemagglutinin.

b) We look through MaDCaT results and look for helices that are not part of matching segments which would lie in the hydrophobic groove of hemagglutinin when matching segments are superimposed. These helices can be used as backbone scaffolds for our peptide designs.

c) We extract these potential backbone scaffolds, and add them to the original query to perform a second MaDCaT search.

d) The ensemble of potential backbone scaffolds is extracted from this second MaDCaT search and used in fixed-backbone design.

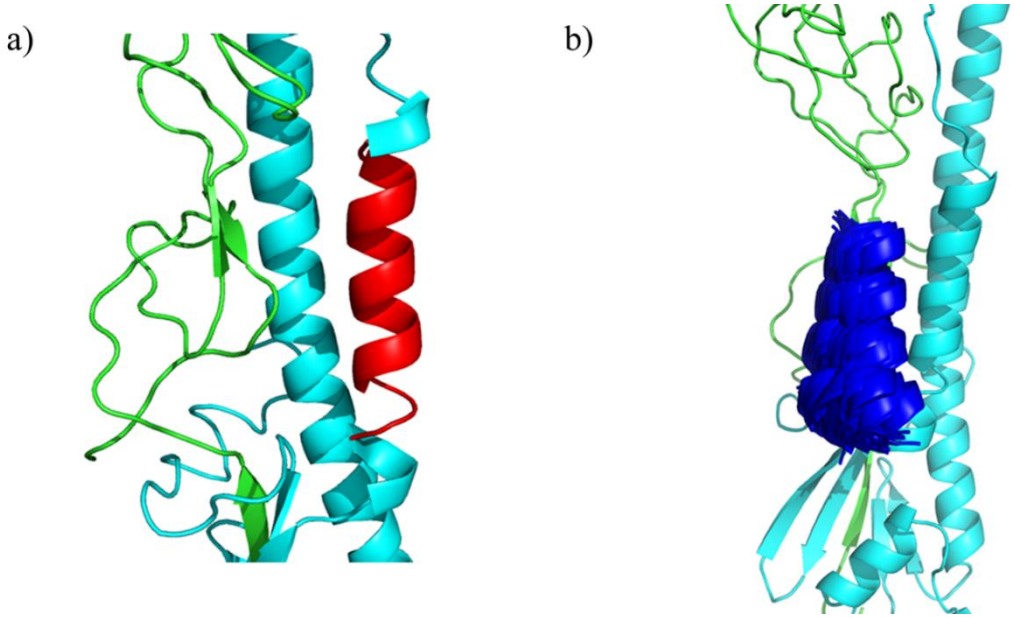


Figure 2

An overview of the helix-dimer scaffold search:

a) The hemagglutinin helix we used as a query to search our database.

b) A cluster of interacting helices that have partners which are structurally similar to the query helix.

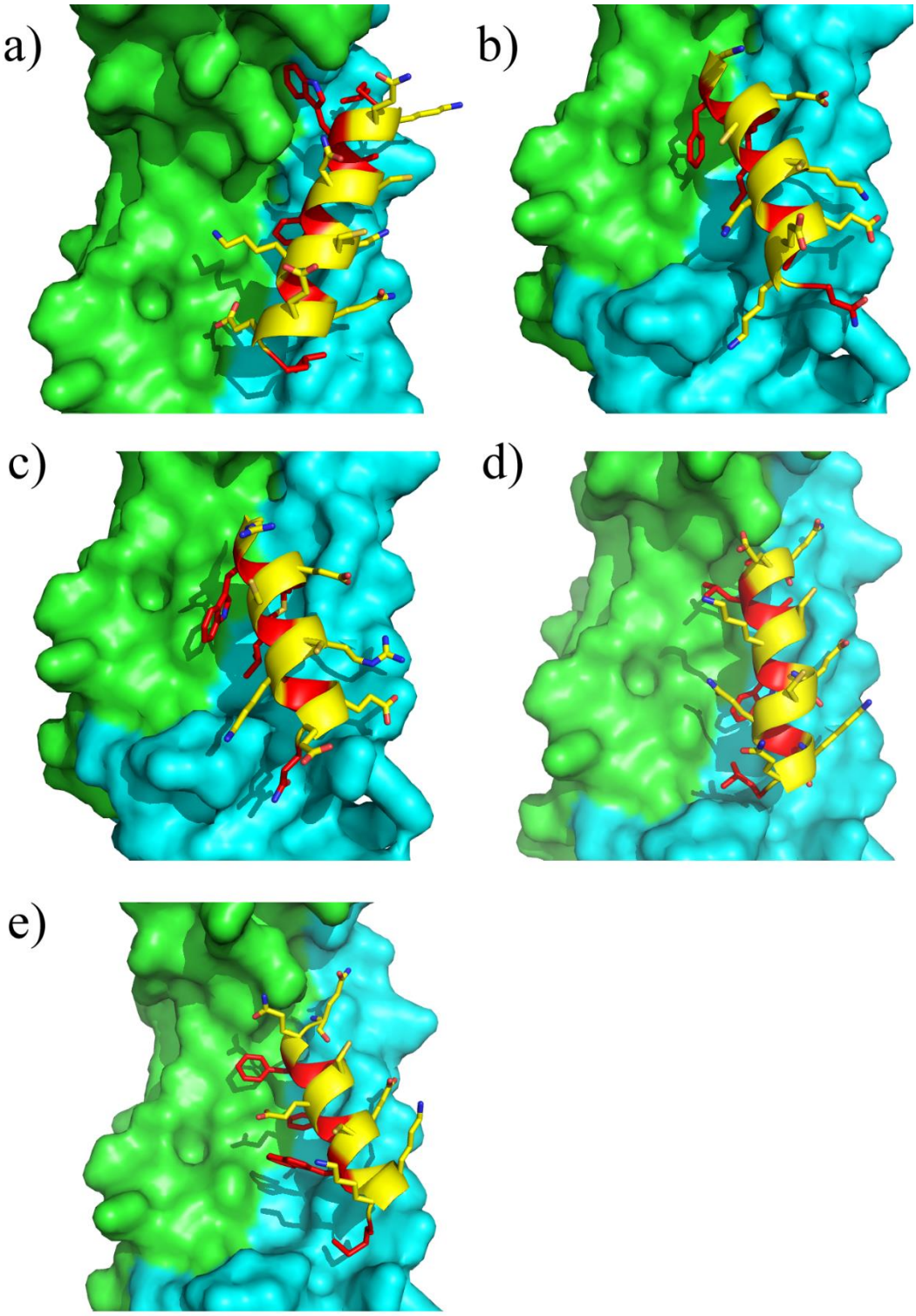
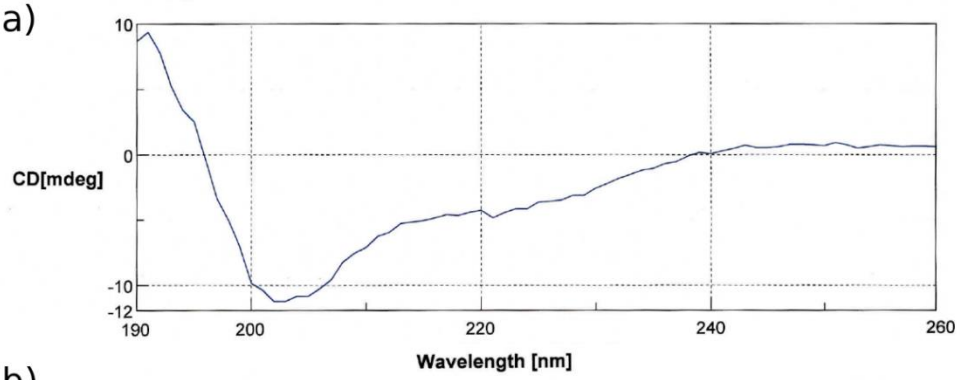

a)
b)
c)
d)
e)

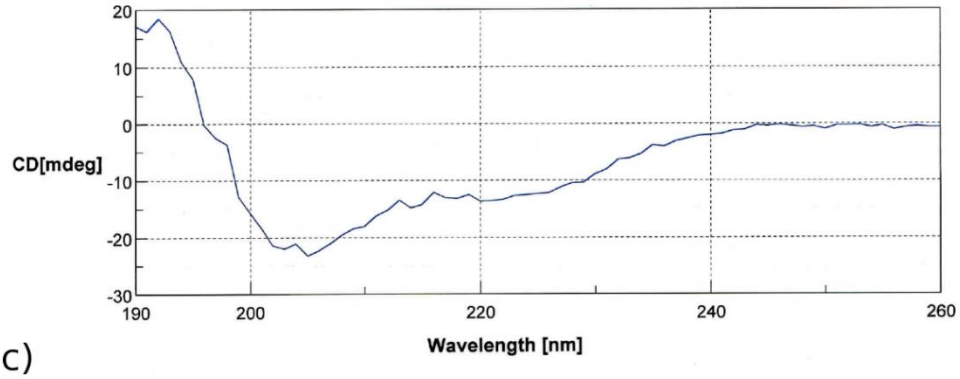
Figure 3

Our five initial designs for peptide binders to hemagglutinin. Hemagglutinin is seen in surface representation with chain A colored green and chain B colored cyan. Our peptides are shown in cartoon representation with sticks for side-chains. Residues designed to make contact with hemagglutinin are colored red while those designed to be solvent exposed are colored yellow. Designs are a) bth_1, b) bth_2, c) bth_3, d) chaim_1, e) chaim_2.

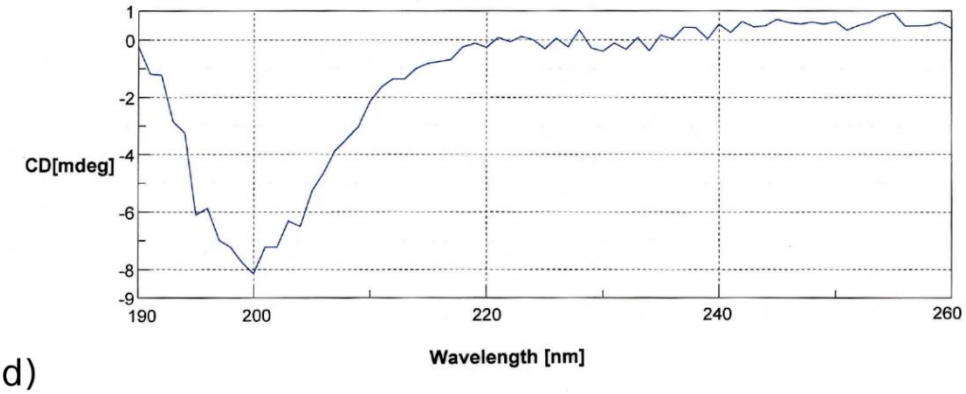
a)


10
0
-10
-12
CD[mdeg]
190
200
220
240
260
Wavelength [nm]

b)


20
10
0
-10
-20
-30
CD[mdeg]
190
200
220
240
260
Wavelength [nm]

c)


1
0
-2
-4
-6
-8
-9
CD[mdeg]
190
200
220
240
260
Wavelength [nm]

d)

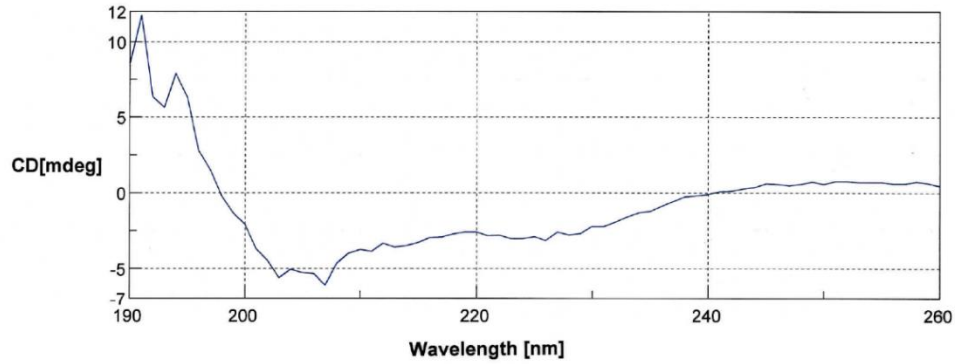

12
10
5
0
-5
-7
CD[mdeg]
190
200
220
240
260
Wavelength [nm]

Figure 4

Circular Dichroism spectra of four of our synthesized peptides with chemical cross-linkers (bth_1 resisted solubilization and was unable to be characterized):

a) bth_2

b) bth_3

c) chaim_1

d) chaim_2

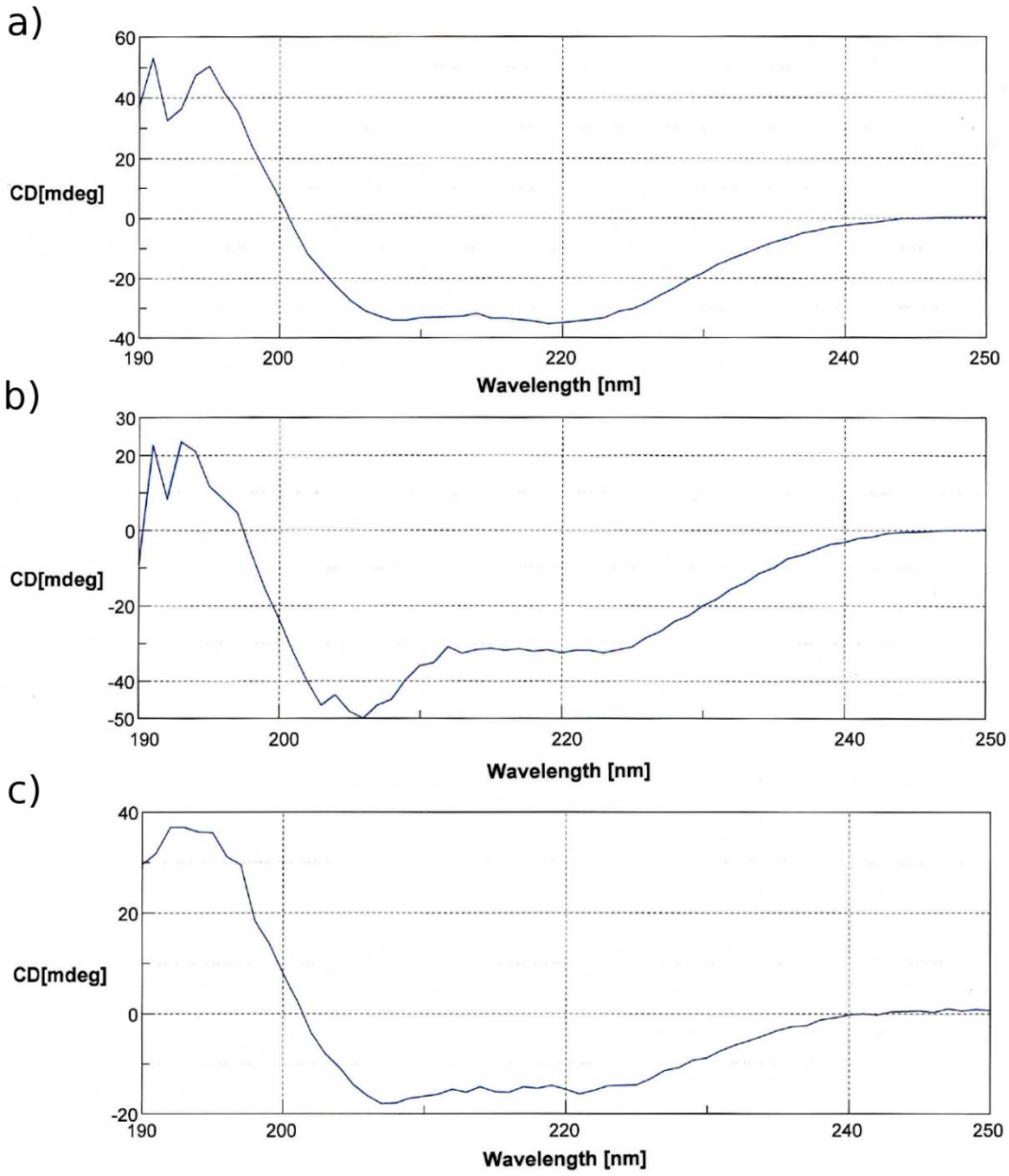


Figure 5

Circular dichroism of our three peptides without chemical cross-linkers:

a) bth_1

b) chaim_1

c) Fleishman

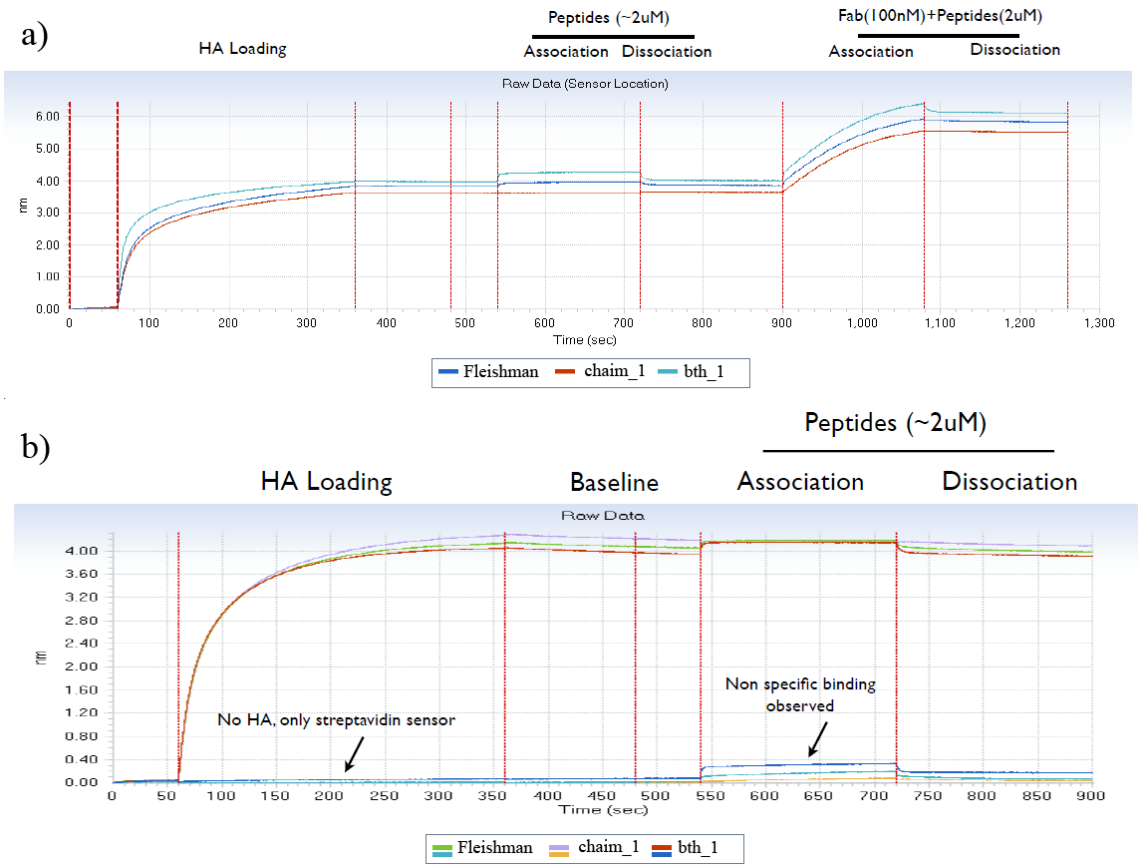


Figure 6

Bio-layer interferometry data examining binding between our helical peptide designs and hemagglutinin.

a) During time period 60 to 360, biotinylated hemagglutinin is flowed across the chip and bound to streptavidin. From 540 to 720 our peptides flow across the hemagglutinin bound to the chip. The bth_1 peptide and Fleishman-inspired helix appear to bind, although the signal drop soon after the peptides stop flowing. From 900 to 1080 an antibody known to bind hemagglutinin flows across the chip along with our peptide designs and binding is detected. There is a slight drop in signal once the flow is stopped, possibly due to our peptides falling off while the antibody remains.

b) As a control, we flow our peptides across the sensor without hemagglutinin bound (cyan, orange, and blue labels). Similar binding is observed to when hemagglutinin is present, indicating peptide binding observed was most likely due to non-specific stickiness of our peptides.

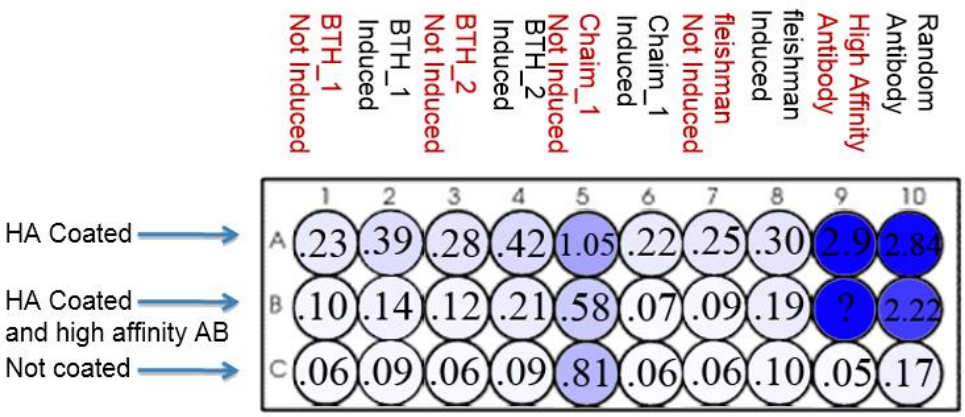


Figure 7

ELISA results for our initial helix designs placed in phage. The first row of wells is coated in hemagglutinin (HA), the second row is coated in hemagglutinin and also has a high-affinity antibody which targets our desired hemagglutinin epitope, and the third row has no hemagglutinin. Signal strength is given in absorbance units at OD450, and shading is proportional to this measure.

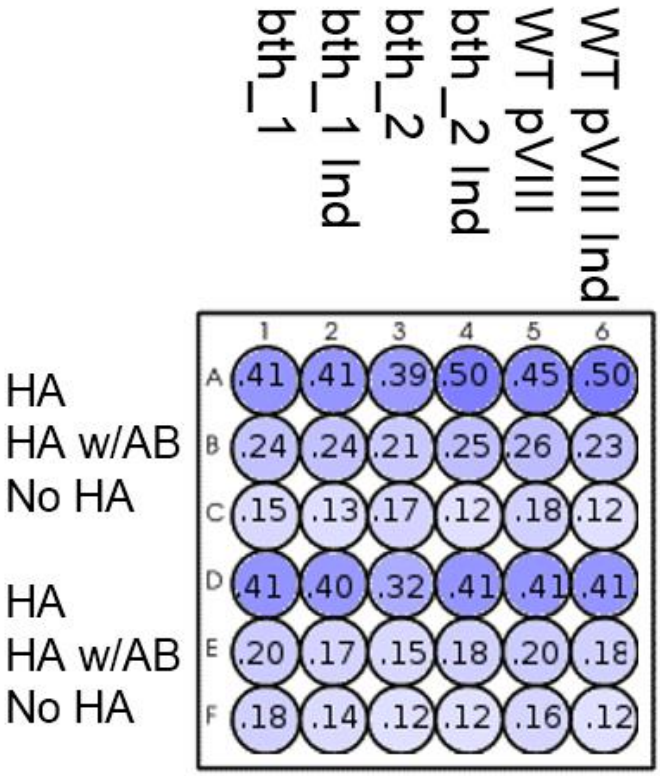


Figure 8

ELISA results obtained for designs bth_1 and bth_2 as well as phage displaying the wild-type pVIII gene. Signal strength is given in absorbance units at OD450 and shading is proportional to strength as in **Figure 7**.

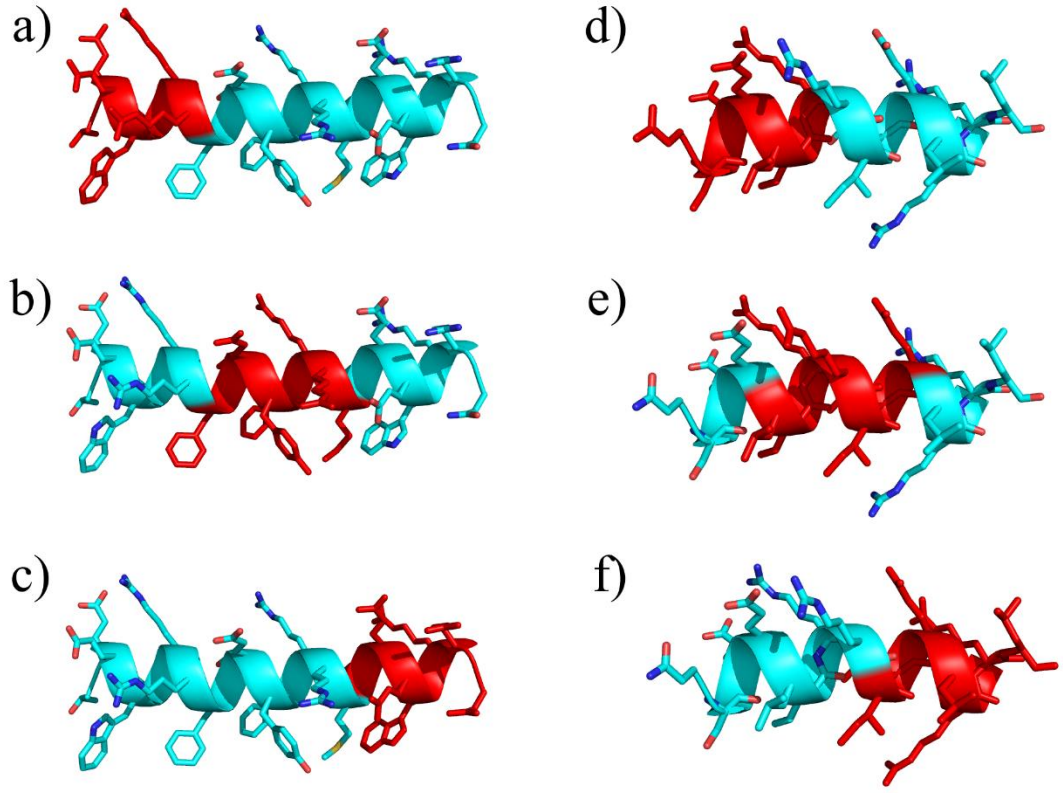


Figure 9

Diversification of initial peptide designs for phage display.

For designs bth_1 (a-c) and bth_2 (d-f) we created three phage libraries where we diversified either the N-terminal, middle, or C-terminal amino acids (diversified residues are shown in red).

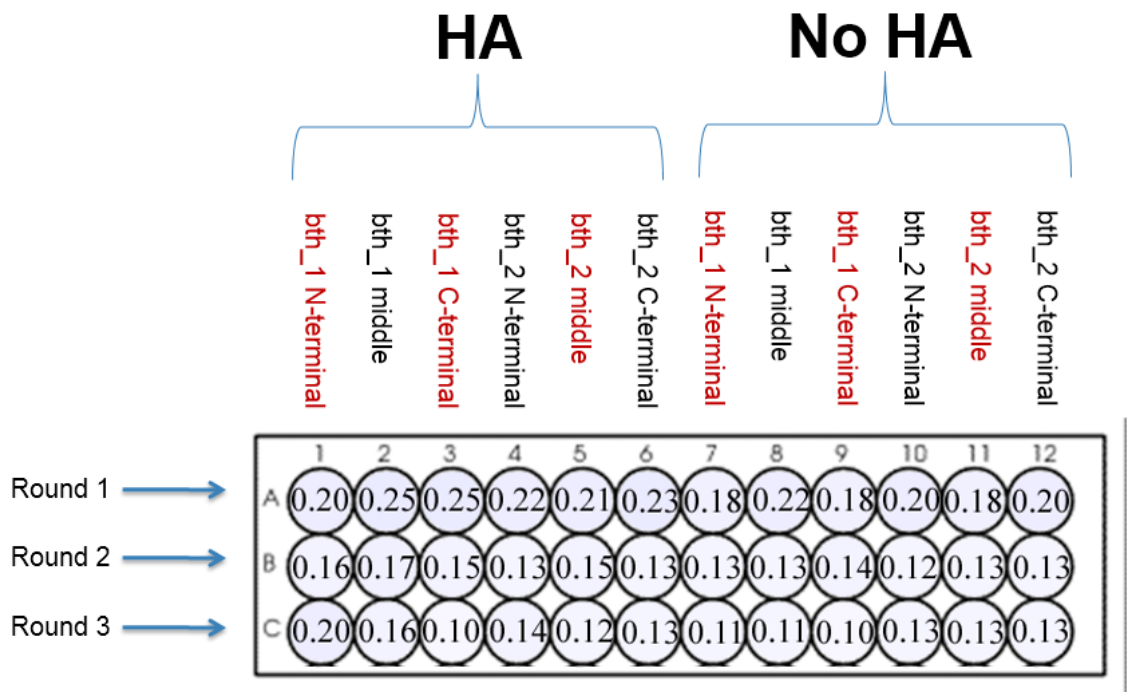


Figure 10

ELISA results after one, two, and three rounds of selection against hemagglutinin. The first six columns give binding results for the given library in wells coated with hemagglutinin. The last six columns serve as a negative control, giving binding results for the given library in wells coated with BSA.

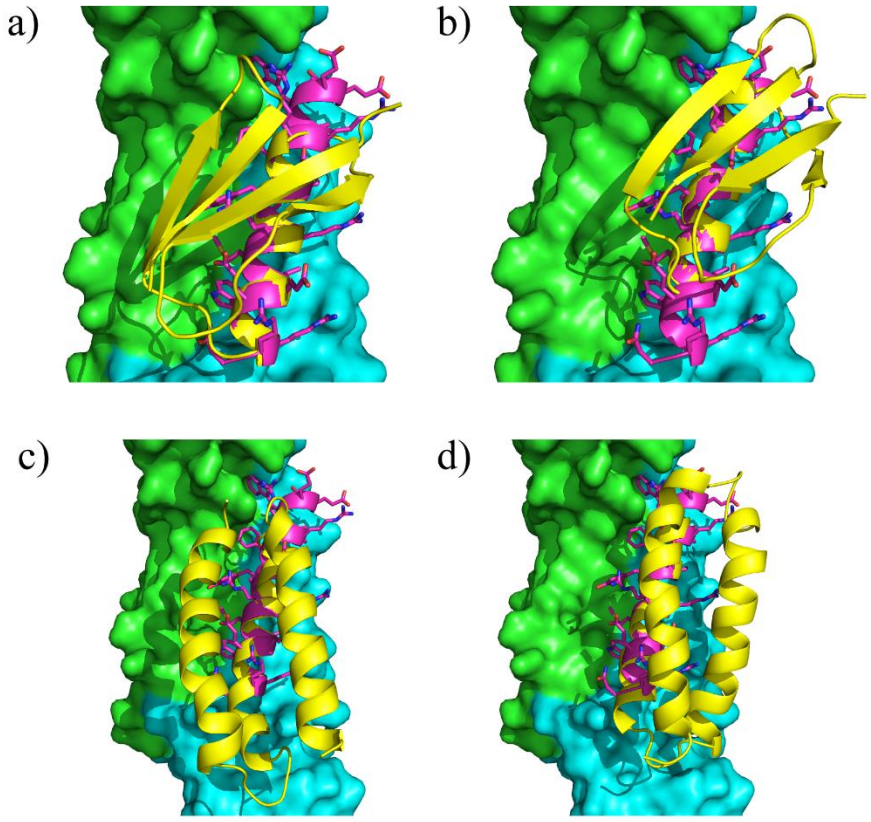


Figure 11

Threading protein scaffolds onto helical peptide designs.

Two proteins were chosen to be used a scaffolds to hold the designed hemagglutinin binder bth_1: a) and b) GB1; c) and d) α_3D. We threaded our designed peptide onto each scaffold in two different registries. The Hemagglutinin molecule is shown in cyan and green, the bth_1 peptide is shown in magenta, and the protein scaffolds are shown in yellow.

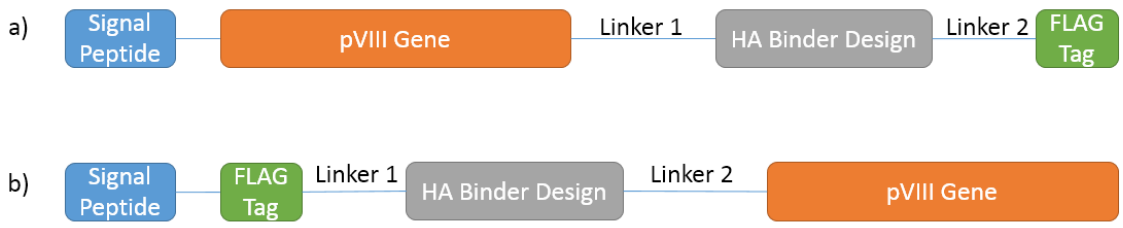


Figure 12

We have designed two possible arrangements for inserting our designed hemagglutinin binder into the phage pVIII gene. a) Binder inserted at C-terminus b) Binder inserted at N-terminus.

The length and composition of the two linkers as well as the sequence of the signal peptide can be modified to increase the yield of properly folding chimeric protein. We can monitor the presence of our chimeric protein on the phage surface through an ELISA assay using anti-FLAG antibody.

4.7 References

1. *Exploring the origins of binding specificity through the computational redesign of calmodulin.* **Shifman, J.M. and Mayo, S.L.** 23: Proceedings of the National Academy of the Sciences, 2003, Proceedings of the Natural Academy of Sciences, Vol. 100, pp. 13274-13279.

2. *Computer-aided design of a PDZ domain to recognize new target sequences.* **Reina, J., et al.** 8: Nature Structure Biology, 2002, Vol. 9. pp 621-627.

3. *Redesign of a protein-peptide interaction: Characterization and applications.* **Jackrel, M.E., Valverde, R. and Regan, L.** s.l. : Protein Science, 2009, Vol. 18. pp 762-774.

4. *Computational Design of a New Hydrogen Bond Network and at Least a 300-fold Specificity Switch at a Protein–Protein Interface.* **Joachimiak, L.A., et al.** s.l. : Journal of Molecular Biology, 2006, Vol. 361. pp 195-208.

5. *Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin.* **Leishman, S.J., et al.** s.l. : Science, 2011, Vol. 332. pp 816-821.

6. *A De Novo Protein Binding Pair By Computational Design and Directed Evolution.* **Karanicolas, J., et al.** s.l. : Molecular Cell, 2011, Vol. 42. pp 250-260.

7. *One thousand families for the molecular biologist.* **Chothia, C.** s.l. : Nature, 1992, Vol. 357. pp 543-544.

8. *Helix-packing motifs in membrane proteins.* **Walters, R.F.S. and DeGrado, W.F.** 37: Proceedings of the National Academy of Sciences, 2006, Vol. 103. pp 13658-13663.

9. **Zhang, S.** A new dictionary of helix-helix interactions in membrane and soluble proteins. *Association of Protein Helices and Assembly of Foldamers: Stories in Membrane and Aqueous Environments.* Philadelphia : s.n., 2013.

10. *Probing designability via a generalized model of helical bundle geometry.* **Grigoryan, G. and DeGrado, W.F.** s.l. : Journal of Molecular Biology, 2011, Vol. 405. pp 1079-1100.

11. *Protein-Peptide interactions adopt the same structural motifs as monomeric protein folds.* **Vanhee, P., et al.** s.l. : Structure, 2009, Vol. 17. pp 1128-1136.

12. *Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways.* **Keskin, O. and Nussinov, R.** 1: Protein Engineering, Design & Selection, 2005, Vol. 18. pp 11-24.

13. *A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.* **Keskin, O., et al.** s.l. : Protein Science, 2004, Vol. 13.

14. *The designability of protein structures.* **Helling, R., et al.** s.l. : Journal of Molecular Graphics and Modelling, 2001, Vol. 19. pp 1043-1055.

15. *Natural selection of more designable folds: A mechanism for thermophilic adaptation.* **England, J., Shakhnovich, B.E. and Shakhnovich, E.I.** 15: Proceedings of the National Academy of the Sciences, 2003, Vol. 100. pp 8727-8731.

16. *Computational Design of Peptides that Target Transmembrane Helices.* **Yin, H., et al.** s.l. : Science, 2007, Vol. 315. pp 1817-1822.

17. *Computational Design of Virus-Like Protein Assemblies on Carbon Nanotube Surfaces.* **Grigoryan, G., et al.** 6033: Science, 2011, Vol. 332. pp 1071-1076.

18. *Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines.* **Bui, H.H., et al.** 361: BMC Bioinformatics, 2007, Vol. 8. doi:10.1186/1471-2105-8-361.

19. *Computational Design of a Protein-Based Enzyme Inhibitor.* **Procko, E., et al.** 18: Journal of Molecular Biology, 2013, Vol. 425. pp 3563-3575.

20. *Protein Design of an HIV-1 Entry Inhibitor.* **Root, M.J., Kay, M.S. and Kim, P.S.** 5505: Science, 2001, Vol. 291. pp 884-888.

21. *Design of a switchable eliminase.* **Korendovych, I.V., et al.** 17: Proceedings of the National Academy of Sciences, 2011, Vol. 108. pp 6823-6827.

22. *Exploring the origins of binding specificity through the computational redesign of calmodulin.* **Shifman, J. and Mayo, S.L.** 23: Proceedings of the National Academy of the Sciences, 2003, Vol. 100. pp 13274-13279.

23. *Accurate Prediction of Peptide Binding Sites on Protein Surfaces.* **Petsalaki, E., et al.** 3: PLoS Computational Biology, 2009, Vol. 5. doi: 10.1371/journal.pcbi.1000335.

24. *CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of unctionally annotated residues.* **Dundas, J., et al.** s.l. : Nucleic Acids Research, 2006, Vol. 34. pp W116-W118.

25. *Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure.* **Capra, J.A., et al.** 12: PLoS Computational Biology, 2009, Vol. 5. doi: 10.1371/journal.pcbi.1000585.

26. *Fpocket: An open source platform for ligand pocket detection.* **Le Guilloux, V., Schmidtke, P. and Tuffery, P.** 168: BMC Bioinformatics, 2009, Vol. 10. doi:10.1186/1471-2105-10-168.

27. *Mining Tertiary Structural Motifs for Assesment of Designability.* **Zhang, J. and Grigoryan, G.** s.l. : Methods in Enzymology, 2013, Vol. 523. pp 21-40.

28. *A Real-Time All-Atom Structural Search Engine.* **Gonzalez, G., Hannigan, B.T. and DeGrado, W.F.** s.l. : Unpublished, 2013.

29. *Studies of the membrane fusion activities of fusion peptide mutants of influenza virus hemagglutinin.* **Steinhauser, D.A., et al.** 11: Journal of Virology, 1995, Vol. 69. pp 6643-6651.

30. *A neutralizing anitbody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins.* **Corti, D., et al.** 6044: Science, 2011, Vol. 333. pp 850-856.

31. *Antibody recognition of a highly conserved influenza virus epitope: implications for univeral prevention and therapy.* **Ekiert, D.C., et al.** 5924: Science, 2009, Vol. 324. pp 246-251.

32. *Structural and functional bases for broad-spectrum nutralization of avian and human influenza A viruses.* **Sui, J., et al.** 3: Nature Structural & Molecular Biology, 2009, Vol. 16. pp 265-273.

33. *Development of α-Helical Calpain Probes by Mimicking a Natural Protein-Protein Interaction.* **Jo, H., et al.** s.l. : Journal of the American Chemical Society, 2012, Vol. 134. pp 17704-17713.

34. *A rapid method for determining dynamic binding capacity of resins for the purification of proteins.* **Do, T., et al.** s.l. : Protein Expression and Purification, 2008, Vol. 60. pp 147-150.

35. *Elucidating the folding problem of helical peptides using empirical parameters.* **Muñoz, V. and Serrano, L.** s.l. : Nature Structural Biology, 1994, Vol. 1. pp 399-409.

36. *Design, structure and stability of a hyperthermophilic protein variant.* **Malakauskas, S. M. and Mayo, S. L.** s.l. : Nature Structural Biology, 1998, Vol. 5. pp 470-475.

37. *Solution structure and dynamics of a de novo designed three-helix bundle protein.* **Walsh, S.T.R., et al.** 10: Proceedings of the National Academy of Sciences, 1999, Vol. 96. pp 5486-5491.

38. *PISCES: a protein sequence culling server.* **Wang, G. and Dunbrack, R.L.** s.l. : Bioinformatics, 2003, Vol. 19. pp 1589-1591.

39. *PDB file parser and structure class implemented in Python.* **Hamelryck, T. and Manderick, B.** 17: Bioinformatics, 2003, Vol. 19. pp 2308-2310.

40. *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.* **Leaver-Fay, A., et al.** s.l. : Methods in Enzymology, 2011, Vol. 487. pp 545-574.

41. *Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries.* **Tonikian, R., et al.** 6: Nature Protocols, 2007, Vol. 2. pp 1368-1386.

42. **DNAWorks. [Online] http://helixweb.nih.gov/dnaworks/.**

43. *Enzymatic assembly of DNA molecules up to several hundred kilobases.* **Gibson, D.G., et al.** s.l. : Nature Methods, 2009, Vol. 6. pp 343-345.

Chapter 5

Super Codons: Creating optimal sets of nucleotide mixtures for use in gene library production

5.1 Abstract

Protein design typically requires the use of targeted mutagenesis libraries to enhance the binding or enzymatic activity of a protein. Proteins with improved binding or enzymatic properties are identified from these libraries through a series of selection steps. To construct the libraries, designers frequently target specific residues for diversification and use degenerate codons at these sites. Degenerate codons are formed by replacing single nucleotides in a codon with equimolar mixtures of nucleotides, thus producing a defined distribution of amino acids at a given site. While this technique has proven to be useful, the resulting distributions are often non-ideal and thus a system providing a more tailored amino acid distribution at each site would create more efficient libraries. To that end, we have developed a novel set of target amino acid distributions which take into account the frequencies of amino acid substitutions observed in nature. Moreover, we develop an algorithm which will calculate four mixtures of nucleotides that will best recapitulate these distributions, or any alternative set of distributions. The distributions which can be made from these four mixtures offer substantial benefits over traditional degenerate codons including: a) more efficiently exploring sequence space, b) taking into account amino acid substitution rates seen in nature, and c) introducing stop codons at lower rates than NNK based libraries. To aid in the adoption of our approach, we have developed a website at supercodons.degradolab.org, which allows researchers to easily create their own nucleotide mixtures to match: (a) their own amino acid distributions, (b) the amino acid distributions from a given multiple alignment, (c) the amino acid distributions observed in antibody sequences, and (d) amino acid distributions based on the substitutions seen in nature.

5.2 Introduction

Gene library production has become a favorite tool of protein designers. Typically, designers will start with a design for a new or improved protein, whether based upon intuition, homology studies, or computational design, and then create a library of mutants that samples sequence space around this initial design. For instance, one study described designing 88 potential protein binders to the influenza hemagglutinin protein using the computational package Rosetta (1). These de novo proteins were assayed for binding, revealing two moderate binders, and further optimization of these binders required generation of libraries in which the designs were randomly mutated. As library size is limited however, randomizing the entire protein lacks the efficiency provided by an approach randomizing fewer, more structurally targeted residues. Antibody researchers frequently use libraries of antibody scaffolds, randomizing portions of the complementarity determining regions (CDRs) to produce antibodies which bind novel targets (2) (3) (4). Other groups have used library approaches to create or modify enzymatic activity of a protein (5) (6) (7).

During library production, researchers must decide on a strategy to introduce diversity into their initial design. Small-scale screening methods require small, focused libraries and thus researchers may choose to make only a few targeted mutations, relying on intuition to limit the number of amino acids tried at a given position, or to perform saturation mutagenesis of all 20 amino acids at only a very few positions. However, for large-scale, high-throughput selections, researchers can explore a larger section of sequence space, with a total diversity of up to 10^7 variants for yeast-display experiments, 10^9 variants for phage-display experiments, and up to 10^{14} variants for ribosome and mRNA display experiments (8). This strategy allows many more positions to be diversified, prompting researchers to use techniques such as degenerate codons (9).

Degenerate codons use various equimolar mixtures of the four standard nucleotides when synthesizing DNA to create defined amino acid distributions. For example, **Table 1** lists a few of the more popular degenerate codons in use, as well as commonly used descriptions for these degenerate codons (10). In **Figure 1** we show the computed amino acid distributions one would expect to observe when using these degenerate codons. Despite offering some level of tunability in specifying what mutations are introduced at a given site, these distributions are still sub-optimal for library generation. For instance, many of the distributions have a majority of amino acids represented at a frequency of 10% or lower. If one were to diversify ten sites using these degenerate codons, then less than one out of ten billion sequences would be the original designed sequence. Even with the large maximum diversity level of 10^9 variants afforded by phage display, one would not expect a single copy of the original protein with weak activity to be present in the library. Ideally a small, but non-negligible, fraction of library members would carry the sequence for the original protein with weak affinity or activity, and only improved mutants would out-compete these during selection. Moreover, most of the degenerate codons preclude certain amino acids altogether. While the NTT codon may be used to place a hydrophobic amino acid at a given position, it is possible that, surprisingly, a charged residue greatly improves binding or activity. Having a distribution which favored hydrophobic amino acids while still allowing for the possibility of a charged or hydrophilic amino acid would be advantageous.

There are two main approaches one could use to realize a given amino acid distribution at a target site – through the use of trinucleotide phosporamidites, or by creating three individual nucleotide mixtures to compose custom degenerate codons. The first approach essentially preforms codons and then mixes these in the proper proportion to match the desired amino acid distribution (11) (12). While this method will ensure a very close match to the

desired amino acid distribution, to-date trinucleotide use is not as wide-spread or cost-effective as simply making individual nucleotide mixtures. Alternatively, a set of three nucleotide mixtures is calculated which, when used to form a codon, will produce an amino acid distribution which closely matches the target distribution. A number of groups have proposed algorithms to calculate these nucleotide mixtures (13) (14) (15) (16). One shortcoming of these approaches is that they create three different nucleotide mixtures for each desired amino acid distribution. Thus, if a researcher wishes to use 10 different amino acid distributions in their library, 30 different nucleotide mixtures will be required. This is time-consuming and costly.

In this present work, we begin by developing a set of amino acid target distributions that address these deficiencies of the standard degenerate codons. Each target amino acid distribution identifies one amino acid as the dominant amino acid, while the rest of each distribution is calculated based upon amino acid substitution probabilities found in nature. By using a target distribution which keeps the original, designed amino acid as the dominant amino acid, researchers can ensure that some small fraction of the library will contain examples of the original design with weak binding or activity. Moreover, as the other amino acids will attempt to be present following the probability of a substitution from this dominant amino acid, a larger fraction of the library will be spent sampling sequences more relevant to the original design, while still allowing for the unexpected beneficial mutation like the aforementioned hydrophobic to charged residue substitution.

We extend our previous work (16) by using constrained optimization to identify four mixtures of nucleotides which will best allow a researcher to produce a set of amino acid distributions that match a given set of target distributions. We then use this procedure to generate a set of four nucleotide mixtures that will be able to generate amino acid distributions that closely match our derived target distributions. Additionally, we have developed a web

portal at supercodons.degradolab.org, which will allow researchers to generate their own nucleotide mixtures to match a) amino acid distributions of their choosing, b) amino acid distributions based on a multiple-alignments and c) amino acid distributions based on observed amino acid frequencies in given antibody positions.

5.3 Results and Discussion

Development of target amino acid distributions

A protein engineer will often further optimize the properties of a designed protein using degenerate codons to randomize portions of the initial design. Degenerate codons allow for some flexibility in what amino acids will be introduced at a given position. For instance, as seen in **Figure 1f**, the NTT codon should introduce an equal number of phenylalanine, isoleucine, leucine, and valine residues at a given location, while disallowing the other sixteen amino acids. The NWW, RVK, and DVT codons on the other hand, will each permit about half of the amino acids while disallowing the other half (**Figure 1c-e**). Meanwhile, the NNN and NNK codons allow for all 20 amino acids, at frequencies between 3% and 10% (**Figure 1a-b**). While allowing for some tunability with regards to what amino acids get introduced where, we believe the distributions allowed by degenerate codons are less than ideal in three distinct ways.

First, while a number of residues of an initial design may be targeted for diversification, often the mutations necessary to enhance binding or activity are found in only a few of these sites. For example, in the two hemagglutinin binders designed in (1), only two or three mutations were necessary to improve the binding of their initial designs by roughly two orders of magnitude. However, traditional degenerate codons will spend the majority of the library exploring sequence space at a significant distance from the original design. For instance, let us imagine choosing ten sites to diversify and using a degenerate codon which will place the original, designed amino acid at that site 10% of the time (as many of the degenerate codons

will). Furthermore, let us assume that we create a phage display library of 10^9 members. Then, using the binomial distribution, we know we can expect only

$$\binom{10}{10} * 0.1^{10} * 10^9 = 0.1$$

members to have the sequence of the initial, weak design,

$$\binom{10}{9} * 0.1^9 * 0.9^1 * 10^9 = 9$$

members to have a single mutation, and

$$\binom{10}{8} * 0.1^8 * 0.9^2 * 10^9 = 364.5$$

members to have just two mutations. In fact, only 0.163% of the library will be composed of members with 5 or fewer mutations from the starting sequence. In other words, a very large proportion of the library will be spent sampling sequence space at a considerable distance from the initial design, as illustrated by **Figure 2a**. Moreover, it is generally assumed that the total number of functional mutants left in the library decreases as the number of mutations from the initial design increases (17). It would be preferable if the library sampled each initial, designed amino acid at a high frequency, while sampling alternative amino acids at low frequencies.

Secondly, the distributions offered by the various degenerate codons frequently do not match with biophysical intuition. For instance, as alluded to by the description "charged hydrophobic," the degenerate codon NWW samples heavily from hydrophobic residues like leucine, isoleucine and valine, while also sampling charged amino acids like histidine, glutamic acid, and aspartic acid. While a dramatic mutation from a small hydrophobic residue to a large, charged residue may on occasion dramatically improve a design, they are also more likely to disrupt the protein fold. It is likely that for most residues chosen for diversification, subtle tweaks will be more amenable. Thus, we propose that distributions that introduce amino acid substitutions with frequencies more akin to what is seen in nature would be more successful.

Thirdly, as shown in **Figure 1c-f**, many of the degenerate codons permit some subset of the amino acids while completely omitting another subset. While many favorable mutations may be fairly conservative in nature, it is also true that some very beneficial mutations will be completely unexpected. For instance, one group was able to improve the catalytic efficiency of a computationally designed Kemp eliminase over 100 fold with four mutations (18). While two mutations were fairly conservative (the hydrophobic to hydrophobic mutation F77I and the asparagine to aspartate N224D), two were rather surprising (the hydrophobic to charged I7D and the small flexible to large and charged G202R). Thus, we believe that amino acid distributions that allow for at least a modest sampling from each of the twenty amino acids will better allow these surprising mutations to be discovered.

In an attempt to fix these deficiencies, we have developed a set of target amino acid distributions. First, we create 20 separate amino acid distributions where in any given distribution, one of the amino acids has a high target frequency (60%). In this way, we increase the fraction of the library which samples sequence space close to the original design, as illustrated in **Figure 2b.** In fact, when diversifying 10 positions using this method, a full 63.3% of the library should have 4 or fewer mutations relative to the initial design, in contrast to the 0.16% of sequences when the original amino acid is present only 10% of the time.

We next address the other two identified deficiencies in degenerate codons – namely the desire for the introduced mutations to make biophysical sense, while simultaneously not precluding any mutations *a priori*. To this end, we first define each target distribution by the amino acid represented at 60%. Next, use the original alignment for the popular amino acid substitution matrix BLOSUM 62 (19) (the default substitution matrix used in protein BLAST) to calculate the frequency of substitution from this majority amino acid to each of the remaining 19 amino acids. We then complete the target distribution by allocating the remaining 40% in

proportion to these substitution frequencies. As the aligned sequences used in generating the BLOSUM matrices come from local alignments of evolutionarily related proteins, we expect our proposed target distributions to better track permissible mutations. At the same time, because every possible amino acid substitution is observed in these alignments, no amino acid will be precluded from one of the target distributions. It should be noted that no attempt is made to correct for differences in the natural abundance of the 20 amino acids. Thus, an amino acid like tryptophan which occurs rarely in known proteins (20), will likewise be a rare substitution in our target distributions. We do this deliberately as we expect designed amino acid frequencies to track natural amino acid frequencies.

The results of these steps can be seen in the 20 target amino acid distributions seen in **Figure 3** and **Supplemental Table 1**. As previously stated, each target distribution selects one amino acid to be present at 60%, while the remaining amino acid probabilities follow biophysical intuition. For instance, looking at the distribution which sets valine to 60%, we see that leucine and isoleucine are the next most favored amino acids. This makes sense as all three are medium-sized, hydrophobic amino acids. On the other hand, tryptophan and histidine are the least favored amino acids. Presumably, the fact that tryptophan is a relatively rare amino acid to start with, and the fact that it is quite large compared to valine, accounts for its low frequency. Histidine's low frequency can be ascribed to the facts that it too is a rare amino acid, and it is a large, slightly basic side-chain as opposed to a medium, hydrophobic one. In the target distribution which set glutamic acid to 60%, one finds the next most popular amino acids to be the charged amino acids of aspartic acid and lysine, while smaller, hydrophobic amino acids are less likely. Similar patterns can be found in the remaining 18 target distributions.

Creating libraries with derived target distributions

With our target distributions in place, we now turn to the question of how best to introduce these distributions into gene libraries. We propose to use mixtures of nucleotides at each of the three positions of a codon in an attempt to best match the desired target amino acid distributions.

A number of groups have studied methods to create mixtures of nucleotides that will approximate a given target distribution (21) (22) (23) (16). These methods create three nucleotide mixtures for each target amino acid distribution, requiring 3*N different mixtures for N target distributions. Each separate nucleotide mixture is not without cost, however. Either one has to make each separate mixture, or ask for it to be made when ordering the DNA construct – a request typically granted at a cost of a couple hundred dollars per mixture. We noted that the four nucleotides of DNA could be combined to make 64 codons which code for the 20 amino acids found in proteins. Likewise, perhaps we could create four mixtures of nucleotides, which we refer to as Super Nucleotides, which could then be used to generate 64 different Super Codons, with each Super Codon producing a different amino acid distribution. Our goal is then to find the set of four Super Nucleotides that would allow us to best approximate each of our target amino acid distributions. By limiting ourselves to only four nucleotide mixtures, we dramatically decrease the cost and difficulty associated with creating gene libraries.

We frame this as an optimization problem. The user supplies a set of target distributions

$$\{\text{TargetDist}_1, \text{TargetDist}_2, \text{TargetDist}_3, \dots, \text{TargetDist}_N\},$$

where each target distribution is simply:

$$TargetDist_i = \{\%Ala, \%Arg, \%Asn, \dots, \%Val\}$$

We then want to find the set of four Super Nucleotides, {SNt_1, SNt_2, SNt_3, SNt_4} which will allow us to best match our target distributions. Each Super Nucleotide is simply a mixture of the four standard nucleotides.

$$SNt_i = \{\%A, \%C, \%G, \%T\}$$

The set of 4 Super Nucleotides can be used create 64 different Super Codons and their corresponding amino acid distributions:

$$SuperCodon_1 = (SNt_1, SNt_1, SNt_1) \rightarrow Dist_1$$

$$SuperCodon_2 = (SNt_1, SNt_1, SNt_2) \rightarrow Dist_2$$

…

$$SuperCodon_{64} = (SNt_4, SNt_4, SNt_4) \rightarrow Dist_{64}$$

We then want to minimize our objective function, which is:

$$\sum_{i=1}^{N} \min_{1 \le j \le 64} CompareFun(TargetDist_i, Dist_j) \quad (1)$$

where CompareFunction is the function we use to compare how similar two amino acid distributions are to one another.

In our previous work (16) we evaluated a number of functions that could be used to compare distributions, and preferred a function which mixed an entropic term with chi-squared term as shown below.

$$CompareFun = \sum_{a=1}^{21} P_{calc}(a) \ln \frac{P_{calc}(a) + \varepsilon}{P_{des}(a) + \varepsilon} + 0.5\,[P_{des}(a) - P_{calc}(a)]^2 \quad (2)$$

The entropic term is asymmetric and has the effect of adding a significant penalty for overshooting an amino acid with a small desired probability. This term has the effect of severely penalizing the formation of stop codons, which will have a desired probability of 0. However, there are other low probability amino acids in our distributions which we want to be represented for reasons detailed in the previous section. We found that the use of this comparison function resulted in many distributions which precluded the appearance of some of these low frequency amino acids. We therefore turned to a second comparison function which had previously looked promising.

$$CompareFun = \sum_{a=1}^{21} (1 - \cos(|P_{des}(a) - P_{calc}(a)|\pi) \qquad (3)$$

As seen in **Figure 4**, the penalties rise slowly for small differences, quickly for medium difference, and then slowly again for large differences. This should have the effect of preferring many small deviations to one large deviation, decreasing the likelihood of a major deviation for a single amino acid. Note that the penalty is not based on a percentage of the desired probability. In other words, the penalty accrued due to realizing a 2% probability at a given position when a 7% probability is desired is the same as realizing a 55% probability when a probability of 60% were desired. Thus as a percentage of the desired probability, the objective function will focus on the dominant amino acid in each distribution.

One potential downside to function 3 is that it does not penalize the introduction of stop codons as heavily as the other function. Each of our distributions sets a target probability of 0% to stop codons, so their introduction will still be penalized. However, we make two slight alterations to further limit the occurrence of stop codons. First, we slightly modify our comparison function so that the penalty for stop codons is twice the normal penalty. Second,

we discard any potential distribution which has a stop codon percentage greater than an arbitrary threshold, which we set to 10%.

With our constrained optimization problem defined, we use a Sequential Least Squares Programming (SLSQP) package to minimize our objective function, using our previously derived twenty amino acid distributions as input. As output, we receive four Super Nucleotides, shown in **Table 2**, and a list of twenty distributions which can be realized using those Super Nucleotides, seen in **Figure 3** and **Supplemental Table 2**. We have gathered key statistics from these distributions in **Table 3.**

The first and most notable characteristic to note is how closely we were able to achieve the desired probability of 60% for each of the twenty amino acids. The actual probabilities for the main amino acid in each distribution range from 56.5% for tryptophan to 68.9% for valine, with an average of 62.2%. These values represent a range of around -6% to +15% of the targeted value, and a mean only 3.7% higher than the ideal target probability. In **Figure** 4, we compare the regions of sequence space sampled by our proposed distributions and the NNK degenerate codon, assuming we create a library which introduces diversity at 10 residues and the initial amino acids follow the same distribution as amino acids in naturally occurring proteins (24). Our proposed distributions should generate a library where over 70% is devoted to sampling sequences that are 4 mutations or less from the initial design. On the other hand, the NNK library would devote almost 90% of the library to sampling designs that are 9 or 10 mutations from the original. Neither library will be able to sample all sequence combinations available by randomizing 10 sites, but our approach should spend significantly more time sampling more meaningful regions of sequence space by concentrating on sequences closer to the original design and more biophysically intuitive substitutions. This success addresses the first deficiency we identified with regards to traditional degenerate codons, ensuring that our

library should sample sequence space closer to the original design that has already shown some evidence of binding or activity.

Next, we look at the other 19 minor amino acids in each distribution to see how well we were able to match the targets we set based upon the data from substitution matrices. To evaluate this, for each distribution we measured the correlation between the desired probabilities for the lower 19 amino acids plus 1 stop codon to the delivered probabilities, as seen in **Table 3**. Most of the distributions correlate nicely with the desired distributions, with 13 of the 20 distributions recording correlations between 0.50 and 0.81. Considering Valine as a typical high-scoring distribution, we see that the top three most desired minor amino acids are isoleucine, leucine, and alanine. The Super Codon produced using the Super Nucleotide numbers (1, 0, 2) from **Table 2** produces a distribution that similarly stresses these amino acids, with these three forming the top three delivered minor amino acids.

To get a better sense of the significance of our correlation values, for each distribution, we calculated 10,000 random distributions for the minor amino acids, and calculated correlation values between these random distributions and our desired distribution. For 15 of our 20 distributions, the correlation value realized by using our Super Codons exceeds 95% of the correlation values obtained between the random distributions and the desired distribution. However, four distributions (serine, tyrosine, cysteine, and tryptophan) have negative correlations, placing them in the bottom half of the correlation values between random distributions and the target distribution. We wondered if these distributions performed poorly due to some inherent incompatibility between the desired distribution and our algorithm, or if their poor performance was a case of the global optimum for all 20 amino acid distributions creating sub-optimal solutions for a few individual distributions. To investigate this, we used our algorithm to target each of the distributions separately so that each distribution had its own set

of Super Nucleotides. The results can be seen in **Supplemental Table 3**. Both tyrosine and serine were able to significantly improve their correlation values under this more generous protocol, with tyrosine going from a correlation of -0.05 to 0.63 and serine going from -0.01 to 0.45. However, both cysteine and tryptophan retained poor correlation scores, indicating that these distributions may be difficult to match by just using a mixture of nucleotides at each codon position. However, for the vast majority of amino acid distributions, our approach of creating four Super Nucleotides was able to create distributions where mutations from an initial amino acid closely followed the distribution found substitution matrices.

The third concern we wanted to address was to make sure each amino acid was represented in each distribution. The lowest probability found in any of our twenty distributions is for methionine in both the cysteine and tyrosine distributions, with a probability of 0.0125%. While a fairly small percentage, in a very modest phage library of size 10^7, we would still expect methionine to be sampled 1250 times at positions using these Super Codons. Thus, even the least frequent amino acid mutation should be present at over a thousand copies for each diversified position in our library.

Finally, we turn to the presence of stop codons. As shown in **Figure 1**, traditional degenerate codons are not immune to the introduction of stop codons. The popular NNK degenerate codon for instance will insert a stop codon at a frequency of 3.13% for instance. If one were to create a library using this degenerate codon at 10 positions, a full 27% of the library would be wasted by premature stop codons. Looking at the distributions from our proposed Super Codons, we see stop codons will be introduced at frequencies ranging from 0.05% for the leucine distribution on up to 8.57% for tryptophan, with an unweighted average of 1.93%. If we weight the various distributions according to how frequently their major, target amino acid occurs in the naturally occurring proteins found in UniProt (24), we get an average of 1.38%.

This means we would expect only about 13% of our library to be wasted on premature stop codons if 10 positions are randomized.

Extending Super Codons to target alternative distributions

While we believe the four Super Nucleotides and twenty target distributions we propose will work well for most gene libraries, we recognize there may be instances when researchers wish to create alternative distributions. To allow for this, we have created a public website at supercodons.degradolab.org which allows researchers to use our algorithm to produce four Super Nucleotides targeting their own distributions. We provide four different ways to specify new target distributions.

BLOSUM-based distributions

This method will produce distributions similar to our proposed distributions. However, the user is able to specify a percentage other than 60% for each distribution's target amino acid. This allows researchers some latitude in setting how far from the initial, designed sequence they would like to sample. By decreasing the percentage of the original amino acid, they will increase the average number of mutations per library member. In addition, this method also permits researchers to specify a subset of amino acid distributions which they care about. Therefore, if a researcher is generating a library targeting ten locations for diversification, it might be worthwhile to create distributions only for the ten amino acids diversified rather than all twenty. By decreasing the number of distributions our algorithm is trying to match, it may be possible to more closely match those distributions which are important for the experiment at hand.

Antibody mutagenesis

One popular use of display libraries is to evolve antibodies which target a given epitope. Antibodies are composed of a conserved scaffold called the constant region and the variable

region responsible for target recognition. The antibody community has collected hundreds of thousands of antibody sequences and have shown that even in the variable regions, different positions have particular preferences for certain amino acids. Thus, when randomizing a position in a given antibody, it may be useful to avoid sampling a given amino acid if that amino acid is almost never seen at that position in any other antibody sequence, while sampling frequently observed amino acids more often. Using standard antibody position naming schemes, we allow the user to specify as many positions as she is interested in, and then use the position-specific amino acid distribution data curated by abYsis (25) as the target distributions. Our algorithm will then generate the best four Super Nucleotides to target those sites as seen in **Figure 5**.

Custom distributions and multiple-alignments

We also allow researchers to simply input their own distributions in comma-separated format, or to drop in a set of multiple alignments. In the case of multiple alignments, our website will identify all sites within the alignment that have conservation less than a user defined threshold (default value of 95%) and then create distributions based upon the frequency of amino acids seen at each position. This may be useful when taking into account evolutionary data or homologous proteins. One group reported using Rosetta (26) to inform the generation of a phage display library by using the sequence profiles of decoys to select degenerate codons (27). These sequences can instead be used directly as input to our server which will then attempt to create four Super Nucleotides that can match the distributions seen at the variable sites, replacing the impreciseness of the traditional degenerate codons.

5.4 Conclusions

The set of twenty amino acid distributions we describe here will be a powerful tool for targeted mutagenesis libraries. Our set of four Super Nucleotides (each Super Nucleotide

composed of a mixture of the four standard nucleotides) is a cost effective way for researchers to create amino acid distributions that closely mirror our derived distributions. By keeping one amino acid in each distribution at a high percentage of 60%, researchers will be able to create libraries which sample sequence space within a few mutations of a starting sequence, rather than spending the bulk of the library sampling sequences much further from the starting design, which are likely less functional. Moreover, other amino acids will be sampled at rates proportional to amino acid transition frequencies observed in the multiple alignments of related proteins which helped create the BLOSUM62 substitution matrix, which we believe should make libraries more efficient. Additionally, no amino acid transition is ever precluded, and thus our method can capture surprising mutations such as a hydrophobic to charged amino acid mutation, that one may not expect a priori (18). And finally, our distributions have a lower rate of stop codon incorporation than other commonly used degenerate codons (**Figure 1** and **Figure 3**), ensuring that ~90% of each library actually produces full-length proteins.

In addition to the public website we have produced, all of our source code is available under GPLv2 license at GitHub: https://github.com/godotgildor/SuperCodons. The command line version of our tool is even more customizable than the website, including allowing the user to try out different objective functions for minimization and varying the number of Super Nucleotides to calculate (not just 4).

Super Codons will provide a powerful tool to the community of protein engineers, allowing researchers to increase the number of directed mutations introduced to their libraries while maintaining larger numbers of functional variants and minimizing the number of premature stop codons. By expanding the number of mutants permitted and focusing the area of sequence space explored, display libraries will become even more powerful tools, enabling proteins with enhanced binding or activity properties to be discovered more readily.

5.5 Methods

Target distribution derivation

To build our twenty target distributions, we first set one amino acid to have a desired probability of 60%. We then aimed to split the remaining 40% in such a way as to mirror amino acid transition probabilities seen in nature. We downloaded the original data used to make the BLOSUM matrices from ftp://ftp.ncbi.nih.gov/repository/blocks/unix/blosum/blosum.tar.Z, and looked at the counts given in the blosum62.out file. That file has a 20 x 20 matrix giving weighted counts for each amino acid pair seen in the multiple alignments used to generate BLOSUM62. If we were making the distribution which set valine to 60% probability for example, then we would look at the counts in blosum62.out for valine. We then would ignore the identity transition, i.e. the valine to valine cell, and scaled the 19 remaining cells so that the sum of their probabilities would = 40%. In other words, each of the 19 remaining cells would be set to:

$$p(aa_i|Val) = c(aa_i|Val) * \frac{40\%}{\sum_{aa_j \neq Val} c(aa_j|Val)}$$

where $c(aa_i \mid Val)$ is the weighted count of amino acid aa_i in the Valine row.

Super Nucleotide and Super Codon Calculation

To find the best set of Super Nucleotides that will match our set of distributions, we set up the problem as a constrained optimization problem. First, we define a Super Nucleotide as one mixture of the four traditional nucleotides:

$$SNt_i = \{\%A, \%C, \%G, \%T\}$$

where the percentages of A, C, G, and T sum up to 100% and are all positive. We then want to find the set of N Super Nucleotides that best allow us to approximate our M desired amino acid distributions

$$TargetDist_i = \{\%Ala, \%Arg, \%Asn, \ldots, \%Val\}$$

and again the percentages sum to 100% and are all positive. Note, in the work described in our paper, N = 4, and M = 20, although our source code leaves these as user defined parameters.

With N Super Nucleotides, one can make N * N * N = N^3 different distributions. Our program will be tasked with minimizing the objective function:

$$Objective\ Function = \sum_{i=1}^{M} \min_{1 \le j \le N^3} CompareFunction(TargetDist_i, Dist_j)$$

that is, for each of our M target distributions, it will find the one of N^3 possible distributions that best matches, and sum the CompareFunction values for all M target distributions.

The function we use to compare our distributions is:

$$CompareFunction = \sum_{a=1}^{21} (1 - \cos(|P_{des}(a) - P_{calc}(a)|\pi)$$

To find the best set of Super Nucleotides that minimizes our objective function, we use the Sequential Least Squares Programming (SLSQP) minimizer available in SciPy v0.12 (28). We supply a pointer to our objective function and an initial guess for the set of Super Nucleotides, and the minimizer then returns a set of Super Nucleotides that is a local minimum of our objective function. Because the final answer depends upon the initial guess, our website currently launches seven separate threads, each with a different, random starting set of Super Nucleotides, and one set where we set one nucleotide to 70% and the remaining nucleotides to 10% in each of the four Super Nucleotides. This last step ensures that we have at least one set of Super Nucleotides where each of the four nucleotides is the majority nucleotide in one Super Nucleotide. We then take the best result of the 8 different minimizations as our final answer. We have found that increasing the total number of threads does not offer much improvement, although we leave that option as a command line switch in our code.

Testing significance of correlations

To test the significance of our correlations between the desired probabilities for the 19 minor amino acids in a given distribution to the delivered probabilities, we created 10,000 decoy distributions by simply drawing a vector of 19 random floating point values from a Dirchlet distribution with $\alpha_i = 1$ for i=1 to 19. We then calculated the correlation coefficient between each of these decoys and the desired distribution, and then calculated the fraction of these values which were less than the correlation between our delivered and the desired probabilities.

5.6 Figures

Codon	Description
NNN	All 20 amino acids
NNK	All 20 amino acids
NWW	Charged, hydrophobic amino acids
RVK	Charged, hydrophilic amino acids
DVT	Hydrophilic amino acids
NTT	Hydrophobic amino acids

Table 1

Common degenerate codons and a brief description. Taken from (10).

Codes use IUB code where each letter represents an equimolar mixture as follows: N: A/C/T/G; K: G/T; R: A/G; V: A/C/G; W: A/T

Super Nucleotide Number	%A	%C	%G	%T	IDT Code
0	5%	6%	5%	84%	(05060584)
1	6%	7%	82%	5%	(06078205)
2	9%	77%	8%	6%	(09770806)
3	83%	5%	7%	5%	(83050705)

Table 2

Optimal Super Nucleotide mixtures found to fit our 20 target amino acid distributions. Each Super Nucleotide is simply a mixture of the 4 standard nucleotides. The last column gives the code used by the firm IDT to specify each given mixture when ordering DNA constructs.

AA Name	Major Amino Acid %	Correlation of Minor Amino Acids	Correlation Significance	Stop Codon %
Gly	67.2%	0.538	99.5%	0.42%
Ala	63.1%	0.633	99.9%	0.42%
Val	68.9%	0.740	100.0%	0.07%
Leu	65.2%	0.720	100.0%	0.05%
Ile	64.8%	0.760	100.0%	0.43%
Met	57.2%	0.752	100.0%	0.24%
Pro	59.3%	0.437	97.3%	0.14%
Phe	63.5%	0.806	100.0%	0.63%
Trp	56.5%	-0.190	21.6%	8.57%
Tyr	62.7%	-0.047	42.7%	7.27%
His	57.5%	0.502	98.7%	0.52%
Lys	62.0%	0.481	98.6%	4.03%
Arg	64.0%	0.268	86.9%	0.51%
Asp	61.3%	0.564	99.5%	0.43%
Glu	59.9%	0.629	99.8%	3.67%
Asn	62.0%	0.549	99.4%	0.43%
Gln	57.5%	0.502	98.8%	4.83%
Ser	65.0%	-0.008	48.6%	1.89%
Thr	63.9%	0.633	99.9%	0.11%
Cys	62.0%	-0.082	36.2%	3.95%

Table 3

Summary of realized amino acid distributions using the Super Nucleotides in **Table 2**. The first column gives the name of the amino acid with a desired probability of 60%. The second column gives the actual probability of that amino acid realized in our distribution. The third column gives the correlation between the desired and realized values for the other 19 amino acids, giving an indication as to the fit of the minor amino acids. The fourth column gives a value for the significance of the given correlation, calculated as described in the **Methods** section. The last column gives the percentage of stop codons present in each of the realized distributions.

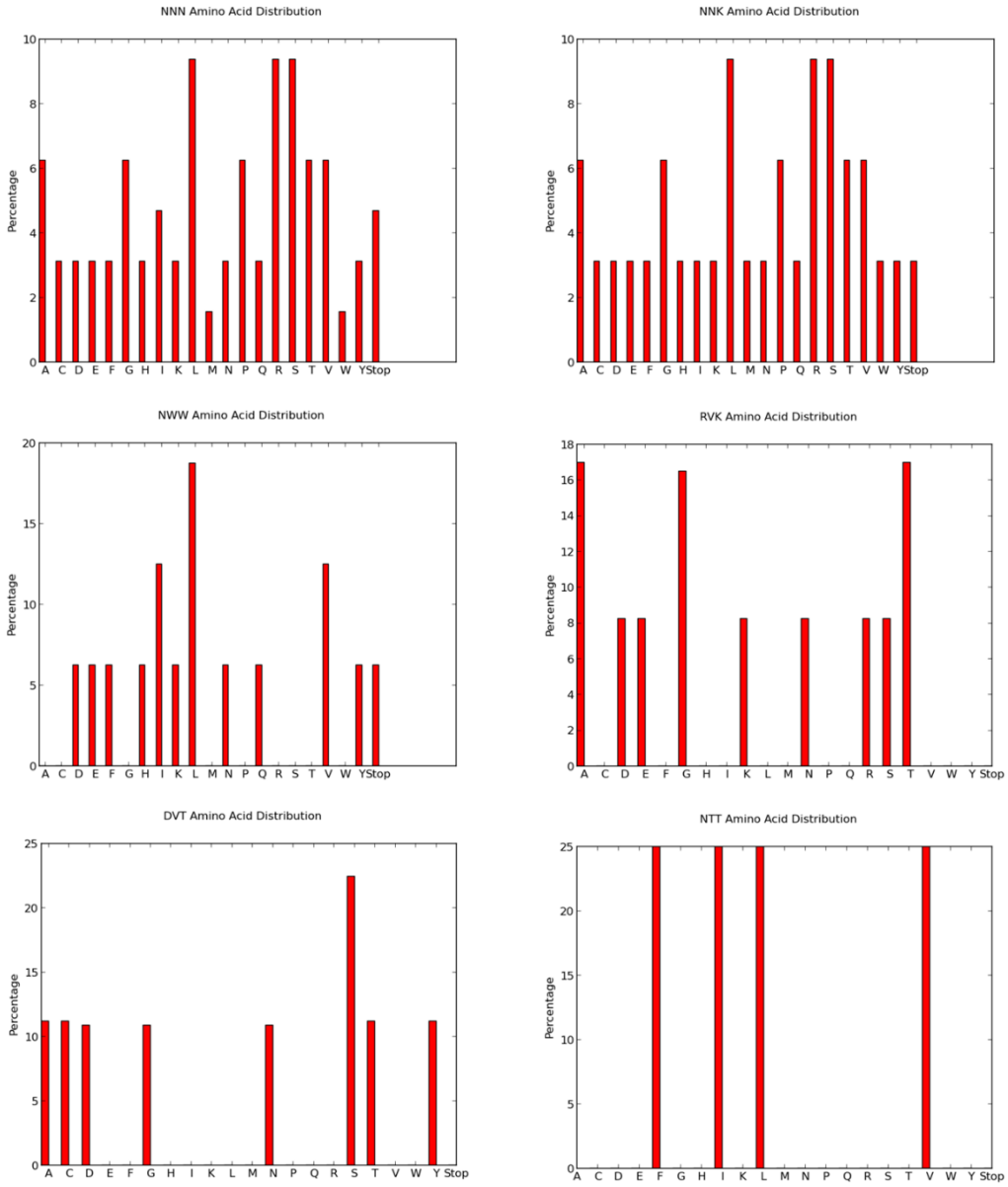


Figure 1

Theoretical amino acid distributions of six commonly used degenerate codons.

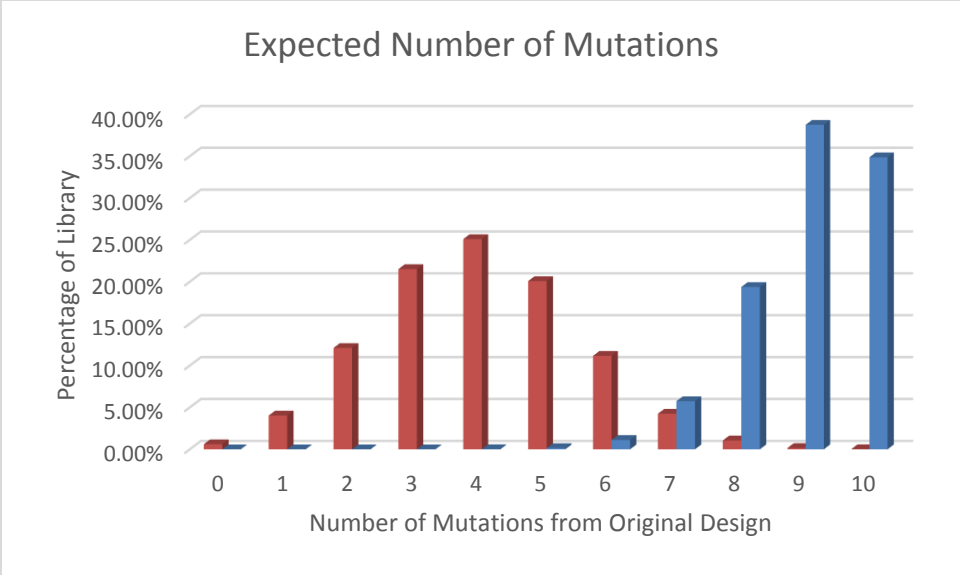
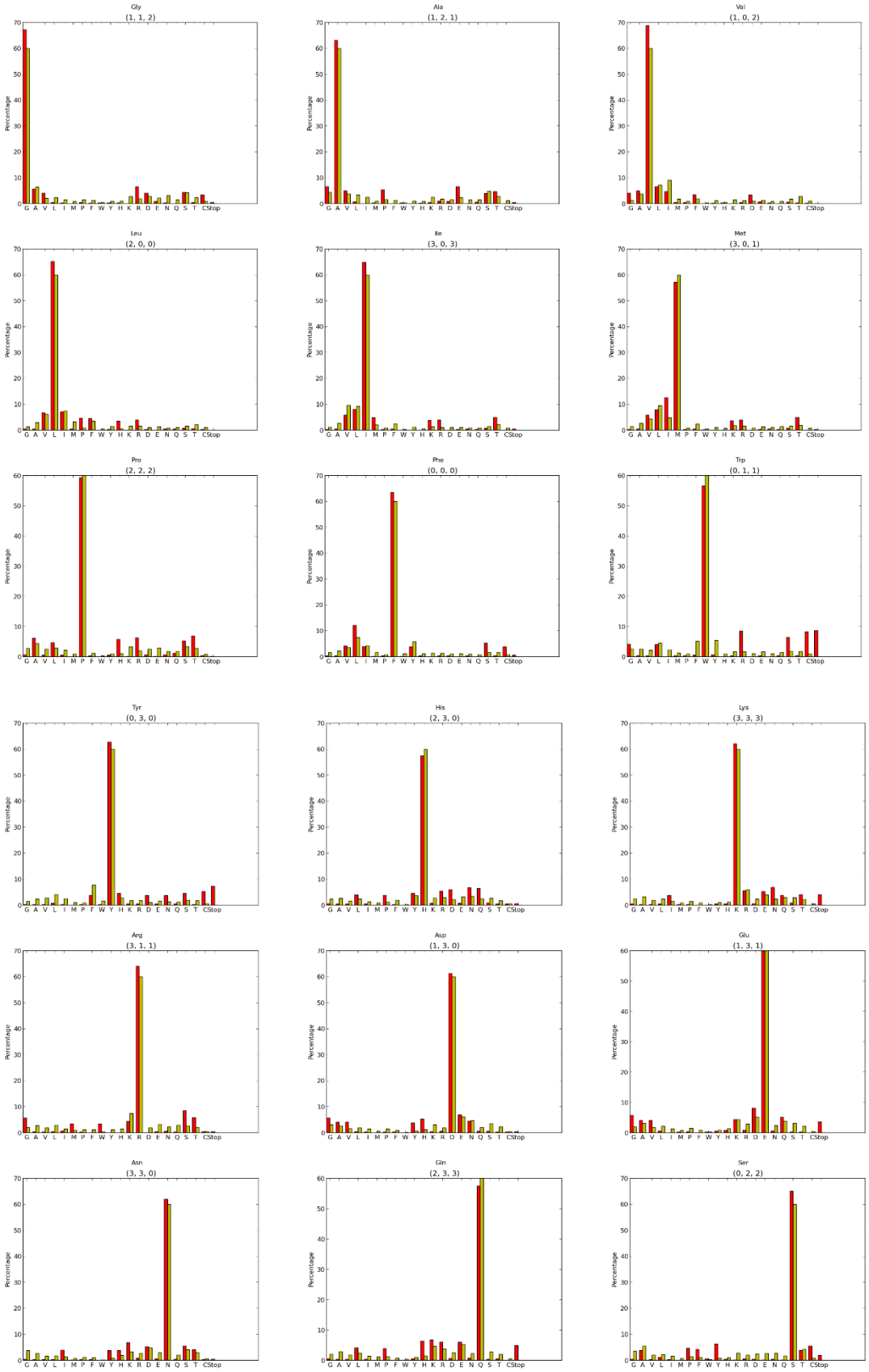


Figure 2

The expected number of mutations from an initial sequence when randomizing 10 positions with

Blue bars - 10% probability of keeping the initial amino acid at each location

Orange bars - 60% probability of keeping the initial amino acid at each location


Gly
(1, 1, 2)
Ala
(1, 2, 1)
Val
(1, 0, 2)
Leu
(2, 0, 0)
Ile
(3, 0, 3)
Met
(3, 0, 1)
Pro
(2, 2, 2)
Phe
(0, 0, 0)
Trp
(0, 1, 1)
Tyr
(0, 3, 0)
His
(2, 3, 0)
Lys
(3, 3, 3)
Arg
(3, 1, 1)
Asp
(1, 3, 0)
Glu
(1, 3, 1)
Asn
(3, 3, 0)
Gln
(2, 3, 3)
Ser
(0, 2, 2)
Percentage
G A V L I M P F W Y H K R D E N Q S T CStop

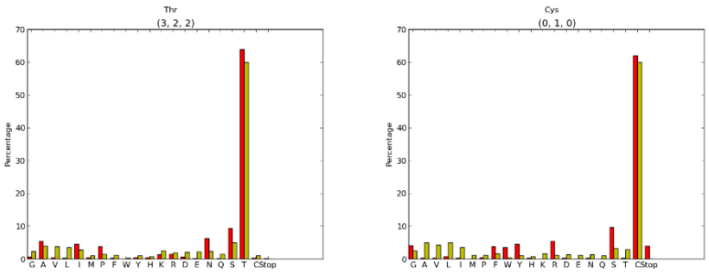


Figure 3

The twenty target amino acid distributions are shown in yellow, while the amino acid distributions able to be realized by the four Super Nucleotides in **Table 2** are shown in red.

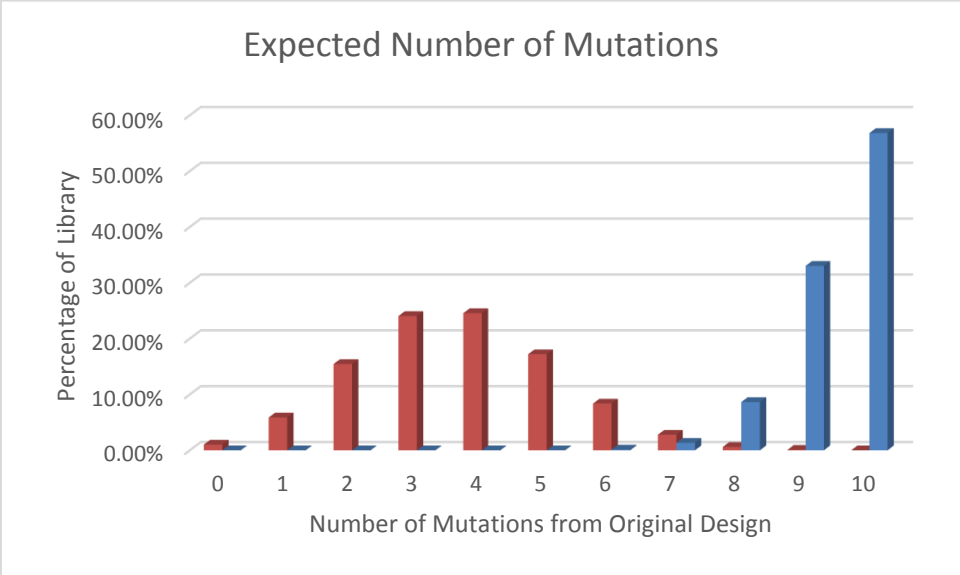


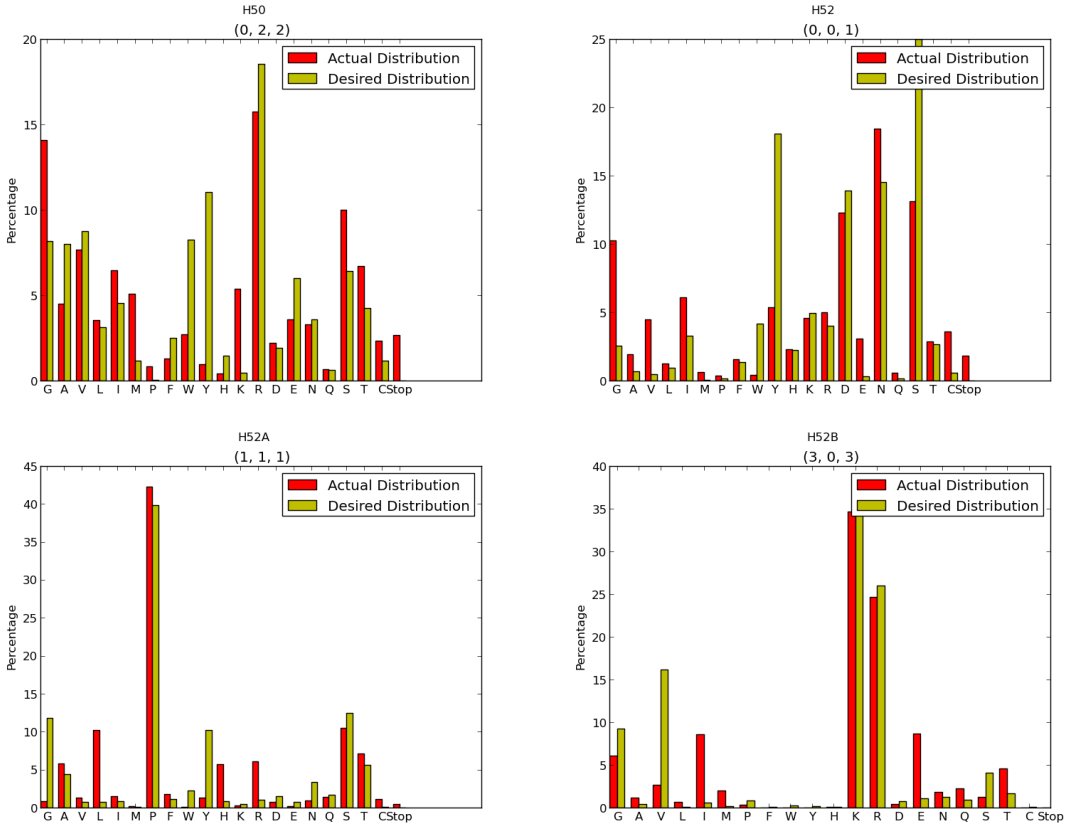
Figure 4

Histograms illustrating what fraction of a library which diversifies 10 residues has the given number of mutations relative to the initial designed sequence.

Blue bars - using the standard NNK degenerate codon

Orange bars - using our proposed Super Nucleotides.

a)



b)

Super NT #	A	C	G	T
0	48%	6%	32%	14%
1	11%	65%	9%	15%
2	18%	14%	44%	24%
3	76%	5%	19%	0%

Figure 5

a) Desired (yellow) and actual (red) distributions for four positions on the CDR H2 of an antibody. Desired distributions are based upon the observed frequency of each amino acid in nature as reported by abYsis (25).

b) The four Super Nucleotides that will allow us to create the distributions in a).

5.7 Supplemental Figures

Distribution Name	G	A	V	L	I	M	P	F	W	Y	H	K	R	D	E	N	Q	S	T	C	Stop
Cys	2.4%	5.0%	4.3%	4.9%	3.4%	1.2%	1.1%	1.6%	0.5%	1.1%	0.7%	1.6%	1.2%	1.3%	1.2%	1.4%	1.0%	3.3%	2.9%	60.0%	0.0%
Asp	3.1%	2.7%	1.6%	1.9%	1.5%	0.6%	1.5%	0.9%	0.2%	0.7%	1.2%	3.0%	2.0%	60.0%	6.1%	4.6%	2.0%	3.5%	2.3%	0.5%	0.0%
Ile	1.1%	2.6%	9.7%	9.2%	60.0%	2.0%	0.8%	2.5%	0.3%	1.1%	0.5%	1.3%	1.0%	1.0%	1.0%	0.8%	0.7%	1.4%	2.2%	0.9%	0.0%
Ser	3.4%	5.6%	2.1%	2.2%	1.5%	0.8%	1.5%	1.1%	0.3%	0.9%	1.0%	2.8%	2.0%	2.5%	2.6%	2.8%	1.7%	60.0%	4.2%	0.9%	0.0%
Gln	2.0%	2.9%	1.7%	2.4%	1.3%	1.1%	1.3%	0.8%	0.3%	1.0%	1.6%	4.6%	3.7%	2.4%	5.3%	2.3%	60.0%	2.8%	2.0%	0.5%	0.0%
Lys	2.4%	3.2%	1.8%	2.3%	1.5%	0.9%	1.5%	0.9%	0.3%	1.0%	1.1%	60.0%	5.9%	2.3%	3.9%	2.3%	2.9%	3.0%	2.2%	0.5%	0.0%
Trp	2.5%	2.5%	2.2%	4.5%	2.2%	1.2%	0.9%	5.2%	60.0%	5.4%	0.9%	1.7%	1.6%	1.0%	1.6%	1.0%	1.4%	1.8%	1.7%	0.9%	0.0%
Asn	3.7%	2.6%	1.6%	1.8%	1.3%	0.7%	1.1%	1.0%	0.2%	0.9%	1.9%	3.2%	2.6%	4.9%	2.9%	60.0%	2.0%	4.1%	2.9%	0.6%	0.0%
Pro	2.8%	4.4%	2.6%	2.9%	2.1%	0.8%	60.0%	1.1%	0.3%	0.9%	1.0%	3.2%	2.0%	2.5%	2.9%	1.8%	1.7%	3.4%	2.8%	0.7%	0.0%
Thr	2.3%	3.9%	3.8%	3.5%	2.8%	1.1%	1.4%	1.2%	0.3%	1.0%	0.8%	2.4%	1.9%	2.0%	2.1%	2.3%	1.4%	4.9%	60.0%	1.0%	0.0%
Phe	1.6%	2.2%	3.5%	7.4%	4.2%	1.6%	0.7%	60.0%	1.2%	5.8%	1.1%	1.3%	1.3%	1.0%	1.2%	1.0%	0.7%	1.6%	1.6%	0.7%	0.0%
Ala	4.4%	60.0%	3.8%	3.4%	2.4%	1.0%	1.6%	1.2%	0.3%	1.0%	0.8%	2.5%	1.8%	1.6%	2.3%	1.5%	1.5%	4.7%	2.8%	1.2%	0.0%
Gly	60.0%	6.4%	2.0%	2.3%	1.5%	0.8%	1.5%	1.3%	0.4%	0.9%	1.1%	2.8%	1.9%	2.8%	2.1%	3.1%	1.5%	4.2%	2.4%	0.8%	0.0%
His	2.3%	2.6%	1.5%	2.3%	1.4%	0.9%	1.1%	1.9%	0.4%	3.6%	60.0%	2.8%	2.9%	2.3%	3.2%	3.4%	2.5%	2.6%	1.8%	0.5%	0.0%
Leu	1.4%	2.9%	6.1%	60.0%	7.4%	3.2%	0.9%	3.5%	0.5%	1.4%	0.6%	1.6%	1.6%	1.0%	1.3%	0.9%	1.0%	1.6%	2.2%	1.0%	0.0%
Arg	2.0%	2.8%	1.9%	2.9%	1.5%	0.9%	1.1%	1.1%	0.3%	1.1%	1.5%	7.4%	60.0%	1.9%	3.2%	2.3%	2.9%	2.7%	2.1%	0.5%	0.0%
Met	1.4%	2.6%	4.4%	9.4%	4.8%	60.0%	0.8%	2.3%	0.4%	1.1%	0.7%	1.7%	1.5%	0.9%	1.3%	1.0%	1.4%	1.6%	1.9%	0.7%	0.0%
Val	1.4%	3.8%	60.0%	7.1%	9.0%	1.7%	0.9%	1.9%	0.3%	1.2%	0.5%	1.5%	1.2%	1.0%	1.3%	0.9%	0.9%	1.8%	2.7%	1.0%	0.0%
Glu	2.0%	3.1%	1.8%	2.1%	1.3%	0.7%	1.5%	0.9%	0.3%	0.9%	1.4%	4.3%	2.8%	5.2%	60.0%	2.3%	3.7%	3.1%	2.1%	0.4%	0.0%

Supplemental Table 1

The twenty target distributions we propose. Each distribution is named for the amino acid present at 60%. The remaining 40% is distributed proportionally based upon amino acid mutation probabilities from the data used to create the BLOSUM 62 substitution matrix.

Distribution Name	G	A	V	L	I	M	P	F	W	Y	H	K	R	D	E	N	Q	S	T	C	Stop
Cys	4.1%	0.4%	0.3%	0.7%	0.2%	0.0%	0.4%	3.8%	3.4%	4.5%	0.3%	0.0%	5.3%	0.3%	0.0%	0.3%	0.0%	9.6%	0.4%	62.0%	3.9%
Asp	5.7%	4.1%	4.1%	0.4%	0.3%	0.0%	0.4%	0.2%	0.0%	3.7%	5.2%	0.5%	0.5%	61.3%	6.8%	4.5%	0.6%	0.6%	0.3%	0.3%	0.4%
Ile	0.4%	0.4%	5.9%	8.0%	64.8%	4.9%	0.3%	0.4%	0.0%	0.0%	0.0%	3.7%	4.0%	0.0%	0.3%	0.4%	0.2%	0.7%	5.0%	0.0%	0.4%
Ser	0.4%	3.9%	0.3%	1.2%	0.3%	0.0%	4.6%	4.2%	0.5%	6.3%	0.4%	0.1%	0.5%	0.4%	0.1%	0.4%	0.1%	65.0%	3.9%	5.6%	1.9%
Gln	0.6%	0.4%	0.4%	4.1%	0.4%	0.0%	3.9%	0.0%	0.0%	0.5%	6.4%	6.7%	6.0%	0.7%	6.0%	0.7%	57.5%	0.4%	0.5%	0.0%	4.8%
Lys	0.5%	0.4%	0.4%	0.5%	3.9%	0.3%	0.3%	0.0%	0.0%	0.4%	0.4%	62.0%	5.6%	0.6%	5.2%	6.9%	3.7%	0.8%	4.2%	0.0%	4.0%
Trp	4.1%	0.4%	0.3%	4.0%	0.0%	0.2%	0.4%	0.5%	56.5%	0.6%	0.0%	0.3%	8.5%	0.0%	0.3%	0.0%	0.3%	6.4%	0.4%	8.3%	8.6%
Asn	0.5%	0.4%	0.4%	0.3%	3.9%	0.2%	0.3%	0.2%	0.0%	3.7%	3.7%	6.9%	0.9%	5.2%	0.6%	62.0%	0.4%	5.5%	4.2%	0.3%	0.4%
Pro	0.6%	6.2%	0.5%	4.7%	0.5%	0.0%	59.3%	0.3%	0.0%	0.4%	5.8%	0.1%	6.3%	0.6%	0.1%	0.7%	1.2%	5.2%	6.9%	0.4%	0.1%
Thr	0.6%	5.4%	0.4%	0.4%	4.6%	0.4%	3.9%	0.2%	0.0%	0.4%	0.4%	1.3%	1.5%	0.5%	0.1%	6.2%	0.1%	9.4%	63.9%	0.3%	0.1%
Phe	0.3%	0.3%	4.2%	12.1%	4.0%	0.2%	0.4%	63.5%	0.2%	3.8%	0.3%	0.0%	0.3%	0.2%	0.0%	0.2%	0.0%	5.3%	0.3%	3.8%	0.6%
Ala	6.6%	63.1%	4.9%	0.7%	0.1%	0.3%	5.4%	0.0%	0.3%	0.1%	0.1%	0.5%	1.0%	0.9%	6.5%	0.1%	0.6%	3.9%	4.6%	0.0%	0.4%
Gly	67.2%	5.7%	4.1%	0.4%	0.3%	0.0%	0.5%	0.2%	0.3%	0.2%	0.3%	0.1%	6.6%	4.1%	0.8%	0.3%	0.1%	4.4%	0.4%	3.4%	0.4%
His	0.6%	0.4%	0.4%	3.9%	0.4%	0.0%	3.9%	0.3%	0.0%	4.5%	57.5%	0.7%	5.5%	6.0%	0.7%	6.7%	6.4%	0.9%	0.5%	0.4%	0.5%
Leu	0.4%	0.5%	6.7%	65.2%	7.2%	0.4%	4.6%	4.5%	0.0%	0.3%	3.5%	0.0%	3.9%	0.4%	0.0%	0.4%	0.4%	0.8%	0.5%	0.3%	0.0%
Arg	5.7%	0.5%	0.4%	0.5%	0.7%	3.4%	0.4%	0.0%	3.4%	0.0%	0.0%	4.4%	64.0%	0.1%	0.4%	0.6%	0.3%	8.5%	5.8%	0.5%	0.5%
Met	0.4%	0.4%	5.9%	7.9%	12.5%	57.2%	0.3%	0.5%	0.2%	0.0%	0.0%	3.7%	3.9%	0.0%	0.3%	0.5%	0.2%	0.8%	5.0%	0.0%	0.2%
Val	4.1%	4.9%	68.9%	6.6%	4.6%	0.4%	0.4%	3.5%	0.0%	0.2%	0.3%	0.1%	0.4%	3.4%	0.7%	0.2%	0.1%	0.5%	0.4%	0.2%	0.1%
Glu	5.7%	4.1%	4.1%	0.6%	0.1%	0.2%	0.4%	0.0%	0.3%	0.5%	0.7%	4.4%	0.9%	8.2%	59.9%	0.6%	5.1%	0.3%	0.3%	0.0%	3.7%
Tyr	0.4%	0.3%	0.3%	0.7%	0.2%	0.0%	0.3%	3.8%	0.3%	62.7%	4.5%	0.4%	0.5%	3.7%	0.4%	3.7%	0.5%	4.5%	0.3%	5.3%	7.3%

Supplemental Table 2

The amino acid distributions which best match the desired distributions given in **Supplemental Table 1** as found by our algorithm.

Distribution Name	G	A	V	L	I	M	P	F	W	Y	H	K	R	D	E	N	Q	S	T	C	Stop
Ala	6.4%	60.8%	5.6%	0.7%	0.2%	0.2%	3.8%	0.3%	0.3%	0.4%	0.2%	0.3%	0.7%	3.2%	4.0%	0.2%	0.2%	7.1%	4.6%	0.3%	0.5%
Arg	4.2%	0.4%	0.5%	2.9%	2.4%	2.2%	2.1%	0.0%	0.3%	0.0%	0.6%	8.4%	61.0%	0.1%	0.9%	0.9%	5.0%	4.1%	3.5%	0.1%	0.4%
Asn	0.8%	0.5%	0.3%	0.2%	2.3%	0.2%	0.4%	0.1%	0.0%	2.3%	4.5%	5.4%	1.1%	6.8%	0.6%	61.9%	0.4%	7.0%	4.9%	0.2%	0.2%
Asp	4.9%	4.9%	3.3%	0.2%	0.4%	0.0%	0.3%	0.1%	0.0%	3.0%	3.7%	0.9%	0.4%	60.6%	8.3%	6.7%	0.5%	0.7%	0.5%	0.2%	0.4%
Cys	6.5%	0.4%	0.6%	0.9%	0.3%	0.0%	0.3%	6.1%	3.3%	4.6%	0.3%	0.0%	5.1%	0.4%	0.0%	0.2%	0.0%	7.1%	0.2%	61.8%	1.7%
Gln	0.7%	0.5%	0.5%	4.0%	0.1%	0.3%	4.0%	0.0%	0.1%	0.1%	4.6%	6.9%	6.1%	0.6%	7.7%	0.5%	61.0%	0.1%	0.5%	0.0%	1.6%
Glu	4.1%	4.9%	4.1%	0.4%	0.4%	0.0%	0.5%	0.0%	0.0%	0.1%	0.7%	6.7%	0.8%	7.6%	61.3%	0.8%	6.0%	0.1%	0.5%	0.0%	0.8%
Gly	60.8%	8.1%	4.1%	0.1%	0.4%	0.2%	0.0%	0.2%	1.6%	0.3%	0.0%	0.6%	4.3%	4.5%	3.6%	0.7%	0.0%	6.1%	1.3%	2.5%	0.6%
His	0.4%	0.2%	0.4%	5.6%	0.5%	0.0%	3.2%	0.5%	0.0%	6.2%	61.0%	0.4%	4.8%	4.6%	0.3%	5.4%	4.6%	0.7%	0.3%	0.4%	0.5%
Ile	0.5%	0.5%	10.4%	9.9%	61.0%	2.5%	0.3%	3.1%	0.0%	0.2%	0.2%	1.8%	1.8%	0.3%	0.3%	1.9%	0.2%	1.8%	2.9%	0.1%	0.3%
Leu	0.3%	0.4%	6.8%	60.8%	8.4%	4.3%	3.0%	4.6%	0.2%	0.3%	1.2%	0.6%	2.8%	0.2%	0.3%	0.3%	2.4%	1.0%	0.8%	0.2%	0.9%
Lys	0.9%	0.5%	0.5%	0.4%	3.8%	0.4%	0.3%	0.0%	0.0%	0.1%	0.4%	61.1%	8.2%	0.6%	6.6%	5.3%	4.4%	0.8%	4.2%	0.0%	1.6%
Met	0.4%	0.4%	6.0%	10.7%	6.7%	60.4%	0.0%	1.2%	0.6%	0.1%	0.0%	3.6%	3.6%	0.0%	0.3%	0.4%	0.0%	1.1%	4.0%	0.1%	0.6%
Phe	0.3%	0.2%	4.1%	9.3%	6.7%	0.7%	0.2%	61.2%	0.4%	6.7%	0.3%	0.1%	0.2%	0.4%	0.0%	0.7%	0.0%	3.7%	0.4%	3.7%	0.7%
Pro	0.5%	5.5%	0.5%	5.7%	0.4%	0.1%	60.8%	0.2%	0.1%	0.3%	3.0%	0.3%	5.8%	0.3%	0.3%	0.3%	3.2%	5.7%	6.2%	0.2%	0.4%
Ser	0.7%	7.9%	0.9%	3.5%	0.4%	0.3%	3.2%	3.9%	2.5%	2.1%	0.1%	0.2%	0.5%	0.3%	0.2%	0.2%	0.1%	61.9%	6.3%	3.0%	1.8%
Thr	0.5%	5.9%	0.8%	0.7%	5.0%	3.2%	3.0%	0.3%	0.1%	0.3%	0.2%	4.3%	2.8%	0.4%	0.4%	3.9%	0.2%	6.8%	60.7%	0.2%	0.4%
Trp	6.4%	0.5%	0.9%	8.7%	0.1%	0.7%	0.0%	0.7%	63.2%	0.2%	0.0%	0.2%	5.2%	0.0%	0.2%	0.0%	0.0%	5.5%	0.4%	4.8%	2.4%
Tyr	0.2%	0.3%	0.7%	1.4%	0.6%	0.0%	0.4%	10.3%	0.1%	62.0%	6.1%	0.1%	0.3%	3.8%	0.1%	3.8%	0.2%	4.3%	0.3%	3.2%	1.9%
Val	2.2%	5.0%	60.5%	8.1%	10.0%	2.6%	0.4%	2.8%	0.0%	0.2%	0.2%	0.5%	0.4%	2.1%	2.2%	0.4%	0.2%	0.7%	1.1%	0.1%	0.3%

Supplemental Table 3
The best calculated distributions when each distribution was treated separately.

AA Name	Natural Abundance of AA
Ala	8.25%
Arg	5.53%
Asn	4.06%
Asp	5.45%
Cys	1.37%
Gln	3.93%
Glu	6.75%
Gly	7.07%
His	2.27%
Ile	5.96%
Leu	9.66%
Lys	5.84%
Met	2.42%
Phe	3.86%
Pro	4.70%
Ser	6.57%
Thr	5.34%
Trp	1.08%
Tyr	2.92%
Val	6.87%

Supplemental Table 4

Natural abundance of each amino acid in the complete UniProt/Swiss-Prot database per (24).

5.8 References

1. *Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin.* **Fleishman, Sarel J., et al.** 2011, Science, pp. 816-821.

2. *High-affinity Human Antibodies from Phage-displayed Synthetic Fab Libraries with a Single Framework Scaffold.* **Lee, Chingwei V., et al.** 2004, Journal of Molecular Biology, pp. 1073-1093.

3. *Design and Use of a Phage Display Library.* **Pini, Alessandro, et al.** 34: Journal of Biological Chemistry, 1998, Vol. 273. pp 21769-21776.

4. *Nature-inspired design of motif-specific antibody scaffolds.* **Koerber, JT, et al.** s.l. : Nature Biotechnology, 2013. pp 916-921.

5. *A general strategy for the evolution of bond-forming enzymes using yeast display.* **Chen, Irwin, Dorr, Brent M. and Liu, David R.** 28: Proceedings of the National Academy of the Sciences, 2011, Vol. 108. pp 11399-11404.

6. *Mechanism-Based Phage Display Selection of Active-Site Mutants of Human Glutathione Transferase A1-1 Catalyzing SNAr Reactions.* **Hansson, Lars O., Widersten, Mikael and Mannervik, Bengt.** 37: Biochemistry, 1997, Vol. 36. pp 11252-11260.

7. *A Single Mutation in a Regulatory Protein Produces Evolvable Allosterically Regulated Catalyst of Nonnatural Reaction.* **Moroz, Olesia V., et al.** s.l. : Angewandte Chemistry, 2013, Vol. 125. pp 6246-6249.

8. *In vitro display technologies: novel developments and applications.* **Amstutz, Patrick, et al.** s.l. : Current Opinion in Biotechnology, 2001, Vol. 12. pp 400-405.

9. *Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.* **Cornish-Bowden, A.** 13: Nucleic Acids Research, 1985, Vol. 10. pp 3021-3030.

10. *Identifying sepcificity profiles for peptide recognition modules from phage-displayed peptide libraries.* **Tonkikian, R, et al.** s.l. : Nature Protocols, 2007, Vol. 2. pp 1368-1386.

11. *Efficient and flexible access to fully protected trinucleotides suitable for DNA synthesis by automated phosphoramidite chemistry.* **Zehl, A, et al.** 23: Chemical Communications, 1996. pp 2677-2678.

12. *Combination of DMT-mononucleotide and Fmoc-trinucleotide phosphoramidites in oligonucleotide synthesis affords an automatable codon-level mutagenesis method.* **Gaytan, P, et al.** 9: Chemistry & Biology, 1998, Vol. 5. pp 519-527.

13. *Optimizing doped libraries by using genetic algorithms.* **Tomandl, D., Schober, A. and Schwienhorst, A.** 1: Journal of Computer-Aided Molecular Design, 1997, Vol. 11. pp 29-38.

14. *Scoring functions for computational algorithms applicable to the design of spiked oligonucleotides.* **Jensen, L.J., et al.** s.l. : Nucleic Acids Research, 1998, Vol. 26. pp 697-702.

15. *Combinatorial codons: a computer program to approximate amino acid probabilities with biased nucleotide usage.* **Wolf, E. and Kim, P.S.** s.l. : Protein Science, 1999, Vol. 8. pp 680-688.

16. *Designing gene libraries from protein profiles for combinatorial protein experiments.* **Wang, W and Saven, J.** 21: Nucleic Acids Research, 2002, Vol. 30. pp e120.

17. *Why High-error-rate Random Mutagenesis Libraries are Enriched in Functional and Improved Proteins.* **Drummond, D.A., et al.** s.l. : Journal of Molecular Biology, 2005, Vol. 350. pp 806-816.

18. *Kemp elimination catalysts by computational enzyme design.* **Röthlisberger, D., et al.** s.l. : Nature, 2008, Vol. 453. pp 190-195.

19. *Amino acid substitution matrices from protein blocks.* **Henikoff, S. and Henikoff, J.G.** 22: Proceedings of the National Academy of Sciences, 1992, Vol. 89. pp 10915-10919.

20. *Non-Darwinian Evolution.* **King, J.L. and Jukes, T.H.** 3881: Science, 1969, Vol. 164. pp 788-798.

21. *Optimal codon randomization via mathematical programming.* **Nov, Y. and Segev, D.** s.l. : Journal of Theoretical Biology, 2013, Vol. 335. pp 147-152.

22. *Optimizing doped libraries by using genetic algorithms.* **Tomandl, D., Schober, A. and Schwienhorst, A.** s.l. : Journal of Computer-Aided Molecular Design, 1997, Vol. 11. pp 29-38.

23. *Scoring functions for computational algorithms applicable to the design of spiked oligonucleotides.* **Jensen, L. J., et al.** 3: Nucleic Acids Research, 1998, Vol. 26. pp 697-702.

24. UniProtKB/Swiss-Prot protein knowledgebase release 2013_09 statistics. *ExPASy.* [Online] [Cited: October 3, 2013.] **http://web.expasy.org/docs/relnotes/relstat.html.**

25. abYsis. [Online] [Cited: September 1, 2013.] **http://www.bioinf.org.uk/abysis/index.html.**

26. *Practically useful: what the Rosetta protein modeling suite can do for you.* **Kaufmann, K.W., et al.** 49: Biochemistry, 2010, Vol. 13. pp 2987-2998.

27. *Engineering a protein–protein interface using a computationally designed library.* **Guntas, G., Purbeck, C. and Kuhlman, B.** 45: Proceedings of the National Academy of Sciences, 2010, Vol. 107. doi: 10.1073/pnas.1006528107.

28. Eric Jones, Travis Oliphant, Pearu Peterson and others. *SciPy: Open Source Scientific Tools for Python.* [Online] 2001. http://www.scipy.org/.

29. Cirino, Patrick C., Mayer, Kimberly M. and Umeno, Daisuke. Generating Mutant Libraries Using Error-Prone PCR. [book auth.] **Frances H. Arnold and George Georgiou.** *Directed Evolution Library Creation.* Totowa, New Jersey : Humana Press, 2003.

30. *An efficient one-step site-directed and site-saturation mutagenesis protocol.* **Zheng, Lei, Baumann, Ulrich and Reymond, Jean-Louis.** 14: Nucleaic Acids Research, 2004, Vol. 32. pp e115.

31. *Automated design of degenerate codon libraries.* **Mena, M. A. and Daugherty, P. S.** 12: Protein Engineering, Design & Selection, 2005, Vol. 18. pp 559-561.

Chapter 6

A Real-Time All-Atom Structural Search Engine for Proteins

6.1 Abstract

Protein designers use a wide variety of software tools for *de novo* design, yet their repertoire still lacks a fast and interactive all-atom search engine. To solve this, we have built the Suns program: a real-time, atomic search engine integrated into the PyMOL molecular visualization system. Users build atomic-level structural search queries within PyMOL and receive a stream of search results aligned to their query within milliseconds. This instant feedback cycle enables a new "designability"-inspired approach to protein design where the designer searches for and interactively incorporates native-like fragments from proven protein structures. We demonstrate the use of Suns to interactively build protein motifs, identify scaffolds compatible with hot-spot residues. The official web site and installer are located at http://www.degradolab.org/suns/ and the source code is hosted at https://github.com/godotgildor/Suns (PyMOL plugin, BSD license), https://github.com/Gabriel439/suns-cmd (command line client, BSD license), and https://github.com/Gabriel439/suns-search (search engine server, GPLv2 license).

This chapter has been submitted to PLoS Computational Biology and also appears in the dissertation of Gabriel Gonzalez. Gabriel Gonzalez is the first author, Brett T. Hannigan is a co-author, and William F. DeGrado is the corresponding author. I developed the client-side implementation of the search program, focused on the importance of interactive design, contributed algorithmic insights to the design of the server-side portion, used Suns to analyze the hemagglutinin hot-spot residues proposed by Fleishman et al., and contributed portions of the text. Gabriel Gonzalez had the initial idea for the search engine, was chief architect of the search algorithm, and wrote the discussion, design, and results sections.

6.2 Introduction

Protein structural bioinformatics rapidly approaches a big data crisis as the last decade has witnessed a dramatic increase in protein structure depositions. In 1993 researchers had just over 23,000 searchable structures at their disposal in the Protein Data Bank (PDB), while today we have over 94,000 [1]. This rapid structural expansion could inform protein design, structure determination, and structure prediction by providing numerous examples of native-like structural interactions in exquisite detail, but researchers lack high-powered computational tools to intelligently explore large structural data sets in detail.

One of the first popular protein structural search tools developed for this purpose was Dali by Holm and Sander [2]. Dali uses distance maps formed by calculating pairwise α-carbon distances to form a two-dimensional representation of a three-dimensional protein. Regions of similarity between two distance maps correspond to similar substructures in their respective proteins. Holm and Sander used Dali to create the Families of Structurally Similar Proteins (FSSP) database [3], which aligns substructures across entries in the Protein Data Bank (PDB) to form families and subfamilies of common folds. Researchers commonly use Dali to compare protein folds and infer homology [4-6].

The more recent MaDCaT search program [7] also uses α-carbon distance maps to search for similar protein backbone arrangements. However, where Dali uses a heuristic approach to detect structural similarity, MaDCaT takes a query backbone structure or motif and finds globally optimal structural matches within an entire structural database. This approach makes MaDCaT ideal for finding the best matches to frequently occurring motifs. These "designable" motifs promise to be excellent design scaffolds, and MaDCaT applied this approach to design a viral-like protein coat for carbon nanotubes from designable interactions [8].

Both Dali and MaDCaT return results after a several minutes of searching. For greater speed, Shyu et. al. developed ProteinDBS [9] in order to provide the first real-time protein backbone search. They use image processing techniques to extract a set of features from α-carbon distance maps and organize their structural database into a tree, allowing quick traversal and parallelism during searches. These optimizations allow them to return search results nearly instantly, but they limit themselves to searching for backbone α-carbons.

We required an all-atom search engine to guide the protein design process, so that we could search for proteins with similar active sites or binding motifs, explore protein scaffolds that can host a specific motif, and discover atomic-scale supporting interactions.

The state of the art for all-atom search is Erebus [10], which permits all-atom rigid substructure searches, but this is insufficient for our design purposes because we desired an interactive search process. Several bottle-necks in the Erebus search workflow impede a fluid design process, including time-consuming assembly of search queries, long search delays, and a web interface for retrieving results.

A truly interactive search tool must remove every single one of these bottlenecks to bring the feedback loop down from minutes to seconds and permit users to rapidly explore multiple design alternatives iteratively in atomic detail. Improved speed and faster feedback lets researchers to ask more sophisticated questions, explore structures more intelligently, and use limited collaboration time more efficiently.

The Suns protein search engine makes it easy to search and browse a database of protein structures at the atomic level. To our knowledge, Suns is the first real-time all-atom structural search engine and also the first to integrate seamlessly into the popular molecular visualization program PyMOL [11], so that researchers can easily click on motifs of interest, click search, and view aligned results within a fraction of a second. We expect Suns to inform and

guide protein design, modeling, and structure determination by lowering the entry barrier to structural search so that it becomes a staple of every structural biologist's toolbox rather than a tool limited to programmers.

6.3 Design and Implementation

Overview

Our structural search engine greatly resembles a web search engine, even though these two types of engines index different types of data: web search engines commonly index linear text strings whereas our search engine indexes three-dimensional protein structures. Despite these differences, we still borrow many principles from web search engines [12] to improve search speed:

1. Divide structures into structural "pages" (3-D volumes) analogous to web pages
2. Divide these "pages" into structural "words" (chemical motifs) analogous to textual words
3. Create a forward index that matches sets of structural words to structural pages
4. Perform slower and more accurate filters after the fast forward index lookup
5. Return only as many results as requested to avoid unnecessary computation

Forward Index

Web search engines derive much of their speed by preprocessing the data set using a forward index that matches words to web pages [12]. The search engine can then tokenize each query into words and consult the forward index to rapidly return all pages that contain every word in the user's search query. Protein search engines can copy this trick, but they must first decide what volume size corresponds to a "page" and what chemical motifs correspond to "words".

Two opposing considerations constrain the choice of page and word size. The forward index resolves pages solely by their word counts, so larger words and smaller pages lead to more unique word counts per page and improves the selectivity of the forward index. However, users prefer the exact opposite: smaller words and larger page sizes increase the power and flexibility of user search queries. Therefore, optimizing a structural search engine requires balancing user needs against the efficiency of the forward index.

We select a compromise suitable for atomic-level search queries: we restrict structural pages to cubes 15 Å wide and we define structural words to be connected chemical substructures ranging from 2 atoms (a hydroxyl) to 9 atoms (an indole ring) (**Figure 1**). Our choice of page size assumes that larger structural patterns of interest can be reduced to a network of bridging local interactions below the 15 Å length scale. Similarly, our choice of word size assumes that users will accept a modest restriction on search queries to groups of chemical motifs instead of groups of atoms. Like web search engines, we permit searches for multiple disconnected words, allowing users to assemble complex queries from these simple chemical building blocks.

Structural words

We specify structural words using PDB files, which contain the specific residue and atom types to match. For example, one structural word consists of a single PDB file containing the Cα-Cβ-Cγ that links the phenyl group of phenylalanine to its backbone atoms. When users search for the three-carbons in phenylalanine's linker, their searches will not match tyrosine's linker, nor will they match three connected ring carbons within a phenylalanine. This allows the search index to optionally resolve motifs that are otherwise chemically identical [13].

Structural words may also match more than one protein element, and in those cases we use multiple PDB files to specify the structural word: one PDB file per matching chemical motif.

For example, one motif we index is a carboxylate, specified using two PDB files we created: one for glutamate's carboxylate and another for aspartate's carboxylate. User search queries for carboxylates will match either of these two groups.

The choice of structural words is customizable and for our public-facing server we select a default set of substructures appropriate for general-purpose searches (**Supplementary Table 1**). The most important searchable substructure matches the four backbone atoms for any protein residue, which permits geometrically exquisite backbone searches that specify all backbone atoms and torsion angles. We partition flexible residues such as lysine and methionine into two separate words, and also isolate important chemical moieties into their own words, such as imidazole and guanidinium groups. Some chemical moieties are shared between residues, such as the hydroxyl group, which matches serine, threonine, and tyrosine. However, every residue except glycine possesses at least one unique structural word so that users can restrict searches to a specific residue.

Database

Our forward index is formally a *record level* inverted index that converts sets of words to matching pages. We supplement the forward index with a custom in-memory database that stores two pieces of information necessary to complete the search. First, the database stores correspondences between words in the forward index and atoms in each structural page. Second, the database also keeps compact representations of every structural page suitable for returning as search results

When the forward index produces a matched page, the database remembers which atoms in that page correspond to the words advertised in the forward index. Sometimes the page contains more instances of a given word than the user required, such as when the user searches for two peptide bonds, and the page contains five. The page must try out every valid

permutation of words that match the user's query, and the forward index minimizes the number of permutations by prioritizing pages that closely match the minimum required word count.

Alignment and RMSD

Suns uses the Kabsch algorithm [14] to rapidly align each permutation to the user's search query. The Kabsch algorithm requires an exact atom-for-atom correspondence between the user's search query and a candidate motif, and Suns compiles this correspondence from precomputed atomic correspondences for each stored motif in the custom database. After alignment, the search engine only returns search results that match the search query within a specified root-mean-square deviation (RMSD) cutoff.

For each result below the RMSD cutoff, Suns aligns the matching page to the search query and return the page as the search result. If a page contains multiple matches Suns aligns each match separately and returns them as separate results. This superimposes every search result and context on the original query for ease of visual comparison and downstream post-processing.

Streaming results

The user may dial in the stringency of desired matches by tuning the RMSD cutoff. The search engine will immediately stream any result within this cutoff, which allows the user to begin visualizing results before the search has completed, improving interactivity.

Additionally, the search protocol requires the user to specify the number of desired results up front. While the user may request an unlimited number of results in theory, in practice the search clients default to 100 search results, similar to how a web search engine will default to 10 search results. This allows the search engine to stop processing the request after supplying the specified number of results, which reduces server load. Also, the search engine

may also optionally specify a search timeout to further reduce server load for users that request a large number of search results.

Data set

The public search engine uses PISCES [15] as the non-redundant protein structure data set, selecting a 20% sequence identity, 1.6 Å resolution, and 0.25 R-factor cutoff, which currently corresponds to 2058 chains. The search engine's available memory limits how many structures it can index, and our stress tests on the largest PISCES data set (90% identity, 3.0 Å, 1.0 R-factor cutoff, 24,218 chains) required 89 GB of memory or 1 GB of memory per 272 protein chains.

6.4 Results

Building motifs

Suns lets users explore the "designable" space of protein motifs by expanding on small initial fragments, such as building a helix N-terminal capping motif beginning from a single guanidinium group. One might begin by searching on the guanidinium fragment from an arginine, which recruits a cluster of nearby carboxylates forming a salt bridge with the arginine (**Figure 2A**). Adding one of these carboxylates to the search query refines the motif further, revealing a preferred rotamer for the arginine when interacting with a carboxylic acid (**Figure 2B**), and adding a preferred rotamer to the search query crystallizes a complete N-terminal capping motif (**Figure 2C**).

The large number of close geometric matches to the final search query suggests that this is a highly "designable" motif. Incremental searching allows users to rapidly explore and prototype designable native-like interactions like these with very little prior knowledge in protein folding or biophysics. Moreover, a user can discover the motif by gradually refining a specification rather than specifying all the necessary interactions up front. This benefits people

who may not even know what designable interactions look like and simply wish to explore what options they have available.

The salt bridge we built this way also matches one of many newly discovered salt bridges by Donald et. al (Figure 8 of [16]). However, we identified this without requiring a curated database of salt bridge and without using a specialized algorithm built to detect electrostatic pairs. We also obtain detailed information from the superimposition of results, which allows us to visualize the structural variability of this salt bridge motif on a per-atom basis.

Discovering motifs

In addition to designing novel motifs, Suns allows users to search for scaffolds which contain motifs similar to a given query. For instance, we were interested in discovering what other structures contained "nest" motifs similar to the one used in chapter 3 to develop the antibody module that binds phosphate. We took the starting structure of the peptide bound antibody that contained a "nest" motif and chose three consecutive backbone segments to be used as a Suns query (**Figures 3A** and **3B**). The resulting matches not only clustered well on our three query backbone segments, but also followed a similar backbone arrangement N-terminal to the query, as can be seen in **Figure 3C**. In **Figure 3D** we view the matching structures in cartoon representation, and immediately see that our query motif is frequently seen in alpha-helices, despite the fact that the query itself is just a single loop. This is a fascinating result because it is known that aspartic acid residues are a popular means of capping alpha-helices [18], so in essence, the initial antibody structure has a single turn of an alpha-helix and then uses the aspartic acid in the bound peptide to "cap" this turn. Moreover, it is also known that phosphate groups can frequently bind near the amino-terminus of alpha-helices [19], providing a further explanation for our success in modifying this loop to bind phosphate.

Assembling larger fragments

Users can build tertiary interactions for proteins as well. To demonstrate this, we search for a valine from glucose binding protein and grow that into three small β strands with three residues per strand.

Beginning from an interior valine from glucose-binding protein, we seed the two adjacent β strands with highly populated residue clusters on each side corresponding to a valine and tyrosine (**Figure 4A**). To grow the three β strands in both directions, we search for pairs of residues at a time to identify new clusters of residues within the search results that we can insert into the sheet (**Figure 4B**). The PyMOL search client permits a qualitative inspection of residue preference at selected positions by cycling through visualizing each residue type. This process not only provides a rough measure of residue preference, but also reveals rotameric preference, the kind of detailed information that a sequence logo would not reveal.

We repeat this process of iteratively searching for pairs of residues at a time and incorporating clusters from the search results until we assemble a native-like fragment of a sheet where almost every residue originates from a unique protein structure (two disconnected threonines were inadvertently drawn from the same structure). This then provides α-carbon coordinates that we feed into the backbone search engine MaDCaT [8], which finds suitable scaffolds to incorporate this fragment. One MaDCaT search result greatly resembles the β sheet built using Suns (**Figure 4C**). This illustrates how the local search capabilities of the Suns search engine complement existing coarse-grained search tools by bridging the gap between the world of smaller atomic interactions and the world of larger secondary-structure interactions.

Connecting hot-spot residues

Suns can also be used to find scaffolds compatible with specified residues to provide an alternative implementation of the hotspot residue approach to design [16]. The user can select

the hotspot of interest within PyMOL, search, and find all proteins in the PDB that position the given hot spot residues in the specified geometry.

For example, Suns recapitulates the local backbone of a designed hemagglutinin binder [16]. **Figure 5A** illustrates how searching for fragments of the original hotspot residues reveals a prominent cluster of α helices matching the designed protein structure, indicating that the secondary structure of the interface could have been predicted solely from designability.

Not every hotspot search will return a single solution for the backbone. Sometimes searching for disembodied residues will reveal multiple distinct ways to thread the backbone between them (**Figure 5B**).

6.5 Availability and Future Directions

We initially built Suns to guide the protein design process, but we are releasing it as a general purpose search engine so that others may reuse it for applications we did not previously anticipate.

The Suns plugin for PyMOL is available at www.degradolab.org/suns, which also includes a tutorial on how to install and use the library. The source code for the client is available separately at https://github.com/godotgildor/Suns under a BSD license.

Users can also automate searches using a command line tool, available at https://github.com/Gabriel439/suns-cmd under a BSD license. Users who wish to incorporate Suns within an automated workflow should use this client instead.

The source code for the search engine is located at https://github.com/Gabriel439/suns-search under a GPLv2 license. Users should report bugs or request new features using the issue tracker at https://github.com/Gabriel439/suns-search/issues or by contacting the Suns mailing list at suns-search@googlegroups.com.

Currently the public search engine only indexes protein structures. We also plan to add support for ligand search queries so that Suns can be used for drug design. While this paper describes a protein-specific application of the search engine, the underlying algorithm can be readily generalized to ligands and other macromolecules.

6.6 Figures

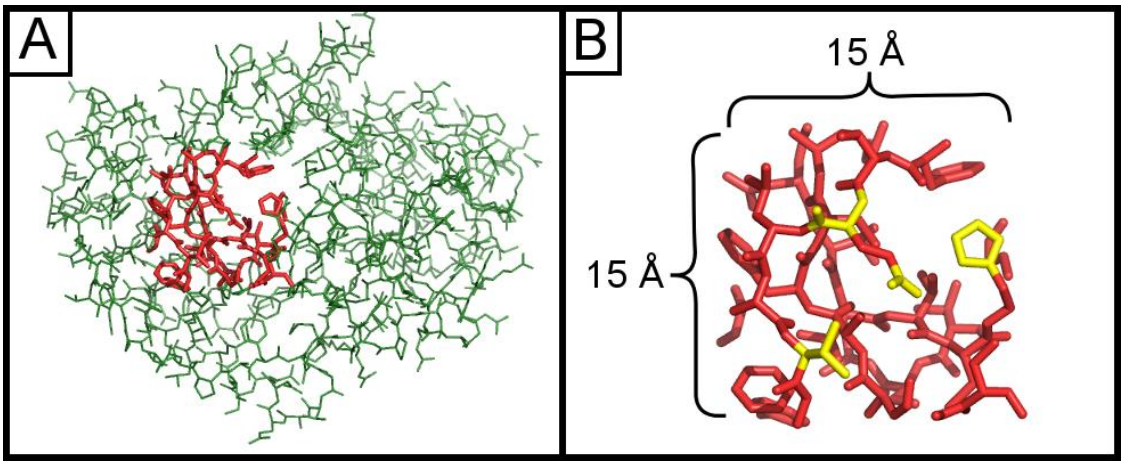


Figure 1 - Subdivision of protein structures.

(A) An interior page highlighted in red from a protein of unknown function (PDB ID = 2FSQ), illustrating the maximum scale of search queries.

(B) Example words (chemical motifs) within the same page highlighted in yellow. Pages are 15 Å x 15 Å x 15 Å cubes.

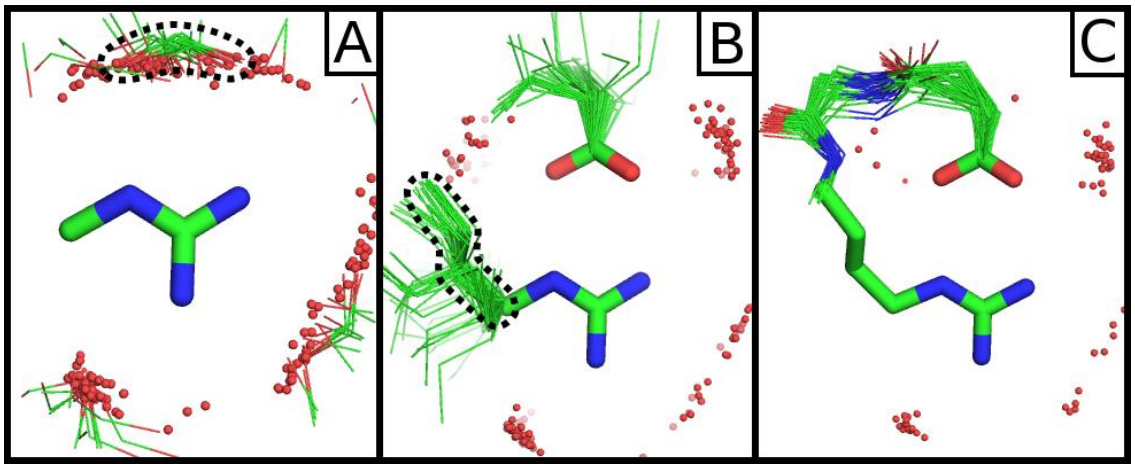


Figure 2 - Incremental assembly of a motif.

(A) An initial search for a guanidinium fragment reveals a cluster of nearby carboxylates. (B) Refining the search with one carboxylate from the results reveals a specific linker preference for both the aspartate and arginine involved in the salt bridge. (C) Adding the most common linker for arginine and repeating the search reveals that this salt bridge is part of an N-terminal capping motif. Search queries are represented as thick sticks and search results are shown as thin lines. Dashed lines highlight clusters in the search results, which are filtered to show the specific residue fragments of interest and neighboring water molecules within 3.0 Å as red spheres. Search parameters and fragments listed in **Supplementary Table 2**.

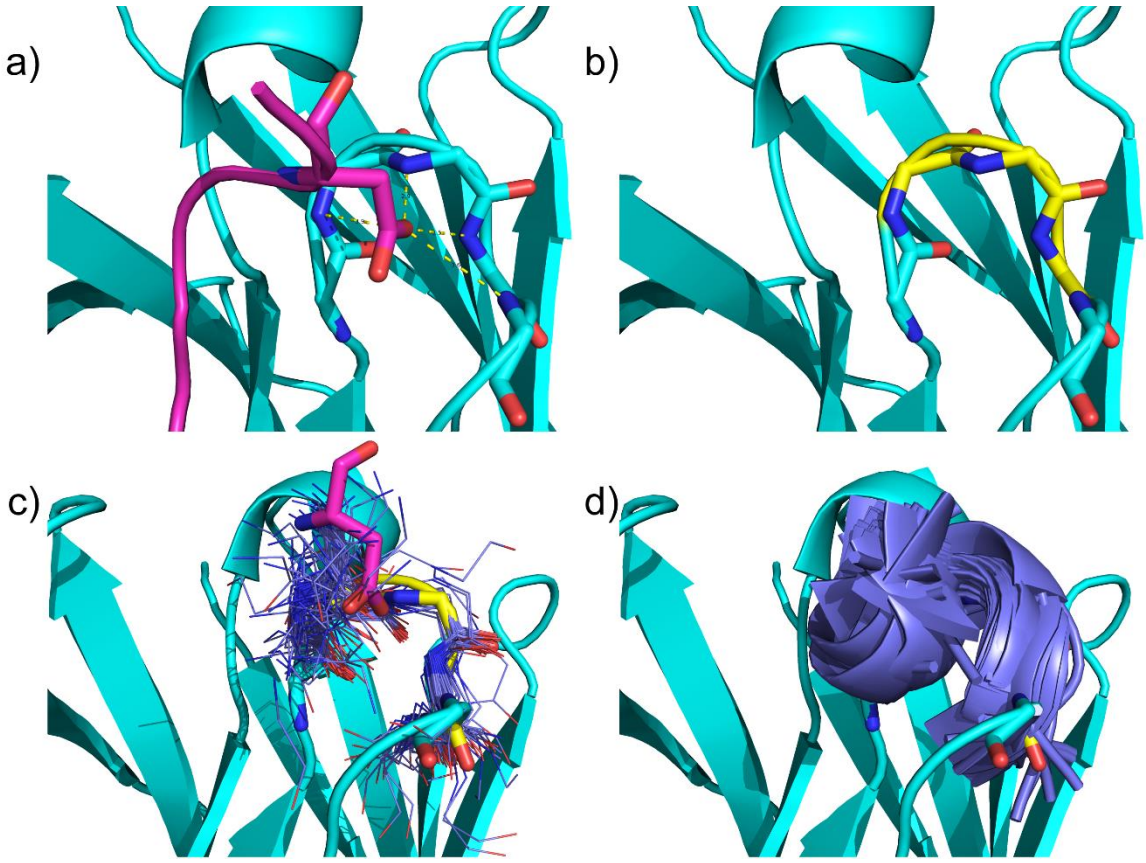


Figure 3 – Identifying alternative "nest"-like motifs

(A) The initial antibody structure (cyan) with select backbone atoms of CDR H2 shown in stick representation. The backbone nitrogen atoms are within hydrogen-bonding distance of one of the carboxyl oxygens of aspartic acid on the bound ligand (magenta). (B) The Suns query segment is highlighted in yellow. (C) Suns results showing matching segments in blue-lines. Note not only clustering over the initial query, but also a looser clustering N-terminus to the query. (D) The Suns results shown in cartoon representation reveals that the query segment is often found in an alpha-helix.

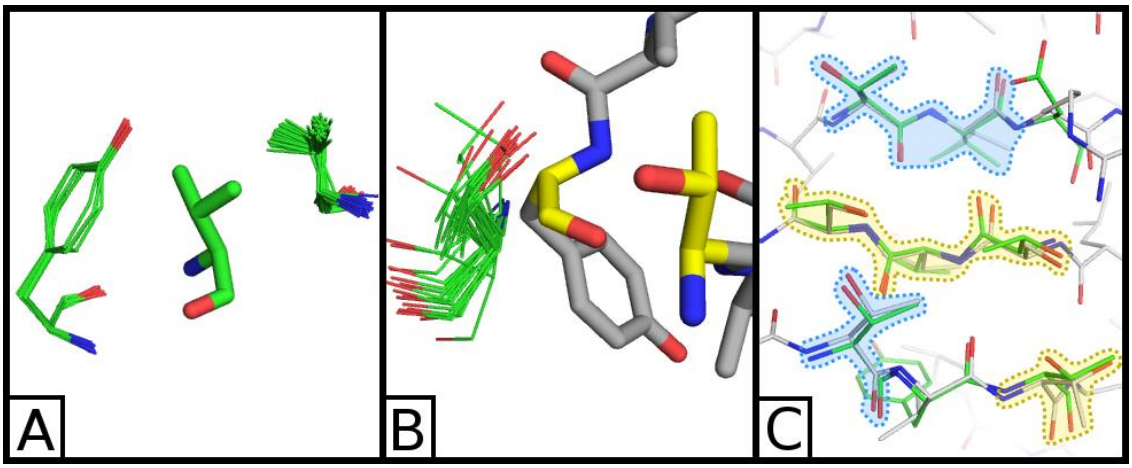


Figure 4 – Building a tertiary interaction.

(A) Three strands are seeded by searching on a valine, which reveals two nearby clusters of valine and tyrosine. (B) Strands are extended one residue in each direction by searching for pairs of residues (colored yellow), yielding clusters of potential inserts (colored green). (C) The final backbone fragment (green) is fed to MadCaT, which identifies multiple compatible scaffolds. One such scaffold (PDB ID=1E54, colored light grey) possesses many exact residue/rotamer matches to the assembled fragment (blue highlights) and many close matches (yellow highlights) that differ by a related residue (threonine to serine or valine to isoleucine) or by varying the rotamer.

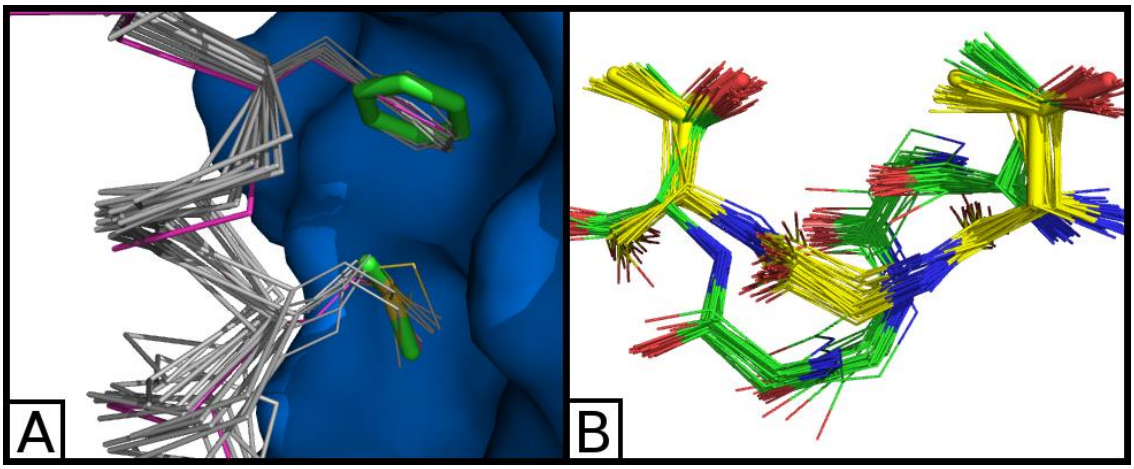


Figure 5 - Finding backbones compatible with hot spot residues.

(A) A Suns search at 0.7 Å RMSD cutoff for two hotspot residues previously identified by RosettaDock [17] for a hemagglutinin binder [16]. The majority of search results are helices that closely match the final designed protein. The search query is shown in thick green sticks, the search result matches are shown as grey α-carbon traces, and the designed hemagglutinin binder is shown as a purple α-carbon trace against a blue hemagglutinin surface. (B) Searching for two threonine side chains at 0.6 Å RMSD cutoff reveals two backbone clusters that can connect them, one corresponding to an α helix (green) and the other corresponding to a β sheet (yellow). The original search query is shown in thick yellow sticks.

6.7 Supplementary Figures

Motif Name	Residue and Atom Names
Alanine	Ala(Cα,Cβ)
Arginine Linker	Arg(Cα,Cβ,Cγ,Cδ)
Asparagine Linker	Asn(Cα,Cβ,Cγ)
Aspartate Linker	Asp(Cα,Cβ,Cγ)
Carboxamide	Asn(Cγ,Oδ,Nδ), Gln(Cδ,Oε,Nε)
Carboxyl	Asp(Cγ,Oδ1,Oδ2), Glu(Cδ,Oε1,Oε2)
Cysteine	Cys(Cα,Cβ,Sγ)
Glutamine Linker	Gln(Cα,Cβ,Cγ,Cδ)
Glutamate Linker	Glu(Cα,Cβ,Cγ,Cδ)
Guanidinium	Arg(Cδ,Nε,Cζ,Nη1,Nη2)
Histidine Linker	His(Cα,Cβ,Cγ)
Hydroxyl	Ser(Cβ,Oγ), Thr(Cβ,Oγ), Tyr(Cζ,Oη)
Imidazole	His(Cγ,Cδ,Nδ,Cε,Nε)
Indole	Trp(Cγ,Cδ1,Cδ2,Cε1,Cε2,Nε,Cζ1,Cζ2,Cη)
Isoleucine	Ile(Cα,Cβ,Cγ1,Cγ2,δ)
Lysine End	Lys(Cδ,Cε,Nζ)

Lysine Linker	Lys(Cα,Cβ,Cγ,Cδ)
Methionine End	Met(Cγ,Sδ,Cε)
Methionine Linker	Met(Cα,Cβ,Cγ)
Peptide Bond	All Residues(Cα,C,N,O)
Phenylalanine Linker	Phe(Cα,Cβ,Cγ)
Phenyl	Phe(Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ), Tyr(Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ)
Proline Ring	Pro(Cβ,Cγ,Cδ)
Serine Linker	Ser(Cα,Cβ)
Threonine Linker	Thr(Cα,Cβ,Cγ)
Tryptophan Linker	Trp(Cα,Cβ,Cγ)
Tyrosine Linker	Tyr(Cα,Cβ,Cγ)
Valine	Val(Cα,Cβ,Cγ1,Cγ2)

Supplementary Table 1 - Default Motif Set.

Default motifs indexed by the public server hosted at suns.degradolab.org. (Motif Name): The common name for the motif. (Residue and Atom Names): The atom names used to define the motif. Some motifs may match multiple residue types, in which case all matching residues are listed with their corresponding atom names.

Figure	Selection / {Search}	Structure	Result ID	Chain	Residue	Atoms	RMSD Cutoff (Å)
2	1	2GBP	N/A	A	Arg4	Cδ,Nε,Cζ,Nη1, Nη2	
	{1}						0.2
	2	3A6R	1	A	Asp61	Cγ,Oδ1,Oδ2	
	{1,2}						0.2
	3	3P02	0	A	Arg325	Cα,Cβ,Cγ,Cδ	
	{1,2,3}						0.3
3A	4	2GBP	N/A	A	Val88	Entire Residue	
	{4}						0.1
	5	4ASM	0	B	Val353	Entire Residue	
	6	2WUR	0	A	Tyr92	Entire Residue	
3B	{4bb,6bb}						0.2
	7	2JCQ	1	A	Thr151	Entire Residue	
	{4bb,7}						0.2
	8	2JCQ	0	A	Thr149	Entire Residue	

	{7sc,8bb}						0.5
	9	3B34	0	A	Thr37	Entire Residue	
	{5bb,8sc}						0.5
	10	3SUU	0	A	Asp102	Entire Residue	
	{6bb,7sc}						0.5
	11	3D9A	0	H	Thr482	Entire Residue	
	{6bb,8sc}						0.5
	12	3Q1I	0	A	Thr561	Entire Residue	
4A	13	†	N/A	B	Met503	Cγ,Sδ,Cε	
	14	†	N/A	B	Phe504	Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ	
	{13,14}						0.7
4B	{7sc,8sc}						0.6

Supplementary Table 2 - Search Parameters for all figures.

(Figure): The figure and sub-figure the selections and searches correspond to. (Selection / {Search}): No braces indicates a saved selection referenced by searches. Braces indicate a search based in terms of previous selections of the form {sel1, sel2, …}. “sc” indicates only the side-chain was taken from the previously saved selection and “bb” indicates only the backbone

atoms were used. (Structure): The PDB ID the selection originated from. (Result ID): The search result serial ID number to disambiguate selections where there are multiple results from the same PDB ID. (Chain): Chain the selection originated from. (Residue): Residue selected. (Atoms): Selected atoms. (RMSD Cutoff): Root-mean-squared deviation cutoff used for a given search. With the exception of initial selections for each figure, all selections are derived from results returned from the preceding search query in the table. †: Structure provided by the David Baker laboratory for their hot spot motif for the hemagglutinin binder [16].

6.8 References

1. (2013) RCSB PDB - Content Growth Report.

2. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233: 123-138.

3. Holm L, Sander C (1994) The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 22: 3600-3609.

4. Prasad BV, Hardy ME, Dokland T, Bella J, Rossmann MG, et al. (1999) X-ray crystallographic structure of the Norwalk virus capsid. Science 286: 287-290.

5. Doolittle JM, Gomez SM (2011) Mapping protein interactions between Dengue virus and its human and insect hosts. PLoS Negl Trop Dis 5: e954.

6. Roy S, Aravind P, Madhurantakam C, Ghosh AK, Sankaranarayanan R, et al. (2009) Crystal structure of a fungal protease inhibitor from Antheraea mylitta. J Struct Biol 166: 79-87.

7. Zhang J, Grigoryan G (2013) Mining tertiary structural motifs for assessment of designability. Methods Enzymol 523: 21-40.

8. Grigoryan G, Kim YH, Acharya R, Axelrod K, Jain RM, et al. (2011) Computational design of virus-like protein assemblies on carbon nanotube surfaces. Science 332: 1071-1076.

9. Shyu CR, Chi PH, Scott G, Xu D (2004) ProteinDBS: a real-time retrieval system for protein structure comparison. Nucleic Acids Res 32: W572-575.

10. Shirvanyants D, Alexandrova AN, Dokholyan NV (2011) Rigid substructure search. Bioinformatics 27: 1327-1329.

11. (2010) The PyMOL Molecular Graphics System. 1.6 ed.

12. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems 30: 107-117.

13. Chen WW, Shakhnovich EI (2005) Lessons from the design of a novel atomic potential for protein folding. Protein science 14: 1741-1752.

14. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography 32: 922-923.

15. Wang G, Dunbrack RL, Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1589-1591.

16. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, et al. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332: 816-821.

17. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, et al. (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of molecular biology 331: 281-299.

18. Muñoz V, Serrano L. (1994) Elucidating the folding parameters of helical peptides using empirical parameters. Nature Structural Biology 1: 399-409.

19. Copley RR, Barton GJ. (1994) A structural analysis of phosphate and sulphate binding sites in proteins. Estimation of propensities for binding and conservation of phosphate binding sites. Journal of Molecular Biology 242: 321-329

Chapter 7

Conclusions and Discussion

7.1 Conclusions and discussion

In this work I have investigated topics associated with the economy of protein structural datasets. Whether datasets are small or large, this means intelligently formulating problems so that the utility of the data is maximized. In chapter 2 we showed how minimal data from transmembrane oligomers could be used to accurately predict their structure. Although the number of proteins solved to atomic accuracy continues to increase at an almost geometrical pace, transmembrane protein structure determination still lags significantly compared to soluble proteins, and techniques such as our which make use of a small set of experimental data will continue to be of use. Recently another group reported impressive results for predicting transmembrane structure using simplified models of carbon hydrogen bonds, although their procedure is limited to proteins with a GxxxG motif (1). It would be interesting to pair their energetic model with our minimal experimental data technique to see if we can expand their technique to include proteins with alternative packing arrangements and improve upon the accuracy of our independent predictions.

In chapter 3 I showed how our knowledge of the anion-binding protein-motif called the "nest" could be leveraged to create an antibody specific for phosphorylated peptides. This work was done prior to the development of our all-atom structural search program, Suns, described in chapter 6. It is hoped that Suns will be an important tool for protein designers to discover structural motifs that can then be used in a similarly modular fashion to create de novo proteins with novel binding or enzymatic properties. To examine whether Suns could have allowed us to discover the "nest" motif without its prior knowledge, I used Suns to look for all phosphorylated

residues in our non-redundant database. In our current default database based upon the PISCES list of proteins with less than 20% identity (2), there is only a single instance of a phosphorylated residue. This suggests that for rare structural motifs or modified amino acids a larger database with a less strict sequence similarity cutoff might be a better choice. I therefore repeated the search for phosphorylated residues using the significantly larger database based upon the PISCES list with a 90% identity cutoff (approximately 23,000 entries vs. 2100 entries) and found almost 20 matches within 0.7 Å RMSD to our query phosphate group. Interestingly, these matches showed the phosphate group largely exposed to solvent rather than involved with binding interactions within the protein, as seen in **Figure 1**. However, the PISCES lists of non-redundant proteins are composed of individual chains, precluding the possibility of capturing details about protein/protein interactions. Because post-translational modifications like phosphorylation are often used for signaling involving protein binding, it may be important to create additional databases which include multiple chains so that details specific to protein/protein interactions can be studied.

In chapter 4 I developed a novel design methodology based upon the concept of "designability" to create protein binders to chosen epitopes. While much of this work was also completed prior to the development of Suns, I believe that Suns will prove to be a powerful tool in discovering designable protein scaffolds for use in designing binders. As shown in chapter 6, Suns was able to identify a designable helix which accommodated the hot-spot residues for one of the successful hemagglutinin binders developed by Fleishman *et al.* and this helix overlays incredibly closely to the helix found in the successful binder. Intriguingly, Suns did not find strongly designable motifs for many of the other hot-spot combinations proposed by Fleishman *et al.*, possibly indicating why designs based upon these hot-spots failed to show binding to hemagglutinin. This suggests an alternative mode of use for Suns – as a sanity-check for design

proposals. Suns could be adapted to accept as input a proposed design, and then automatically break the input structure into smaller sub-structures, identifying which motifs are frequently observed in natural proteins and which portions are rare and therefore should be modified. With such a mode in place, Suns would have identified the successful binder as promising due to its use of the designable helix as an interaction motif, while many of the designs using alternative hot-spots would have been flagged as problematic.

The Super Codons tool developed in chapter 5 will also be a powerful tool for use in conjunction with the design methodology described in chapter 4. Despite continued improvements in energy functions used in computational protein design (3), for the foreseeable future it is likely that the majority of computationally designed proteins will exhibit modest binding affinity or enzymatic activity. Thus, a final gene library step will be necessary to isolate proteins with improved characteristics. As discussed at the end of chapter 4, our lab is currently in the process of creating constructs which place our designed helical binders onto protein scaffolds. Because the amino acid distributions supplied by Super Codons focus libraries on sequences close to the computationally designed sequence, we are considering moving straight to libraries for these protein designs, bypassing the screening for binding using solely the initial sequence. Using Super Codons, a phage library of 10^9 members formed by diversifying 10 residues would still be expected to contain around 6 million phage particles with the initial protein sequence. Consequently, our initial sequence will be well represented in our library, and our selection procedures should isolate binders with affinity greater than or equal to the affinity of that initial design.

Finally, in chapter 6 I described the development of Suns – the first real-time, all-atom, protein structural search algorithm. As shown, this tool should offer powerful insights into

nature's preferred design motifs. However, presently Suns is only indexed for elements found in natural amino acids. This leaves out the ability to search for other important protein interactions found in the PDB, including interactions with post-translationally modified amino acids, DNA and RNA, and chemical ligands. However, Suns was designed to be easily modified. I was able to add the ability to search for phosphate groups to Suns in a matter of minutes, not hours, for example. Modifications for indexing interactions with chemical ligands will be a bit more challenging, mostly due to the process of determining what small chemical group we should use to form our index "words," but even this should be possible in a matter of weeks. By adding the capability to search for these additional interaction types, Suns will be even more useful for tasks such as drug-discovery and design.

7.2 Figures

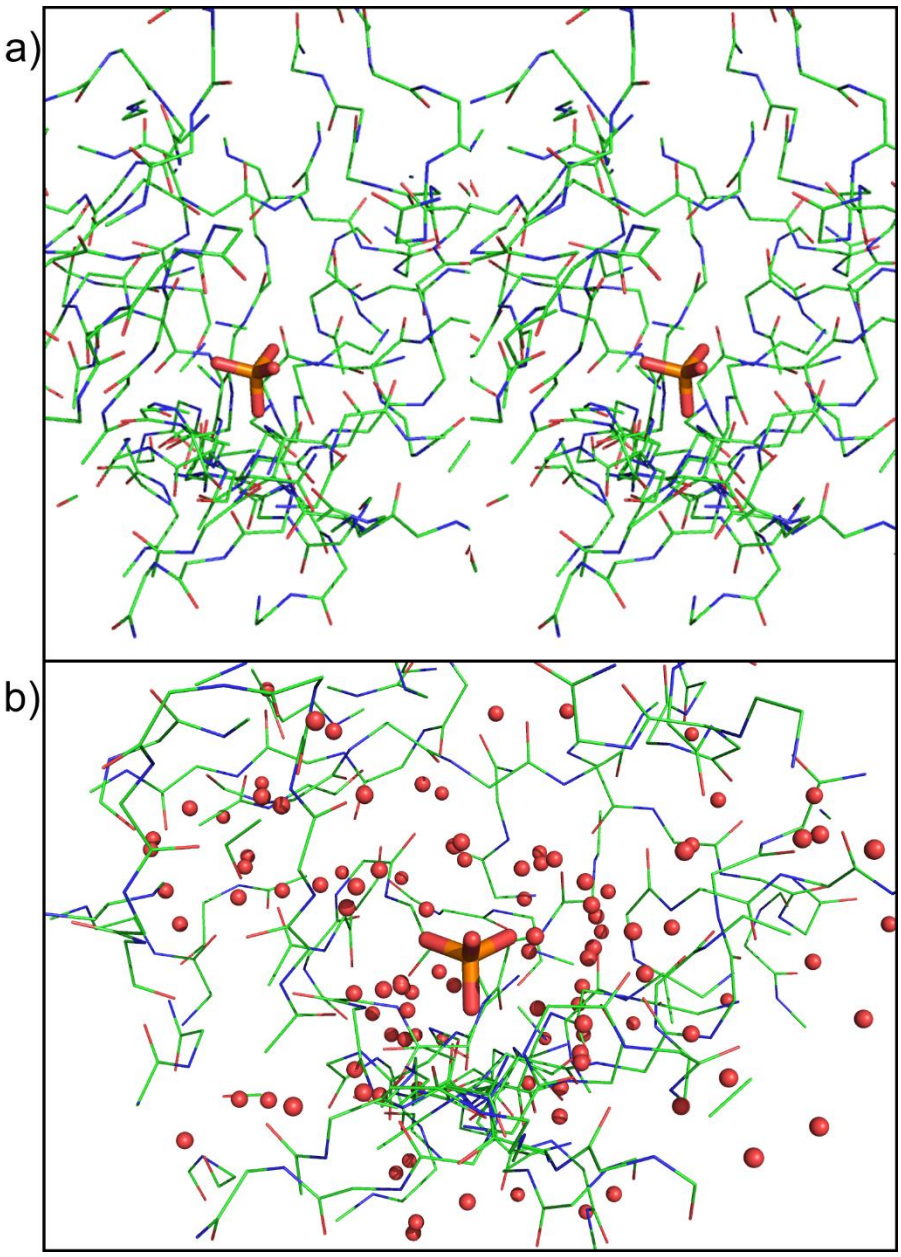


Figure 1

Suns search result for phosphate query. The phosphate group is primarily exposed to solvent, rather than being involved in protein interactions. **a)** Stereo image showing large cavity surrounding aligned phosphate groups. **b)** Image showing the same phosphate groups surrounded by water molecules depicted as red spheres.

7.3 References

1. *GAS_{right} is optimized for Cα hydrogen bonding: analysis and structural prediction of a frequent transmembrane motif*. **Mueller, B.K., Subramaniam, S., Senes, A.** 2013 Submitted

2. *PISCES: a protein sequence culling server.* **Wang, G., Dunbrack, R.L.**, 19: Bioninformatics, 2003, pp 1589-1591

3. *Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects.* **Li, Z., Yang, Y., Zhan, J., Dai, L., Zhou, Y.**, 42: Annual Review of Biophysics, May, 2013, pp 314-335