ENGINEERING CROSS-REACTIVITY IN THE ANTIBODY RESPONSE TO HIV AND INFLUENZA

By

Alexander Mario Sevy

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMICAL AND PHYSICAL BIOLOGY

September 30, 2018

Nashville, Tennessee

Approved:

Spyros Kalams, M.D. (Chair)

Melanie Ohi, Ph.D.

Terry Lybrand, Ph.D.

James Thomas, M.D.

Jens Meiler, Ph.D. (Co-advisor)

James E. Crowe, Jr., M.D. (Co-advisor)

Acknowledgments

This thesis would not have been possible without the support of many people along the way. First, a thanks to my mentors, James Crowe and Jens Meiler. Working under two PIs is not always easy, but Jim and Jens manage to make it work since they bring complementary skills to the table. I have always appreciated Jens for his scientific rigor and flexibility in expanding his research, and Jim for his scientific leadership and big-picture vision for the future of the field. I have learned a lot from both of your leadership styles and I wouldn't be the scientist that I am today without your guidance.

Second, I want to thank everyone in both the Crowe and Meiler labs for their friendship and support. In the Crowe lab, I want to specifically thank Gopal Sapparapu, Erica Parrish, Iuliia and Pavlo Gilchuk, Ross Troseth, Jinhui Dong, Cinque Soto, and many others. You have helped me work through so many problems in the lab, both technical and non-technical. In the Meiler lab, thanks to Amanda Duran, Rocco Moretti, Darwin Fu, and Steven Combs. Also thanks to Nicholas Wu at Scripps for his collaboration and help in many of these projects. Of course, between the two labs there is a small group of people who went through all the highs and lows of trying to juggle two mentors along with me. Thanks to everyone in the Crowe-Meiler interface - Jessica Finn, Marion Sauer, David Nannemann, Jordan Willis, Amandeep Sangha, and Nina Bozhanova. Also I want to acknowledge all of my friends from outside the lab who helped make grad school life bearable. I was lucky to have a good group of friends who were always willing to help distract me from whatever was going on in the lab. Without those happy hours as distraction I might have gone crazy.

Thanks to my family for their support, not only during grad school but during my entire life. My parents have always encouraged me to pursue higher education, and they were both great

role models for me growing up. Thanks to Mom and Rod, Dad and Jorita, Julia, Shane, Blake, and Troy for your support through my whole life. Even if you don't exactly understand what I'm working on, I know that all of you are always there for me. Last, but definitely not least, the biggest thanks go to my fiancée Mildred. You were always there by my side, through all the highs and lows, always there to offer some encouragement or advice or just to listen. You are the biggest reason why I have made it to this point.

Summary

This thesis describes my work in using computational protein modeling to design broadly reactive antibodies. Antibodies are a key component of the human immune response to infectious disease. By studying human antibodies against pathogens such as HIV and influenza, we have been able to learn a great deal about the mechanisms by which antibodies protect us from these viruses. The best antibody response is one that is potent and broad, covering a large number of the many diverse viral variants. Unfortunately, natural human antibodies are rarely perfect in that they don't cover the entire spectrum of possible viral variants.

To address this shortcoming, I used computational design within the ROSETTA software to re-engineer and improve human antibodies to improve their coverage of large viral panels, also known as their breadth. Prior to this work, the computational methods for designing an antibody against a large viral panel were very limited, due to the computational intensity of such simulations. Previously existing methods were limited to only redesigning a small portion of the antibody and only modeling a limited number of viral proteins. In my thesis work I developed two new methods for increasing the scale of computational design against many viral proteins, a technique known as multistate design, and applied one of these methods to an anti-influenza system to define the molecular limits of breadth and affinity.

In Chapter I, I introduce the topics that will be discussed at length in this thesis. I briefly review the structure and function of antibody molecules, as well as mechanisms of generating antibody diversity and antibody-antigen recognition. I also introduce the two major pathogens that were used as target systems for my thesis work, influenza and HIV. I describe the major characteristics of these two viruses, as well as what is known about the antibody response to both of these pathogens and the antibodies that can achieve broad neutralization. I provide a brief

description of structure-based reverse vaccinology, which is a paradigm that uses knowledge of broadly neutralizing antibodies to build better, more informed vaccines. I then describe the protein modeling techniques that are related to this work, either directly or tangentially, and the specific use of protein modeling in antibody design. Finally, I summarize the significance and innovation of the work described in this thesis, and how it ties together all of these seemingly unrelated fields.

Chapter II is the first research chapter, which is largely a reproduction of (Sevy et al., 2015). As previously stated, the field of protein multistate design was limited when I began my thesis work, in terms of the size of the design problem that could be addressed. I developed a new multistate design method within the ROSETTA software suite, known as the RECON method, that enables more efficient searching through sequence space in a multistate design problem. I benchmarked this method on two test cases – multistate design of promiscuous proteins that naturally bind many targets, measuring recovery of the native sequence; and multistate design of antibodies encoded by the same germline gene, measuring recovery of the germline, polyspecific sequence. I compared the results of the RECON method to an existing method for multistate design in ROSETTA and show that RECON recovers more biologically relevant sequences and does so in a fraction of the computing time.

Chapter III extends the work done in Chapter II by reoptimizing the RECON method to run in parallel on many computing cores, allowing much larger panels of viral proteins to be simulated. I applied the optimized algorithm to designing anti-influenza antibodies against a large viral panel of influenza HA proteins of subtype H1. One antibody in particular, called C05, showed promising computational results and the designed variants were expressed and tested for their binding activity. Variant antibodies showed increased affinity against one member of the panel, with a 5x increase in affinity, and increased breadth to a new member of the panel. This

improvement was achieved without losing affinity or neutralization potency for other antigenic strains recognized by wild-type C05. A crystal structure of a C05 double mutant confirmed that the ROSETTA models were accurately positioning the mutated side chains.

Chapter IV is largely a reproduction of (Sevy et al., 2018). In this chapter I collaborated with another graduate student to developed another method for performing multistate design against large viral panels, called BROAD. The BROAD method uses ROSETTA to create structural models of an antibody against a large viral panel, and trains a support vector machine to learn the ROSETTA score function for quick approximation. We then used linear optimization to find the optimal antibody sequence for both breadth against the panel and for antibody stability. Using this algorithm, we were able to improve the predicted breadth of a target HIV antibody, known as VRC23, from 53% experimentally determined breadth to up to 100% predicted breadth. In addition, we found that BROAD sampled new amino acids that were never sampled using structure-based multistate design. BROAD introduced amino acids into the gp120 binding site that mimicked known broadly neutralizing antibodies even though no such information was provided to the algorithm as input.

In Chapter V I describe my work on engineering cross-reactive HIV and influenza antibodies. Based on structural similarity between human antibodies targeting the influenza receptor-binding site and the HIV membrane-proximal external region, I hypothesized that it would be possible to engineer an antibody that was cross-reactive to both antigens, despite the fact that the antigenic proteins have largely different overall folds. We collected B cells from HIV-infected donors after influenza vaccination to isolate the time point at which naturally occurring cross-reactive antibodies would be boosted, and sequenced the antibody repertoire from five such donors. I then used ROSETTA modeling to predict the likelihood that the sequenced antibodies

would adopt the bound conformation of either an influenza or HIV antibody, and identified many with predicted cross-reactivity between both targets. These putative cross-reactive clones were improved even further for predicted influenza-HIV cross-reactivity by applying the multistate design method from Chapter II and III to optimize the sequence.

Chapter VI uses protein engineering in ROSETTA to design molecules with greater binding breadth, but approaches the problem differently from previous chapters. Rather than re-engineering an antibody sequence, I used computational modeling to design cyclic peptides that recapitulate the activity of an antibody CDRH3 loop. I designed peptides that bind to group 1 and 2 influenza HA based on the CDRH3 loop of antibody C05. In addition, these peptides expand the breadth of binding to two new subtypes, H4 and H7, that were not recognized by the IgG molecule. This represents a new strategy for engineering breadth into antibodies.

Lastly, Chapter VII summarizes all of the major findings from this thesis and places them in the context of the fields of broadly neutralizing antibodies and protein design. I examine shortcomings in the computational protocols and what can be done to improve them. I conclude by discussing implications of the findings from this thesis and future directions that could be taken for these projects.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER I.

Introduction

Adapted from Sevy, A. M. & Meiler, J. Antibodies: Computer-Aided Prediction of Structure and Design of Function. *Microbiol Spectr* **2,** (2014).

Author contributions: I am the first author of the manuscript titled "Antibodies: Computer-Aided Prediction of Structure and Design of Function" in the Microbiology Spectrum journal (Sevy and Meiler, 2014). All figures were either created for use in this thesis or are reprinted with permission from the publisher.

Introduction to antibodies

The adaptive immune response is the mechanism by which humans respond to infection by viruses and bacteria. This facet of the immune system is remarkable in both its speed in responding to a novel pathogen, and its memory to be able to respond for a lifetime to a pathogen seen only once. The immune system is able to mount such an effective response through the activity of B and T cells. In this thesis I will primarily focus on the activity of B cells and the immunoglobulin molecules they produce, known as antibodies. Antibodies are extraordinary molecules, as they recognize their targets with extreme precision through chemical interactions between the antibody and antigen molecules. I will briefly review the structure and function of antibody molecules, as well as mechanisms of antibody diversity and antibody-antigen recognition.

Antibody structure

The fundamental structural unit of the antibody is the immunoglobulin (IG) domain of 70-110 amino acids that adopts the characteristic IG β-sandwich fold. Antibodies are homodimers of heterodimers, where each heterodimer consists of one heavy and one light chain (Figure I.1), each chain having multiple IG domains (Harris et al., 1997). The antibody can be divided into two segments, the constant fragment (Fc) and the variable fragment (Fv). The constant domain is named as such since it is virtually identical between antibodies of the same isotype, whereas the variable domain can vary greatly between antibodies and is responsible for antigen specificity. The mammalian antibody heavy chain consists of four IG domains, the first two domains comprising the Fv and the next two domains comprising the Fc. The mammalian light chain consists of two IG domains, which interact with the two N-terminal IG domains of the heavy chain to form heterodimers. These heterodimers homo-dimerize via the C-terminal IG domains of the heavy chain to form the complete antibody. This domain arrangement ensures that the variable domains of heavy and light chain co-localize in space to form the paratope. Each of these variable domains contains three complementarity-determining regions (CDRs), referred to as the CDRH1-3 on the heavy chain, and CDRL1-3 on the light chain. The six CDRs form the combining site that is responsible for antigen recognition.

Mechanisms of antibody diversity

To respond to the virtually limitless space of antigenic proteins, antibodies must have extreme diversity to create a unique antibody specific to each pathogenic threat. Such diversity is generated by four main mechanisms. The amino acid sequence of the variable region is determined by a process called somatic recombination, where an IG domain is assembled by combining randomly chosen gene segments, known as the Variable (V), Diversity (D), and Joining (J) gene

segments (Tonegawa, 1983). The heavy chain is encoded by recombination of V, D, and J genes, whereas the light chain is encoded only by V and J genes (Figure I.2). This process generates large combinatorial diversity, as there are 43 V, 23 D and 6 J genes encoding the human heavy chain (Matsuda et al., 1998). The human light chain has 33-38 V and 4-5 J genes, depending on subtype, κ or λ (Lefranc et al., 2005; Murphy et al., 2012). The second mechanism of antibody diversity is the pairing of heavy chain with light chain.

The third mechanism is a process which occurs during V(D)J recombination known as junctional diversity. During gene recombination the enzymes RAG1 and RAG2 remove nucleotides from single-stranded DNA at the recombination site (Oettinger et al., 1990; Schatz et al., 1989), and the enzyme terminal deoxyribonucleotidyl transferase (TdT) nonspecifically adds back nucleotides during the DNA repair process (Desiderio et al., 1984; Isobe et al., 1985). Through the activity of these enzymes the final DNA sequence of the recombined gene is altered from that of the germline sequence, creating new levels of diversity during the recombination process itself. In addition the reading frame of the D gene can be changed during recombination, leading to new amino acid combinations in the CDRH3 (Benichou et al., 2013). Since junctional diversity occurs in the third CDR loop of each chain, this loop is the most diverse of the CDR loops, and consequently is frequently involved in antigen recognition.

The germline-encoded antibodies produced by these three diversification mechanisms are further modified in the fourth mechanism of diversity, known as somatic hypermutation or affinity maturation. After B cells are activated by recognition of an antigen, the enzyme activation-induced cytidine deaminase (AID) is expressed at high levels (Muramatsu et al., 2000). Expression of this enzyme leads to extremely high mutation rates during B cell division and replication, up to 1 mutation per 1,000 base pairs per division (Rajewsky et al., 1987), with the mutations focused in

the region encoding the CDR loops (Jolly et al., 1996). B cell mutants with higher-affinity binding to their antigens then proliferate and undergo further maturation to produce antibodies with extremely high affinity for the antigen (Victora and Nussenzweig, 2012).

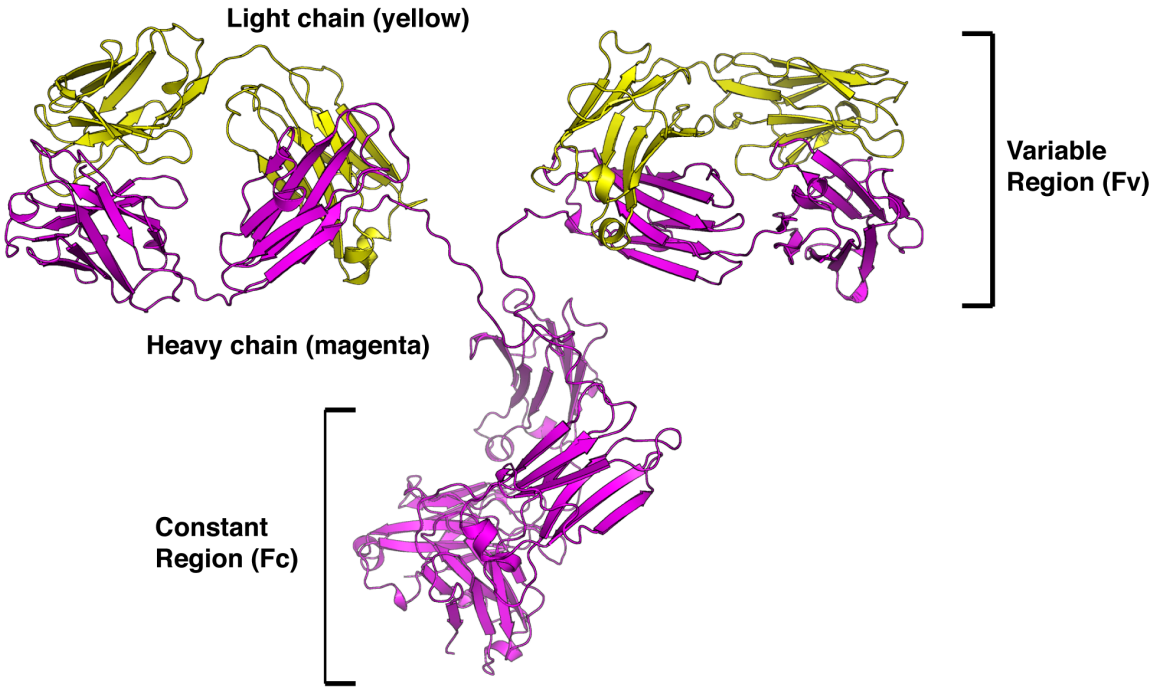


Figure I.1. Structure of an antibody molecule. All human antibodies consist of a heavy chain (magenta) paired with a light chain (yellow), which dimerize to form a full immunoglobulin molecule. Domains can be separated into variable (Fv) and constant (Fc) regions. The structure shown above is from PDB ID 1IGT. Figure is adapted from (Sevy and Meiler, 2014).

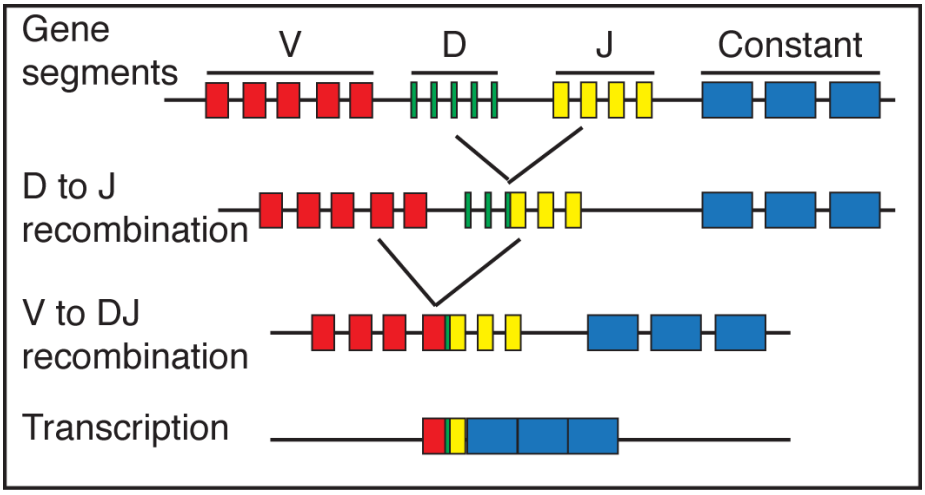


Figure I.2. Somatic recombination of antibody gene segments generates diversity in the variable region. Antibody heavy chains are encoded by recombination of Variable (V), Diversity (D), and Joining (J) gene segments to generate large combinatorial diversity. Light chains are encoded by V and J gene segments. Figure is adapted from (Sevy and Meiler, 2014).

Mechanisms of antibody-antigen recognition

Although the theoretical number of potential antibody structures is large, it is finite, and must contend with an infinite set of antigens, presented by pathogens that can undergo rapid cycles of antigenic shift. A key question in immunology is the mechanism by which a limited set of antibodies can respond to an unlimited set of antigens. One mechanism is the ability of many antibodies to recognize multiple distinct targets, known as multi-specificity. Many antibodies have been studied that are able to bind multiple, often structurally distinct targets, including small molecules (Sethi et al., 2006; Tapryal et al., 2013; Yin et al., 2003), peptides (Kramer et al., 1997), and proteins (Bostrom et al., 2009; Fagète et al., 2012; Garcia-Rodriguez et al., 2006). Multi-specificity can be achieved by various molecular mechanisms. In some cases, multi-specificity can be imparted by structural flexibility, wherein an antibody can adopt a number of distinct conformations, each functioning to recognize a certain target (Foote and Milstein, 1994; James et

al., 2003). This is analogous to the conformational selection and induced fit models of protein-protein recognition, where interaction of an antibody with an antigen requires a conformational change from the unbound to bound state. According to this paradigm it is thought that germline antibodies are highly flexible and multi-specific, and affinity maturation reduces the flexibility of CDR loops while preconfiguring the combining site for specific antigen recognition (Babor and Kortemme, 2009; Schmidt et al., 2013; Sethi et al., 2006; Willis et al., 2013; Xu et al., 2015).

However, in other cases the germline flexibility model is insufficient to explain multi-specificity. In one case it was shown that an antibody can recognize two unrelated targets by differential positioning within a rigid paratope (Sethi et al., 2006). Another *in silico* study supported the idea that rigidification of CDR loops is not a driving mechanism of antibody maturation on a repertoire-wide scale (Jeliazkov et al., 2018). While the germline flexibility model states that germline antibodies are multi-specific and lose their reactivity after affinity maturation, there are several examples of the opposite phenomenon, where the germline antibody is mono-specific and affinity maturation increases multi-specificity and imparts binding to a new target (Corti et al., 2011; Fu et al., 2016; Mouquet et al., 2010). The literature suggests that antibody multi-specificity is a complex phenomenon that can be achieved through many different mechanisms.

Introduction to influenza

Influenza virus is a yearly threat to global public health. Global pandemics caused by influenza have been among the deadliest events in human history, including the Spanish Flu pandemic of 1918 that caused 50 million deaths, ~3% of the world's population at the time (Taubenberger and Morens, 2006). Even today the seasonal circulation of influenza causes as

many as 56,000 deaths and 710,000 hospitalizations annually (Rolfes et al., 2016). The influenza virus is a member of the *Orthomyxoviridae* family, and is an enveloped, double stranded RNA virus with a segmented genome (Acheson, 2011). The influenza A genome contains eight segments, expressing a total of 11 proteins. This thesis will primarily focus on the hemagglutinin protein (HA), which is the viral spike protein responsible for host cell recognition and fusion, since it is the primary target of antibody response.

Hemagglutinin glycoprotein

HA is a trimeric glycoprotein on the surface of the viral capsid. It is initially synthesized as an HA0 precursor, which is proteolytically cleaved into HA1 and HA2 subunits by trypsin and other proteases (Skehel and Wiley, 2000). The full HA spike is a trimer of heterodimers of the HA1 and HA2 subunits, which are covalently linked by disulfide bonds. The HA1 subunit, also referred to as the globular head domain, is composed of hypervariable loops that are a major target of the antibody response (Sahini et al., 2010). This domain also contains the receptor-binding site, where HA recognizes its host cell receptor, sialic acid. The HA2 subunit, also referred to as the stem domain, is composed of a long α helix and contains the hydrophobic fusion peptide. This subunit is highly conserved due to its involvement in membrane fusion. To initiate membrane fusion HA first binds sialic acid on the cell surface, which causes internalization of the virion into an endosome. As the pH of the endosome drops to 5-6, the HA2 subunit undergoes a conformational change exposing the fusion peptide, which inserts into the membrane and induces fusion (Russell, 2014).

Taxonomy

Influenza viruses can be classified into four types, known as type A – D, although only type A – C infect humans (Ducatez et al., 2015; Hampson and Mackenzie, 2006). Influenza A

viruses are responsible for the majority of epidemics and pandemics, and therefore will be the primary focus of this thesis. Influenza A can be further broken down into subtypes based on the HA and neuraminidase (NA) proteins. HA has 18 subtypes (H1 – H18) and NA has 11 subtypes (N1 – N11) (Petrova and Russell, 2018). The combination of HA and NA proteins in a given virus is what determines the common nomenclature of influenza viruses (*i.e.* H1N1, H3N2, etc.). Since the HA protein is the primary focus of this thesis I will discuss only these subtypes in detail. The 18 HA subtypes can be classified into two groups, group 1 and 2 (Figure I.4). Only three HA subtypes are currently circulating among humans – H1, H3, and influenza B. Influenza type B is less divergent than type A and as such is not divided into subtypes, but rather into two lineages known as B/Yamagata and B/Victoria (Petrova and Russell, 2018). Although there are currently two influenza A subtypes circulating, there are others that have circulated in humans in the past and have the potential to re-emerge, such as H2 (Figure I.3). In addition there are zoonotic subtypes that are able to infect humans and periodically emerge and cause epidemics, such as H5, H7, and H9 (Kumar et al., 2018).

Antigenic drift and shift

Influenza viruses have such high diversity due to two major mechanisms of mutation. The first is known as antigenic drift – this occurs when the viral polymerase makes errors in copying the viral genome and introduces point mutations in the viral genes. The HA gene is estimated to accumulate roughly 5 nucleotide mutations per year (Klein et al., 2014). This results in a slow accumulation of mutations and can cause epidemics if the mutated amino acids are sufficient to evade antibody recognition. The second mechanism is known as antigenic shift, and occurs when gene segments from two or more viruses co-infecting the same host recombine to form a new virus

(Kumar et al., 2018). This mechanism is responsible for most of the global pandemics, as the new virus is much different from anything that has previously circulated.

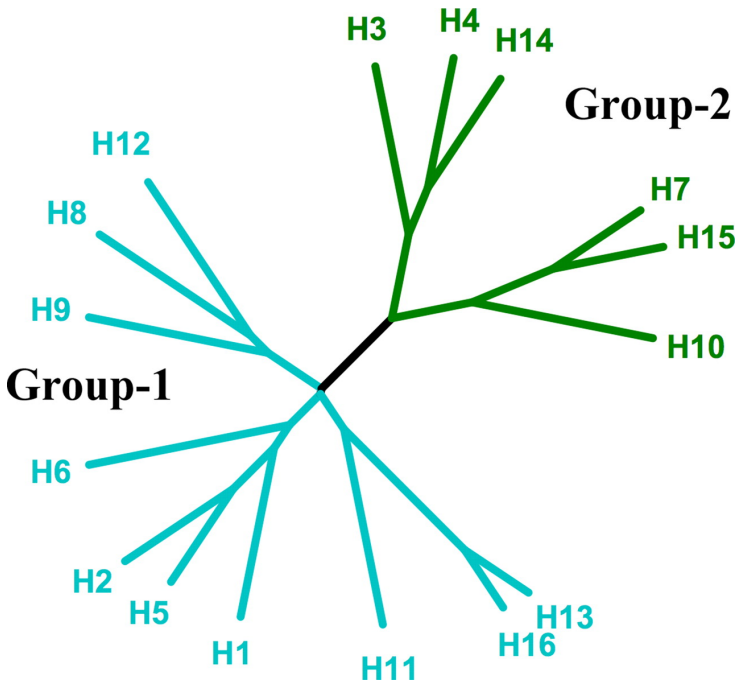


Figure I.4. Phylogenetic tree showing the taxonomy of influenza A HA subtypes H1 – H16. Shown in color are the two groups, group 1 (cyan) and group 2 (green). Figure adapted from (Russell et al., 2008).

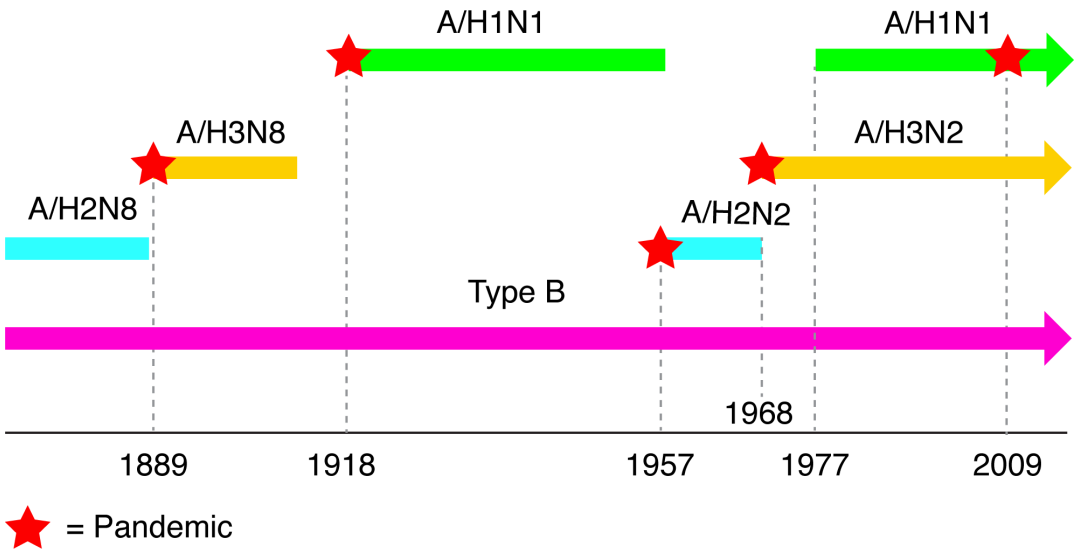


Figure I.3. Annual circulation of influenza viruses in humans. Circulating strains of influenza types A and B are shown. Stars show years with pandemic viruses. Figure adapted from (Hampson and Mackenzie, 2006).

Broadly neutralizing antibodies to influenza

For a universal influenza vaccine to be successful, it must elicit broadly neutralizing antibodies (bnAbs) that cross-react between different seasonal variants. However, due to the antigenic variability of the HA glycoprotein, universal bnAbs against influenza HA have been elusive. Early studies on anti-HA antibodies focused on antibodies targeting the head domain, which is the major immunogenic component of the HA molecule (Caton et al., 1982; Wiley et al., 1981). These studies identified four major antigenic sites on the globular head domain, mainly consisting of protruding loops which are subject to continual antigenic mutation. Early anti-influenza head domain antibodies were primarily strain-specific (Nakajima et al., 1983; Underwood, 1982). However, recent work has identified the conserved receptor-binding site of the globular head of HA as a broadly neutralizing epitope. Several receptor-binding site antibodies have been identified that bind and neutralize divergent strains both within a subtype (Hong et al., 2013; Lee et al., 2014; Xu et al., 2013, Whittle et al., 2011) and across subtypes (Ekiert et al., 2012; McCarthy et al., 2018). These antibodies achieve broad reactivity by mimicking the host cell receptor sialic acid in their recognition of the receptor-binding site. This molecular mimicry is achieved either by placing an aspartic acid residue in the position of the carboxylate group of sialic acid, or by placing a hydrophobic residue in the position of the acetamide moiety of sialic acid (Lee et al., 2014). The discovery of bnAbs targeting the receptor-binding site was a promising sign for the potential of a universal influenza vaccine, as this domain is highly immunogenic and is a promising vaccine candidate.

In addition to the receptor-binding site, there have been many bnAbs that bind to the highly conserved stem region of HA. The first anti-influenza bnAb was isolated from mice in the early 1990s (Okuno et al., 1993), and was shown to bind to the stem region and prevent viral fusion and

entry (Dreyfus et al., 2013). Stem-binding antibodies have been isolated that cross-react across group 1 and 2 viruses (Corti et al., 2011; Ekiert et al., 2009; 2011; Sui et al., 2009) and even across types to influenza B (Dreyfus et al., 2012). In several cases it has been shown that this class of antibodies begin as relatively strain-specific in their germline-encoded sequence, and are only able to expand breadth to new subtypes upon exposure and subsequent rounds of affinity maturation (Corti et al., 2011; 2013; Liao et al., 2013; Pappas et al., 2014). In addition, in studies of

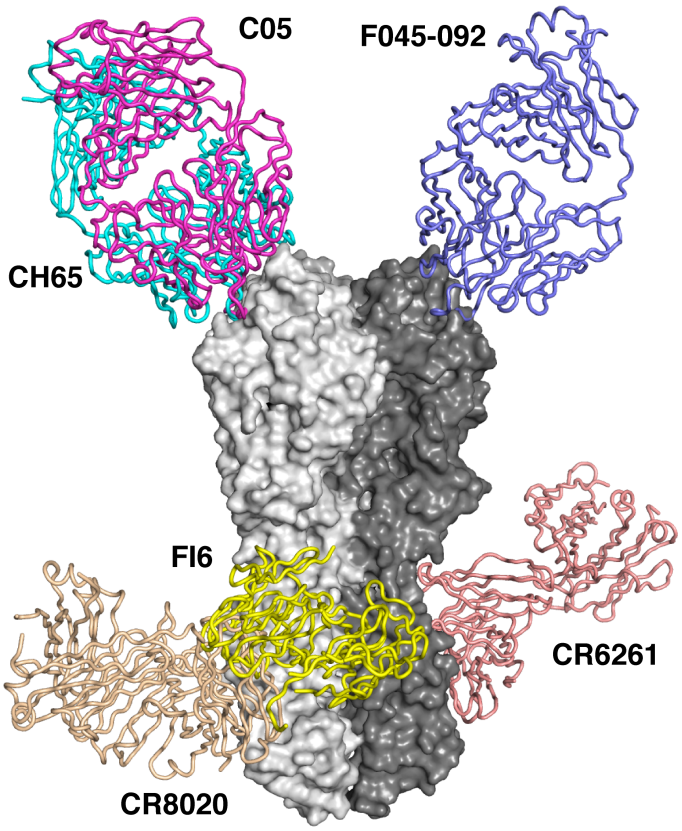


Figure I.5. Binding epitopes of broadly neutralizing antibodies to influenza. HA protomers are shown in shades of gray. Shown above are antibodies CH65 (cyan, PDB ID 5ugy), C05 (magenta, 4fp8), and F045-092 (blue, 4o58), targeting the receptor-binding site. Shown below are antibodies FI6 (yellow, 3ztn), CR8020 (tan, 3sdy), and CR6261 (salmon, 3gbm), targeting the stem domain.

vaccination or infection with a newly emerging virus, it has been shown that the anti-stem response predominates, as they are the only class of pre-existing antibodies with any recognition of the novel virus (Ellebedy et al., 2014; Wrammert et al., 2011). Characteristics of anti-influenza bnAbs discussed in this section are summarized in Figure I.5.

Introduction to HIV

The HIV/AIDS crisis has been an unprecedented global health threat since the first observation of AIDS in 1981 and discovery of human immunodeficiency virus (HIV) as the causative agent in 1983 (Hemelaar, 2012). It is estimated that since the beginning of the HIV/AIDS epidemic 76.1 million people have been infected with HIV and 35 million have died from AIDS-related illnesses (UNAIDS, 2017). In 2016 it was estimated that 36.7 million people were living with HIV worldwide (UNAIDS, 2017). The pandemic began with at least 7 individual transmissions of simian immunodeficiency virus (SIV) to humans in Africa, most likely through consumption of bush meat (Hahn et al., 2000). The separate transmissions led to a wide variety of HIV lineages, divided most broadly into HIV type 1 (HIV-1) and type 2 (HIV-2) (Hemelaar, 2012). All references to HIV in this thesis are referring to HIV-1, unless otherwise noted, as HIV-2 is less virulent and is not distributed globally (Gilbert et al., 2003; Reeves and Doms, 2002). HIV-1 can be further subdivided into four groups, referred to as M, N, O and P, of which group M is responsible for the majority (>90%) of infections worldwide. Group M viruses can then be divided into subtypes (or clades) A-K (Hemelaar, 2012).

HIV is an enveloped virus with a positive strand single-stranded RNA genome, member of the *Retroviridae* family (Acheson, 2011). It infects human immune cells by recognition of the CD4 receptor and either the CCR5 or CXCR4 co-receptor. As a retrovirus it uses the enzyme reverse

transcriptase to create a double stranded DNA copy of its RNA genome, which is later integrated into the host genome. In addition to reverse transcriptase, HIV expresses a variety of structural and nonstructural proteins that play key roles in the viral replication cycle. Among these are integrase, which catalyzes insertion of the viral genome into the host genome; protease, which cleaves pro-protein products; and a variety of accessory proteins that interact with host restriction factors (Vif, Vpu, Tat, Rev, and Nef) (Strebel, 2013). The most relevant viral protein to the work discussed in this thesis is the Env glycoprotein. The Env protein is the viral spike protein on the surface of the virion that mediates host cell recognition and entry. As the vast majority of known antibodies are directed against Env, the work in this thesis primarily focuses on this protein.

The Env glycoprotein is a trimeric complex synthesized as a single precursor protein known as gp160, that later undergoes proteolytic cleavage by furin into two subunits known as gp120 and gp41 (Ward and Wilson, 2015). The gp120 subunit is composed of five highly glycosylated variable loops (V1 – V5) that are highly mutated and mask the CD4 binding site core. The gp41 subunit contains the transmembrane domain, as well as the membrane-proximal external region (MPER), two heptad repeat domains (HR1 and 2), and an intra-membrane C-terminal domain (CTD). Gp41 houses the fusion machinery that is responsible for fusing the viral and cellular membranes (Wilen et al., 2012).

Broadly neutralizing antibodies to HIV

The antibody response to HIV is dominated by non-neutralizing antibodies targeting glycan epitopes or hypervariable antigenic loops on the trimeric Env protein (Horwitz et al., 2017). This is especially true of the early antibody response, which is primarily directed towards the gp41 domain of Env (Liao et al., 2011; Tomaras et al., 2008). However, advances in B cell

immortalization and screening technologies have made it possible to isolate bnAbs from human donors. Up to 20% of infected patients developed a bnAb response, which is typically occurs 2-4 years post-infection (Mikell et al., 2011; Sather et al., 2009). The first generation of bnAbs against HIV consisted of antibodies such as b12, which targets the CD4 binding site of gp120 (Saphire et al., 2001); 2G12, which targets the glycan face on gp120 (Calarese et al., 2003); and antibodies 2F5, 4E10, and Z13 targeting the membrane-proximal external region (MPER) of gp41 (Bryson et al., 2009; Conley et al., 1994; Zwick et al., 2001). Discovery of these bnAbs represented an advance in identifying the vulnerable epitopes on the Env protein. The CD4 binding site is highly conserved due to its role in recognition of the host cell receptor CD4 to initiate viral entry. Antibodies targeting this site are highly potent due to direct competition for CD4 binding and can mimic CD4 recognition. The MPER is also highly conserved based on its role in viral fusion, which can be blocked by bnAb binding.

Although discovery of these bnAbs was promising, each had its limitations, either limited breadth of neutralization, limitation to clade B neutralization, or low potency (Binley et al., 2004; Zwick et al., 2001). The second generation of bnAbs included antibodies with much greater breadth (up to 98% of a viral panel) and potency against highly conserved Env epitopes (Huang et al., 2012). The second generation identified more broad and potent bnAbs targeting the previously defined vulnerable epitopes, such as VRC01, NIH45-46, 3BNC117, and 3BNC60 targeting the CD4 binding site, and 10E8 targeting the gp41 MPER (Huang et al., 2012; Scheid et al., 2011; Wu et al., 2010). In addition a new vulnerable epitope, the V1/V2 apex, was discovered through isolation of the PG9 and PG16 bnAbs (Walker et al., 2009), and further defined by isolation of the CH01-04 bnAbs (Bonsignori et al., 2011). The characteristics of broadly neutralizing antibodies against HIV are summarized in Figure I.6.

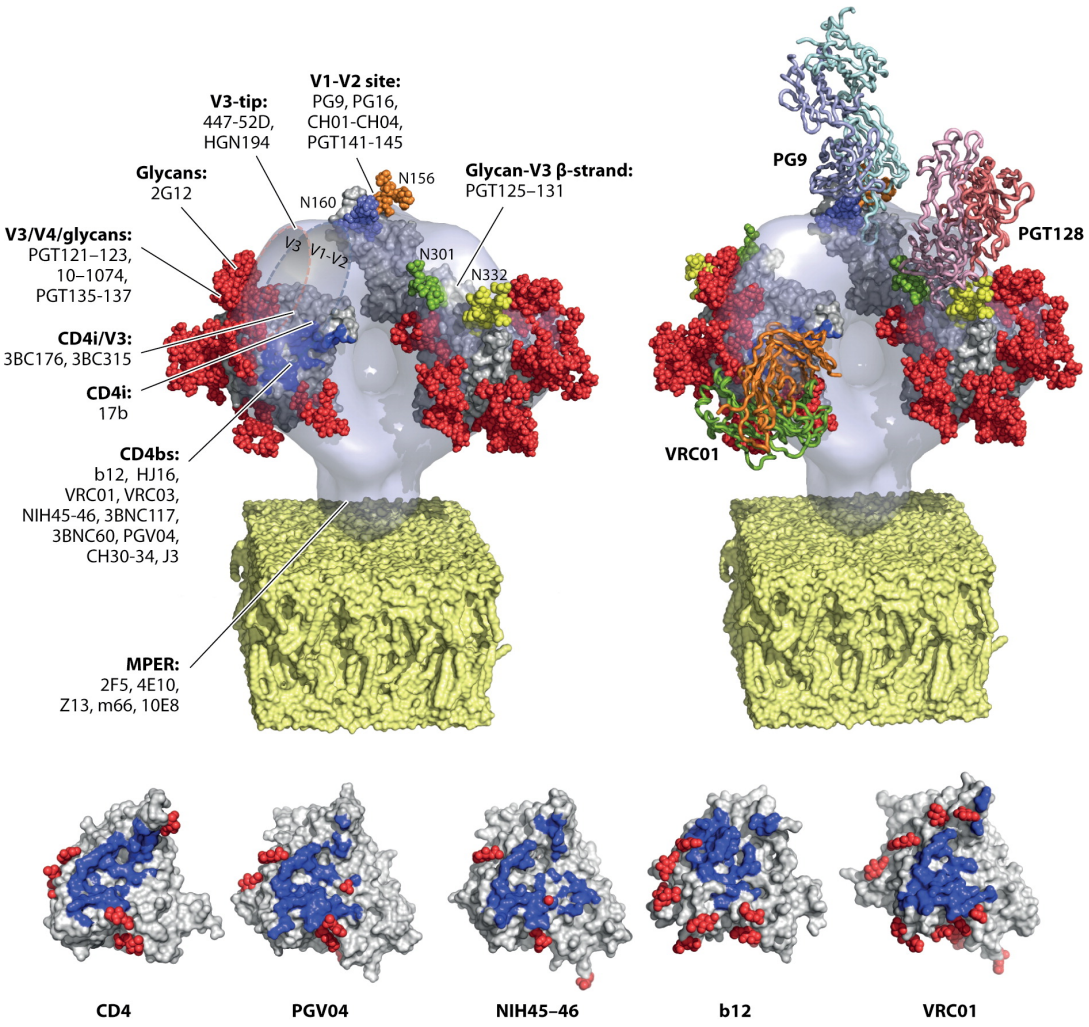


Figure I.6. Epitopes of broadly neutralizing antibodies targeting HIV. Left: architecture of the Env trimer, with glycans shown in red and vulnerable epitopes specified by name. Model is from Electron Microscopy Data Bank code EM-5019. The cellular membrane is shown below in yellow. Right: Binding poses of three bnAbs, PG9 (PDB ID 3u2s), PGT128 (3tyg), and VRC01 (3ngb) overlaid onto a model of the trimer. Bottom: binding footprint of CD4 (1gc1) and CD4 binding site antibodies PGV04 (3se9), NIH45-46 (3u7y), b12 (3dnl), and VRC01 (3ngb) on the gp120 subunit. Glycans are shown in red. Figure adapted from (Corti and Lanzavecchia, 2013).

Structure-based reverse vaccinology

One of the primary applications of highly potent anti-viral bnAbs is the identification of vulnerable epitopes on the antigenic surface that can be exploited to formulate an effective vaccine. Antibody-based therapeutics used as passive immunotherapy have been very successful in treatment of certain cancers and autoimmune disorders (Carter and Lazar, 2018). However, to this point there are only two therapeutic monoclonal antibodies approved by the FDA for anti-viral use (American Academy of Pediatrics Bronchiolitis Guidelines Committee, 2014; Markham, 2018). Although bnAbs have shown promise in suppressing viremia in treatment of HIV (Barouch et al., 2013; Caskey et al., 2015; Shingai et al., 2013) they are still far from being an ideal anti-retroviral treatment. Therefore, the power of bnAbs in treating disease lies not in direct administration, but in using these bnAbs to engineer vaccines, a paradigm known as structure-based reverse vaccinology.

Structure-based reverse vaccinology involves determining the three-dimensional structure of an antibody in complex with its antigen and designing a vaccine immunogen based on the precise epitope. This concept was first applied to designing a vaccine for HIV, due to the failure of traditional vaccines to prevent HIV infection (Flynn et al., 2005; Pitisuttithum et al., 2006). The first proof-of-principle experiments targeted bnAbs 2F5 and 4E10, which bind continuous epitopes in the HIV MPER. These experiments showed that it was possible to transplant the epitope onto an acceptor protein scaffold, and that these constructs maintain high-affinity binding to bnAbs (Correia et al., 2010; Ofek et al., 2010). Later experiments extended this work onto the discontinuous epitope of bnAb b12 (Azoitei et al., 2011), which is significant since most antibodies bind discontinuous epitopes (Rubinstein et al., 2008). In addition to HIV this strategy has also been

used to design a vaccine immunogen for respiratory syncytial virus, which was shown to induce neutralizing antibodies from macaques after immunization (Correia et al., 2015).

Although structure-based reverse vaccinology has been validated as a feasible strategy, many challenges remain for designing an effective HIV vaccine. BnAbs targeting HIV typically have a very high load of somatic hypermutation, in some cases up to 44% mutation at the amino acid level (Corti and Lanzavecchia, 2013). These somatic mutations are required for potent and broad neutralization (Klein et al., 2013; Mouquet et al., 2010; Scheid et al., 2011). In addition it was found that the majority of HIV bnAbs do not bind to Env when reverted to the germline amino acid sequence (Hoot et al., 2013; McGuire et al., 2014; Xiao et al., 2009). Therefore, a vaccine strategy using immunogens based on fully matured bnAbs is unlikely to work if these immunogens are unable to stimulate the germline precursors. This observation gave rise to a new strategy of germline-targeting immunogen design, where immunogens are designed not to bind to the mature bnAbs but rather to their germline forms. Germline-targeting immunogens have shown promise in inducing CD4 binding site bnAbs in mice, not only targeting the germline but inducing a gradient of bnAbs along the affinity maturation pathway (Briney et al., 2016; Jardine et al., 2016; 2015; 2013).

Recently, the same strategy of structure-based vaccinology has been applied to influenza. Traditionally influenza vaccines have been much more effective than HIV, therefore the need for a structure-based vaccine has not been as great. However, with seasonal influenza vaccination efficacy between roughly 30-60% (Belongia et al., 2016), there is clearly room for improvement. Influenza A virus subtypes H1 and H3 and two type B virus lineages currently circulate among human populations (Hannoun, 2013). However, subtypes H5, H7, and H9 circulate among livestock and periodically emerge into human populations and cause epidemics (Kumar et al.,

2018). Ideally a "universal influenza vaccine" would induce protective immunity for all known subtypes of influenza to prevent new viruses from emerging.

One approach has been to create an influenza antigen reactive to bnAbs targeting all different antigenic variants within a given subtype. Giles *et al.* created a broadly reactive antigen through a genetics-based approach of making a consensus HA incorporating amino acids of different clades within the H5 subtype, which was shown to elicit a pan-H5 response (Giles and Ross, 2011; Giles et al., 2012a; 2012b). This approach has also been shown to elicit a pan-subtypic response in influenza viruses of the H1 and H3 subtypes (Carter et al., 2016; Wong et al., 2017).

Another approach has been to extend the structure-based design methods developed in the HIV field to elicit bnAbs against influenza. Antibodies targeting the stem domain of influenza HA are a fitting application of these methods, as the stem domain tends to be the most broadly reactive epitope on HA but is poorly immunogenic in the context of whole virus (Ellebedy et al., 2014; Krammer and Palese, 2013; Sui et al., 2011). Two separate groups engineered stable HA stem domains that can recognize anti-stem bnAbs. These immunogens were shown to adopt a similar conformation as the native stem domain and could protect mice and ferrets from challenge with heterosubtypic influenza viruses (Impagliazzo et al., 2015; Yassine et al., 2015). Similar approaches have engineered full-length HA molecules with hyperglycosylated head domains to increase the immunogenicity of the stem domain (Eggink et al., 2014).

Computational protein modeling

Overview of protein modeling techniques

Computational modeling is a powerful method for modulating the activity of proteins and other macromolecules. The number of possible protein sequences is orders of magnitude larger

than the number of proteins with determined structure, with ~10^5 protein structures currently in the Protein Data Bank (PDB) and ~10^8 protein sequences in the UniProt database (Berman et al., 2000; The UniProt Consortium, 2017). Given this disparity there will always be a need to make predictions about the structure and function of a given amino acid sequence using molecular modeling. Protein modeling can be characterized into two major tasks. The first is the folding problem, which involves predicting the 3-dimensional structure of a protein based on its amino acid sequence. Solving the folding problem has long been a holy grail of structural biology (Dill and MacCallum, 2012). The difficulty of this problem is explained by Levinthal's paradox, which states that the total conformational space of a small protein is at least 10^{300}, impossible to completely sample in the timescale in which proteins are known to fold (Levinthal, 1969). Therefore, protein folding algorithms must approximate the biased sampling pattern used by nature, the parameters of which are still undefined. Early attempts at solving the folding problem used physics-based force fields and coarse-grained representations of the amino acid side chains to simplify sampling (Levitt, 1976; Levitt and Warshel, 1975). As computing power has increased the capabilities of physics-based approaches have also increased, eventually simulating the dynamics and folding of proteins on a timescale of milliseconds (Shaw et al., 2010). However, with the increase in the number of experimentally determined structures, it has become possible to speed up simulations using knowledge-based rather than physics-based energy potentials. These knowledge-based, or statistical, potentials use conformational statistics of known structures in the PDB to approximate the free energy of a given protein structure (Simons et al., 1999; Sippl, 1990). The ROSETTA modeling software was one of the first to use statistical potentials in protein modeling, which allowed simulation of much larger macromolecules than previously possible (Kuhlman et al., 2003; Simons et al., 1999).

The second major task in protein modeling is known as protein design, or the inverse folding problem. This asks the question, given a protein conformation, what amino acid sequences can fold into that conformation? This task shares many similarities to the protein folding problem, such as use of both physics-based and knowledge-based potentials, and an intractable search space to fully explore. The first applications of protein design showed that it was possible to redesign a protein sequence while maintaining the same 3-dimensional fold (Dahiyat and Mayo, 1997; Desjarlais and Handel, 1995; Harbury et al., 1998). Once it became clear that modeling software could recapitulate an existing protein fold, the goal shifted to designing an amino acid sequence that could adopt a new fold, which was accomplished in a seminal paper by Kuhlman *et al.* (Kuhlman et al., 2003). This work designed a new protein fold that had not been seen in any protein in the PDB up to that point. Since development of this new fold, there have been many successes in the protein design field, including design of small proteins to bind influenza hemagglutinin (Fleishman et al., 2011b; Strauch et al., 2017), design of proteins that self-assemble into nanocages (King et al., 2012), and HIV immunogens targeting broadly neutralizing antibodies (Jardine et al., 2013).

Challenges in modeling antibody structure and function

Although there are applications of protein modeling in many disciplines, this thesis will specifically focus on modeling and design of antibodies. The large number of theoretically possible antibodies and the large number of antibodies actually present in humans prohibit a comprehensive experimental characterization of antibody structure and dynamics. While great progress has been made in antibody structure determination by X-ray crystallography – currently around 2,000 depositions in the PDB contain the phrase “antibody” – the number of experimental structures available in the PDB will always be small compared to the total immune repertoire, leaving room

for structure prediction of important antibodies with unknown structure. As antibody structures in the PDB have increased exponentially in recent years (Figure I.7), computational biologists have gained a greater understanding of the molecular determinants of proper loop folding and antigen binding, ultimately allowing high-throughput, accurate structural modeling on a scale unavailable to experimental methods alone. Understanding the structural determinants of the antibody-antigen interaction – *i.e.* how the paratope engages the epitope – is critical for understanding antibody function and processes such as affinity maturation.

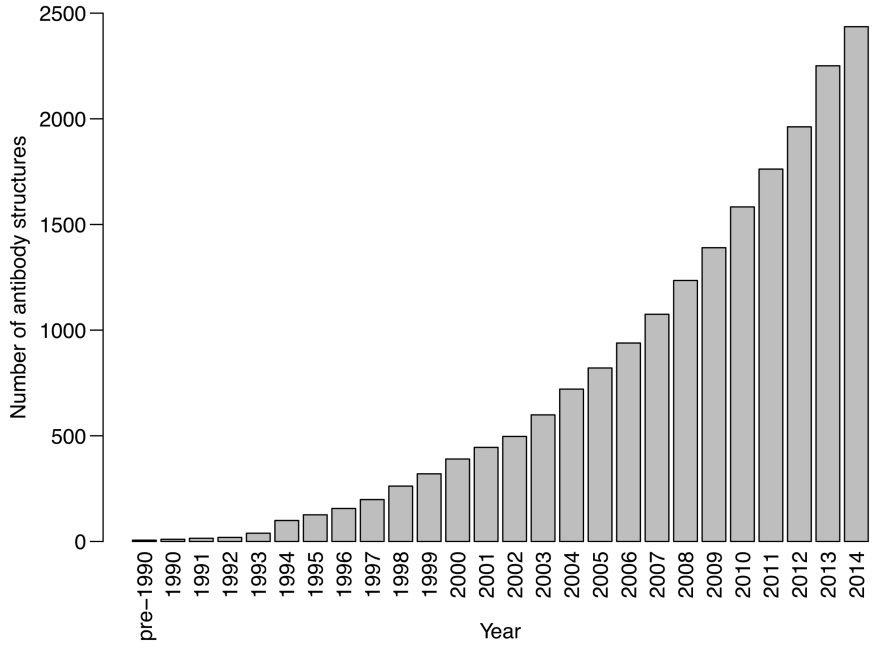


Figure I.7. Exponential growth of antibody structures in the PDB. Advances in antibody isolation and characterization technologies have led to an explosion in available antibody structures. This abundance of structural information can be used to improve computational modeling technologies. Figure is adapted from (Sevy and Meiler, 2014).

In silico affinity maturation

The computational design of antibodies is not only the most stringent test of our understanding of the rules that govern antibody structure and interaction, it also has exciting applications in designing an antibody optimized for a given epitope (affinity maturation) or an antibody that recognizes multiple similar target epitopes (broad neutralization). Through this approach the relation between sequence, structure, and activity of antibodies can be better understood, as the sequence and structural space can be explored in a more comprehensive manner than possible by analysis of naturally occurring antibodies only. For this paradigm to reach its full potential, knowledge of the optimal antibodies to engage an epitope and the relation between sequence, structure, and activity inferred from computational design must be integrated.

Affinity maturation, as previously described in detail, is a process by which the variable region of an antibody undergoes high levels of mutation to select for a variant with increased binding affinity for its target. As affinity maturation is a stochastic process, it can be simulated by computational algorithms that generate random mutations and measure their fitness by an energy function. In an early example of *in silico* affinity maturation, Clark *et al.* were able to use computational design to mature an antibody and generate candidate sequences with higher predicted affinity (Clark et al., 2006). Using a combination of side chain repacking and electrostatic optimization, a triple mutant was created with 10x higher affinity. A comparable increase in affinity was achieved by Lippow *et al.* by redesigning an anti-lysozyme antibody along with the therapeutic antibody cetuximab (Lippow et al., 2007). The design protocol was also able to predict mutations in bevacizumab that had been previously shown to increase affinity. The designed mutations affected primarily the electrostatic nature of the binding interface, either by removing a poorly satisfied polar residue at the interface or adding a polar residue at the solvent-

facing periphery of the interface. A similar approach was used to increase the species cross-reactivity of an antibody, rather than increasing affinity for a previously targeted antigen. By analyzing sequence differences between two serine protease orthologs, Farady *et al.* created novel antibody designs by restricting the search space to positions that contact points of difference between orthologs (Farady et al., 2009). In this manner they were able to target positions that would be most likely to establish new contacts across the binding interface to enable interaction at a reasonable affinity. This method was able to create antibody mutants with increases in affinity of over two orders of magnitude.

One significant limitation of most computational design protocols is that they require a high-resolution crystal structure of the antibody-antigen complex, or alternatively high-resolution structures of each component separately. However, several antibody designs have been made for complexes that do not have a solved structure available, using a combination of comparative modeling, protein-protein docking, and design. Barderas *et al.* used experimental epitope mapping data to dock a comparative model of an anti-gastrin antibody onto the surface of its target (Barderas et al., 2008). They then used the docked models to estimate regions of antibody-antigen interaction, and created mutants using both phage display and *in silico* affinity maturation to mutagenize antibody residues in contact with the antigen and produce designs with high predicted affinity. In several cases the *in silico* suggested mutations matched the mutations seen by directed evolution, and overall the designs were able to increase affinity to nanomolar levels. Another case used docking of an anti-dengue antibody with an NMR-mapped epitope to identify and rationally design mutations in the antibody CDR loops (Simonelli et al., 2013). The authors used this information to create several types of antibody mutations, including those that abolish binding, those that increase affinity for a single target, and those that increase breadth of binding to multiple serotypes.

Design of antibody breadth

In certain cases, it is desirable to design an antibody for increased affinity against not only one target, but multiple targets. As a naturally evolved antibody will typically undergo affinity maturation against a single antigen, designing an antibody *in silico* against a variety of antigens can surpass the biological limitations of antibody evolution. This approach of designing a protein against multiple targets is known as multistate design, for the multiple protein states included in the simulation. Computational approaches to multistate design are capable of determining the protein sequence optimal for binding an arbitrary number of binding partners (Havranek and Harbury, 2003; Leaver-Fay et al., 2011a). This technique has been applied to explore the changes in antibody sequence and conformation responsible for the shift from a polyspecific, germline antibody to one with higher affinity for a single target. In complementary works, Babor *et al.* and Willis *et al.* used multistate design to show that antibody germline sequences are optimal for conformational flexibility of both CDR loops and framework residues, allowing binding of multiple targets, whereas affinity matured antibodies have decreased flexibility (Babor and Kortemme, 2009; Willis et al., 2013). They also identified the key residues responsible for either mono- or multi-specificity for several commonly seen germline genes. These studies validate the biological relevance of design algorithms, since sequences can be both computationally matured and reverted to germline by using different sets of antigens as inputs.

Significance and Innovation

In this thesis I describe my work using molecular modeling to engineer cross-reactivity into antibodies targeting influenza and HIV. As previously described, there are a number of broadly neutralizing antibodies against both influenza and HIV that have been identified by

experimental methods. Many of these antibodies have structural data as well, providing a precise description of the molecular details of the antibody-antigen interaction. However, none of these antibodies are universally effective. Due to the high level of sequence variability of both HIV Env and influenza HA, it has to this point been impossible to identify a "silver bullet" antibody that can potently neutralize all potential variants of either virus. In my work I used molecular modeling and design to define the limitations of antibody potency and breadth.

As previously discussed there have been many reports of using computational design to improve antibody affinity (Barderas et al., 2008; Clark et al., 2006; Lippow et al., 2007; Marvin and Lowman, 2003; Willis et al., 2015). In addition there are experimental methods that can simulate affinity maturation *in vitro* to generate antibodies with greater affinity against a single target (Boder et al., 2000; Daugherty et al., 2000; Nelson et al., 2007; Wu et al., 2017). However, it remains challenging to not only improve the affinity of an antibody against a single target, but to increase its breadth by considering a diverse panel of viral variants. Computational methods for this task, known as multistate design, have been described, but are all very computationally expensive and limited in both the number of antigens, or states, that can be considered, and the size of the binding site to be redesigned (Leaver-Fay et al., 2011a; Yanover et al., 2007). In this thesis I describe new methods for multistate design that can be applied on a much larger scale than previously feasible. Although the focus in this thesis is on antibody design, these methods were developed to be sufficiently general to be applied to any protein multistate design problem.

This work is significant on both a basic science and translational level. As the therapeutic market for monoclonal antibodies increases, the importance of methods for antibody engineering also increases. Many antibody therapeutics have been engineered by computational and experimental methods to improve their therapeutic properties (Bostrom et al., 2009; Lehmann et

al., 2015; Wang et al., 2014; Wu et al., 2007). Antibody engineering in a multistate context has great implications for modulating the behavior of these antibody therapeutics. In addition, knowledge of broadly neutralizing anti-viral antibodies can be used to formulate an effective vaccine. By applying multistate design to existing antibodies, it is possible to improve antibody breadth beyond natural evolution and devise vaccines to elicit antibodies with these characteristics. It is also clear that multi-specific antibodies play a significant role in the antibody response to HIV and in the development of broadly neutralizing antibodies (Liao et al., 2011; Liu et al., 2015; Mouquet et al., 2010; Williams et al., 2015). I use multistate design to analyze how bnAb precursors may mature against different viral targets and how they could be targeted by a vaccination strategy. I also apply the algorithm to influenza antibodies to learn about the molecular details of the tradeoff between breadth and affinity.

CHAPTER II.

Design of protein multi-specificity using an independent sequence search reduces the barrier to low energy sequences

Adapted from Sevy, A. M., Jacobs, T. M., Crowe, J. E. & Meiler, J. Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *PLoS Comput. Biol.* **11,** e1004300 (2015).

Author contributions: I wrote the algorithm described in this chapter and ran all computational benchmarks, under the mentorship of Jens Meiler and James Crowe. I designed all experiments, analyzed data with my co-mentors, and created all figures in this chapter. All figures are reprinted with permission from the publisher.

Abstract

Computational protein design has found great success in engineering proteins for thermodynamic stability, binding specificity, or enzymatic activity in a 'single-state' design (SSD) paradigm. Multi-specificity design (MSD), on the other hand, involves considering the stability of a protein in multiple conformations recognizing multiple binding partners, i.e. to stabilize multiple protein states simultaneously. We have developed a novel MSD algorithm, which we refer to as REstrained CONvergence in multi-specificity design (RECON). The algorithm allows each state to adopt its own sequence throughout the design process rather than enforcing a single sequence on all states. Convergence to a single sequence is encouraged through an incrementally increasing

convergence restraint for corresponding positions. Compared to MSD algorithms that enforce (constrain) an identical sequence on all states the energy landscape is simplified, which accelerates the search drastically. As a result, RECON can readily be used in simulations with a flexible protein backbone. We have benchmarked RECON on two design tasks. First, we designed antibodies derived from the same germline gene against their diverse targets to assess recovery of the germline, polyspecific sequence. Second, we design "promiscuous", polyspecific proteins against all binding partners and measure recovery of the native sequence. We show that RECON is able to efficiently recover native-like, biologically relevant sequences in this diverse set of protein complexes.

Introduction

Computational protein design is an invaluable tool for protein engineers seeking to create a protein with novel properties. Protein design, also known as the inverse folding problem, involves searching for a sequence that stabilizes a desired, given conformation. Besides the obvious goal – to give the protein increased thermodynamic stability (Kuhlman et al., 2003; Miklos et al., 2012; Yang et al., 2014a) – protein design often pursues the goal of creating new function. This can include for example redesigning an antibody to recognize a new variant of a target protein (Farady et al., 2009), designing an enzyme to bind the transition state for a new chemical reaction (Siegel et al., 2010), or redesigning a DNA-binding protein to recognize a different DNA sequence (Ashworth et al., 2010). Most success in protein design has been achieved through a single-state design (SSD) task, i.e. the free energy minimization of a single protein conformation to increase its stability (Harbury et al., 1998; Kortemme et al., 2004; Kuhlman et al., 2003).

Multistate design approaches

In contrast to SSD, multistate design (MSD) minimizes the free energy of multiple protein conformations ("states") simultaneously. This enables negative design, which involves destabilizing a certain conformation to shift relative occupancy to alternate conformations, which is useful in designing proteins with binding selectivity. MSD has been applied successfully in a number of cases, including the design of a protein conformational switch (Ambroggio and Kuhlman, 2006), design of selective b-ZIP binding peptides (Grigoryan et al., 2009), and design of an enzyme with DNA cleavage specificity (Ashworth et al., 2006), among others (Allen et al., 2010; Havranek and Harbury, 2003).

Algorithmic requirements for multistate design

All MSD algorithms have at their core a fitness function that defines the favorability of a given sequence based on its corresponding energy in each state. The major challenge in fixed backbone MSD is efficient optimization of side chain rotational isomer ("rotamer") placement, using the fitness function as the objective function. As more states are considered it becomes increasingly difficult to find the minimum energy sequence on a fixed backbone. As the same sequence on all states is constrained, extensive sampling in sequence and rotamer space is required. This is often accomplished via thorough but slow genetic algorithms (Havranek and Harbury, 2003; Humphris and Kortemme, 2007; Leaver-Fay et al., 2011a).

Challenges in expanding the scope of multistate design

This difficulty in reaching the global minimum in a basic fixed backbone design problem precludes the possibility of using alternate sampling strategies, such as iterating between backbone minimization and rotamer optimization. However, these techniques have been used in SSD to great effect and are often critical to find the lowest energy conformation and sequence (Harbury et al.,

1998; Kuhlman et al., 2003). In result, MSD algorithms can arrive at an incorrect solution even after successful sequence optimization just because the fixed backbone precludes the lowest energy sequence and conformation from being sampled. This issue can be partially resolved by the inclusion of multiple backbone conformations as separate states (Davey and Chica, 2014). However, there is a need for a method that can more efficiently reach the optimal MSD solution for an arbitrary number of input states without relying on the commonly held "fixed backbone assumption".

Multi-specificity design as single-state design with restraints

To this end, we have developed a novel MSD algorithm, referred to as REstrained CONvergence in multi-specificity design (RECON). The algorithm is based on a different conception of MSD, wherein each state independently explores sequence space to reach its energetic minimum. A step-wise increasing convergence restraint is applied such that corresponding positions in different states converge on the same amino acid. By encouraging sequence convergence between different states rather than enforcing a single sequence, we hypothesize that energetic barriers to the fittest solution collapse, reducing the ruggedness of the energetic landscape in a MSD problem to SSD-like complexity. In result the search efficiency and speed are substantially increased allowing for the sampling of additional degrees of freedom. Further, we hypothesize that including backbone conformational sampling reduces the chance that the low energy and possibly correct solutions are excluded from the search space.

Results

The restrained convergence algorithm

The RECON algorithm allows separate states to explore their own local sequence and

conformational space to optimize free energy, while restraining corresponding residues in different states with a convergence restraint to encourage sequence convergence. Convergence restraints are kept small in early rounds, to allow each state to explore its own lowest energy sequence, and ramped up in later rounds to encourage sequence convergence between different states. This is followed by a greedy selection step, which evaluates all candidate amino acids at positions that fail to converge, and selects the one that results in the lowest fitness when applied over all states. This greedy selection is included in order to ensure that one multi-specific sequence is generated from each design trajectory. Backbone minimization steps can be included between design rounds to relieve slight clashes between side chains. Pseudocode describing the implementation of the algorithm is shown in Figure II.1. Individual states optimize rotamer placement using a simulated annealing Monte Carlo search, sampling from a predefined rotamer library (Dunbrack, 2002; Leaver-Fay et al., 2011b). However, we emphasize that this method can be applied to any multi-specificity problem using an arbitrary optimization method and scoring function.

Reduction of energy barriers in restrained multi-specificity design

By allowing each state to determine its optimal sequence independently, we can collapse the energy barrier to reaching a "compromised" sequence that results in low energy in all states. We propose a scenario in which encouraging sequencing convergence in this way can reduce the energetic barrier and enable convergence on a low energy solution (Figure II.2). In this scenario, two separate mutations from residue identity A to B are needed for the lowest fitness over both states. Each single mutation will encounter a high energy penalty and rarely selected by a genetic algorithm – only when both mutations are stochastically placed together will the solution emerge, which may take a large number of evaluations. However, when sequence convergence is encouraged rather than enforced, each state will identify an intermediate solution in early rounds,

```
Main
Begin
    // Perform initial relaxation of all states to relieve clashes in the crystal structure
    Relax( state1 )
    Relax( state2 )
    // Iterate through design rounds with increasing restraint values to encourage sequence
    // convergence
    for restraint_value in ( 0.5, 1.0, 1.5, 2.0 )
        // Optimize sequence of state 1, with state 2 as a reference
        OptimizeSequence( state1, state2, restraint_value )
        // Optimize sequence of state 2, with state 1 as a reference
        OptimizeSequence( state2, state1, restraint_value )
        // Minimization step to relieve clashes between design rounds
        Minimize( state1 )
        Minimize( state2 )
    // Greedy selection to pick the optimal amino acid at each position if the restraints are
    // insufficient to cause convergence
    GreedySelection( state1, state2, designable_positions )
End

OptimizeSequence( optimize_state, reference_state, restraint_value )
Begin
    // Optimize sequence of optimize_state using a simulated annealing rotamer optimization
    // algorithm, with the objective function being the energy of the state plus sequence
    // restraints
    SimulatedAnnealing( optimize_state, objective_function = EnergyWithRestraints( optimize_state,
    reference_state, restraint_value ) )
End

EnergyWithRestraints( optimize_state, reference_state, restraint_value )
Begin
    // Measure the energy of optimize_state, adding a convergence restraint for
    // positions which match between optimize_state and reference_state
    restrained_energy = energy( optimize_state )
    for position in designable_positions
        if optimize_state.residue_at( position ) == reference_state.residue_at( position )
            restrained_energy -= restraint_value
    return restrained_energy
End

GreedySelection( state1, state2, designable_positions )
Begin
    for position in designable_positions
        // If positions do not agree between states, force greedy selection
        if state1.residue_at( position ) != state2.residue_at( position )
            // Place amino acid from state 1 onto state 2 and measure fitness
            energy_sequence_1 = state1.energy() + state2.replace( position, state1.residue_at( position )
            ).energy()
            // Place amino acid from state 2 onto state 1 and measure fitness
            energy_sequence_2 = state1.replace( position, state1.residue_at( position ) ).energy() +
            state2.energy()
            // If sequence from state 1 has lower fitness, accept this amino acid
            if energy_sequence_1 < energy_sequence_2
                state2 = state2.replace( position, state1.residue_at( position ) )
            // Else accept amino acid from state 2
            else
                state1 = state1.replace( position, state2.residue_at( position ) )
End
```

Figure II.1. Pseudocode describing the implementation of the RECON algorithm.

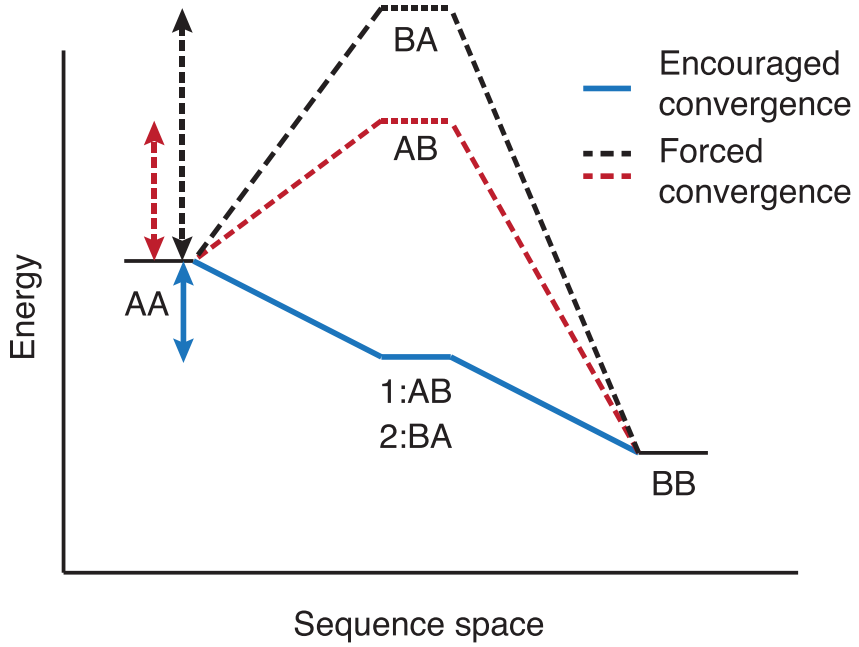


Figure II.2. Schematic showing proposed energy landscape of forced vs. encouraged sequence convergence in MSD. By allowing each state to maintain its own sequence and explore sequence space independently, RECON is able to provide an intermediate solution in an MSD problem, enabling more rapid determination of a low energy solution. Dashed lines represent forced convergence, where both states must adopt the same sequence (either AB or BA), whereas the solid line represents encouraged convergence, where state 1 can adopt sequence AB while state 2 adopts BA. This creates a lower energy intermediate state leading to more rapid adoption of the optimal solution, sequence BB.

and in later rounds the most favorable solution will be selected from the differing states. This collapses the barrier on the pathway to a favorable solution and reduces the steps necessary to find that solution.

Germline gene reversion benchmark

To benchmark RECON, we considered two types of design problems. In the first, mature antibodies derived from a common germline gene were entered into MSD in complex with their target antigens. It has been shown that MSD of mature antibodies results in a higher rate of germline sequence reversion than SSD, implying that the germline sequence is near-optimal for polyspecificity (Babor and Kortemme, 2009; Willis et al., 2013). Therefore, we designed each

antibody against its respective targets and used germline sequence recovery as an indirect measure of the rate of recovery of an optimal solution. We used antibodies derived from three different germline genes - V_H1-69, V_H3-23, and V_H5-51. The number of antibody-antigen complexes per germline gene ranged from 3 to 6 (Table II.1).

Table II.1. Complexes used in common germline antibody benchmark.

V_H germline gene	**Variable positions[a]**	**Antibody**	**Ligand**	**PDB ID**
V_H1-69	40	D5	gp41	2CMR
		F10	H5/Vietnam/1203/2004	3FKU
		CR6261	H5/Vietnam/1203/2004	3GBM
		8066	gp41	3MA9
		8062	gp120	3MAC
		1281	gp41	3P30
V_H3-23	31	Pertuzumab	ErbB2	1S78
		G6	VEGF	2FJG
		Apu2.16	Ubiquitin	3DVN
		E2	MT-SP1	3BN9
V_H5-51	30	2219	UG1033	2B1A
		K1-70	TSHR	2XWT
		Ustekinumab	IL-12	3HMX

RECON was benchmarked on three sets of mature antibodies derived from the same V_H gene. Effective MSD should result in reversion of mature antibodies to the polyspecific germline sequence.

[a]Germline sequence and positions varying from germline are inferred from IMGT/3D Structure-DB (Kaas et al., 2004).

Promiscuous protein design benchmark

The second task was to design a set of "promiscuous" proteins, proteins that have been crystallized in complex with multiple binding partners, against each of these partners. Similar to polyspecific germline antibodies, promiscuous proteins have been shown to have a native sequence that is near-optimal for binding to all of the partners (Fromer and Shifman, 2009; Humphris and Kortemme, 2007). Therefore an effective MSD protocol would result in a high rate of native sequence recovery. A set of five promiscuous proteins derived from a study done by Humphris *et*

al. was used (Humphris and Kortemme, 2007), in addition to two broadly neutralizing anti-influenza hemagglutinin antibodies (Table II.2) (Corti et al., 2011; Ekiert et al., 2009).

Design algorithms included in benchmark

Benchmark cases were designed using three separate design methods. First, design was performed using RECON with a fixed backbone. Fixed backbone design has to this point been the standard in MSD due to the complexity involved in recalculating rotamer interactions for each backbone movement. However, using fixed backbone design alone is prone to false negatives, as sequences that may be highly favorable with a small shift in backbone conformation are discarded. One of the unique advantages of RECON is its ability to incorporate iterative rounds of rotamer packing and backbone minimization. Therefore, we included such an iterative protocol as the

Table II.2. Complexes used in promiscuous protein benchmark.

Promiscuous protein	Binding partner	PDB ID	Designable positions[a]
CR6261	H5/Vietnam/1203/2004	3GBM	19
	H1/BrevigMission/1/1918	3GBN	
FI6v3	H1/California/04/2009	3ZTN	21
	H3/Aichi/2/1968	3ZTJ	
CheY	FLiM	1F4V	15
	CheA	1FFG	
	CheZ	1KMI	
Elastase	Elastase Inhibitor	1EAI	25
	Elafin	1FLE	
	Hybrid Squash Inhibitor	1MCV	
FYN SH3 Domain	HIV-1 NEF Protein	1AVZ	7
	SAP	1M27	
PapD Chaperone	PapE	1N0L	28
	PapK	1PDK	
	PapD Homodimer	1QPP	
Ran	Importin beta	1IBR	24
	Exportin CSE1P/KAP60P	1WA5	

RECON was benchmarked on a set of promiscuous proteins that have been crystallized in complex with multiple partners. As the native sequence is near optimal for binding of all partners, MSD should recover the native sequence at a high rate.

[a]Residues determined to be at the interface with all binding partners. See methods for details on interface residue calculations.

second approach in our benchmark. For comparison purposes, all complexes were also designed using the existing MSD application in ROSETTA (MPI_MSD), which operates on a fixed backbone (Leaver-Fay et al., 2011a). MPI_MSD differs from RECON in that it uses a genetic algorithm to create and advance mutations and a user-defined fitness function to assess fitness of each sequence. However, as both methods are built into the ROSETTA framework, they sample from the same rotamer library and use the same scoring function and are therefore suitable for comparison. In addition to native sequence recovery, we used the fitness of the top ten designs, defined as the sum of ROSETTA energies of all complexes, to analyze how effectively each protocol reached an energetic minimum. This fitness function has been previously used in studies of design of protein multi-specificity (Humphris and Kortemme, 2007). We use the term "design" to refer to sequence optimization of existing protein-protein complexes - however, it is important to note that these sequences were not experimentally characterized, and results reported are purely *in silico*.

Common germline derived antibodies

For common germline gene-derived antibodies, RECON was consistently able to recover the germline sequence at a higher rate than MPI_MSD (Figure II.3). Germline sequence recovery for RECON ranged from 55 – 94% using fixed backbone and 51 – 95% using backbone minimization, while recovery for designs using MPI_MSD ranged from 32 – 64%. When comparing RECON fixed backbone to MPI_MSD, it appeared that designs created by RECON, although higher in native sequence content, were also energetically less favorable. We therefore subjected all fixed backbone designs to a single round of ROSETTA relax energy minimization to relieve frustrations and allow for direct comparison of fitness of RECON incorporating backbone minimization to fixed backbone designs (Supplementary Table II.1). These post-minimization fitness values show that the energetic gap between RECON- and MPI_MSD-generated designs

was substantially closed, and that designs generated by any method occupied similar ranges of fitness. We observed that MPI_MSD tended to produce designs with the lowest fitness - however, it is important to note that rotamer optimization within ROSETTA is a stochastic process, with no guarantee of reaching the global minimum. Therefore, a protocol that performs hundreds of rounds of rotamer optimization, such as MPI_MSD, would be expected to produce better energies than one performing four rounds of optimization, such as RECON, independent of the sequence identity of structures being optimized.

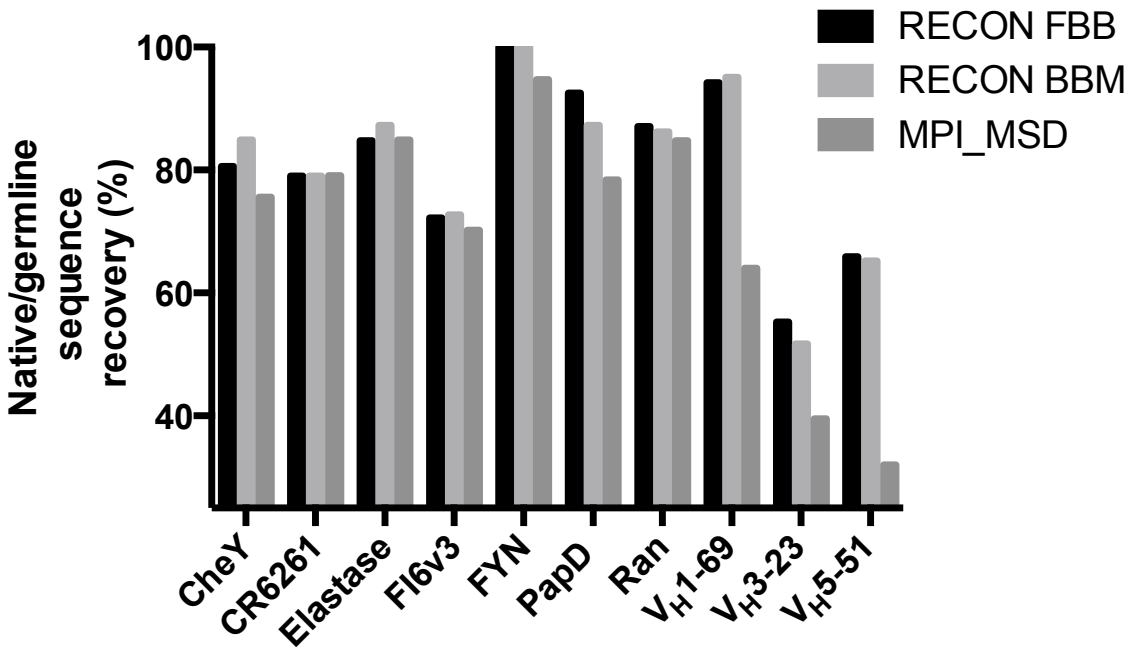


Figure II.3. Native/germline sequence recovery of designed complexes. 100 designs were generated using RECON, with both fixed backbone (FBB) and backbone minimized (BBM) protocols, and MPI_MSD. Sequences of the top 10% of models were compared to either the native sequence or, in the case of common germline-derived antibodies, to the germline sequence. See methods for details of native sequence recovery calculations.

Promiscuous protein complexes

RECON was able to recover the native sequence at a very high level for all promiscuous protein complexes – native sequence recovery ranged from 57 – 100% and 58 – 100%, using fixed backbone and backbone minimization, respectively. MPI_MSD generated designs with native sequence recovery ranging from 56 – 94% (Figure II.3). In most cases fitness of designs generated by RECON fixed backbone and MPI_MSD were very similar, suggesting that both methods have reached the energetic minimum (Supplementary Table II.1). Even though all methods reached a similar level of native sequence recovery and energetic fitness in a majority of these benchmark cases, RECON was able to reach these minima by searching a compressed sequence space, allowing for increased computational efficiency.

Importance of ramping convergence restraints on algorithm performance

We hypothesized that gradually ramping the convergence restraints will allow for sequence divergence in early rounds of design and enforce convergence in later rounds, leading to an improved result as it smoothens the energy landscape. To confirm the effects of gradually increasing the weight of the convergence restraint, we performed a control in which sequences were designed independently for each state with no convergence restraint, followed by sequence selection by the greedy selection used at the end of RECON (Supplementary Table II.2). This greedy selection algorithm performed significantly worse than RECON with gradual ramping convergence restraints, with worse native sequence recovery in all benchmark cases but one. In addition, in many benchmark cases fitness was significantly worsened for designs generated by this greedy selection protocol. These results indicate that ramping convergence restraints throughout the design protocol is critical for the increased performance of RECON.

Sequence recovery at positions that fail to converge

Based on the decreased performance of this greedy selection algorithm, it would be expected that RECON works best at positions where amino acids converge between different states by the end of the protocol and are not greedily selected. We therefore evaluated the convergence of amino acids at each position for the V_H5-51 benchmark set. We report the number of times a position failed to converge in 100 design trajectories for the 30 designed positions in this benchmark set (Supplementary Table II.3). The results suggest that most positions tend to be consistent in their patterns of convergence, and that the majority (21 out of 30) reach a common amino acid solution by the end of the protocol. The results of the greedy selection protocol suggest that failure to converge leads to a decrease in performance of the algorithm and selection of non-native amino acids. We therefore compared germline sequence recovery for positions that failed to converge in at least half of the design trajectories, as compared to those that converged in more than half of the trajectories, to determine whether these positions are substantially decreasing overall germline sequence recovery (Supplementary Table II.3). Surprisingly, positions that failed to converge actually showed a higher rate of germline sequence recovery than those that were able to converge through the application of convergence restraints (Supplementary Table II.3). These results indicate that, although the greedy selection algorithm should not be applied without first ramping convergence restraints to encourage convergence, the use of greedy selection for positions that fail to converge is not a limiting factor for obtaining high native sequence recovery.

RECON is able to circumvent high-energy intermediates

In the scenario proposed in Figure II.2, we hypothesize that RECON is able to circumvent high-energy intermediate sequences by encouraging rather than enforcing sequence convergence. We therefore analyzed the sequence trajectory of an example from the FI6v3 benchmark to support

this scenario (Figure II.4A). In early rounds, the two states diverge in sequence to explore their own energy landscapes. As restraints are increased in later rounds the two states converge on a compromised sequence that is the multi-specific solution for both, only adopting mutations when they are beneficial to both states. Although fitness values continue to decrease after encountering the compromised sequence, this is primarily due to the stochastic nature of rotamer optimization, such that increased optimization will result in a lower score. We focused on a set of complementary mutations that diverged in early rounds with a low convergence restraint, to test the hypothesis that the sequence preference of one state results in a high energy on the other state, and vice versa (Figure II.4A, highlighted in red). We found that the sequences preferred by state 1 (TSY) and state 2 (QQW) indeed resulted in higher energy when forcing one state to adopt both sequences than when each state was allowed to adopt its own sequence (Figure II.4B). This lowers the barrier to reaching the "compromised" sequence, adopting residues favorable to both state 1 and state 2, which in this case is the sequence QQY. Although this barrier is not as large as proposed in Figure II.2, we expect that this barrier will be lower in cases where two binding partners have highly similar binding surfaces, as is the case in our benchmark sets. However, when binding surfaces are more dissimilar, and therefore finding compromise residues is more critical to a favorable binding energy, we expect this barrier to be larger, and the benefit of an independent sequence search to be even greater.

Computational efficiency of design methods

In addition to measuring the sequence recovery and energetic fitness, we compared the computational efficiency of these three design protocols. We argue that, although in certain cases all methods were able to reach the same energetic minimum, RECON provides an added benefit in that it reached this minimum in a fraction of the time required to run MPI_MSD. To this end

we compared CPU hours of runtime for generating the previously discussed designs (Table II.3). As expected, RECON using a fixed backbone was the most efficient of the three protocols, followed by RECON incorporating backbone minimization, and MPI_MSD. This increase in efficiency is due to the reduction in search space by allowing each state to adopt its own sequence.

Generation of evolutionary sequence profiles

We hypothesize that RECON is able to operate at higher efficiency by restricting sampled sequences to more relevant sequence space. We further believe that our conception of "relevant" sequence space is reflected in an ensemble of biologically observed sequences, and that RECON should recover not only a native protein sequence, but also biologically tolerated mutations. To address this question we generated a position-specific scoring matrix (PSSM) of amino acid frequencies in evolutionarily related proteins to each benchmark protein using a PSI-Blast query (Altschul et al., 1997). Among the promiscuous proteins we restricted this analysis to non-antibodies, since the full-length sequence of a mature antibody is unlikely to have a large number of meaningful evolutionary counterparts. However, since antibodies in the common germline-encoded benchmark set were only designed in positions deriving from the V_H gene, we were able to derive a PSSM from other common V_H-encoded antibodies in the database. We then compared the PSSM to the amino acid frequency in corresponding positions in designed sequences to estimate how well the design protocol mimicked evolution. We measured agreement of sequence profiles using a modified Sandelin-Wasserman similarity to yield a percent similarity for each designed position that could then be averaged over the protein (Sandelin and Wasserman, 2004). Figure II.5A shows a comparison of positions in the V_H5-51 benchmark where designs either agreed or disagreed with evolutionary sequence profiles - the degree of agreement could then be quantitated by the percent similarity calculated over each position.

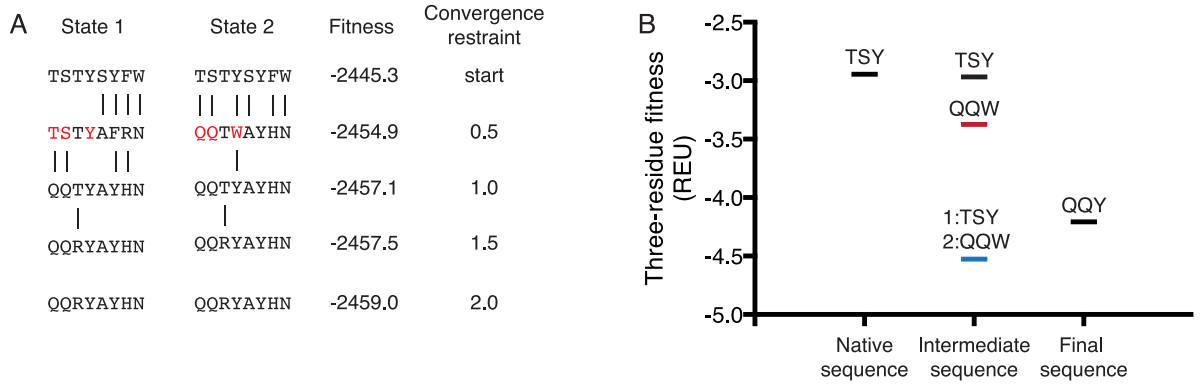


Figure II.4. Encouraging sequence convergence in RECON can avoid high-energy sequence intermediates. A. An example design trajectory of RECON in the FI6v3 benchmark through four design rounds is shown. Sequences tend to diverge in early rounds when convergence restraints are kept low, whereas in later rounds when restraints are increased states are encouraged to adopt a single solution. The figure displays one example from the fixed backbone design protocol, with convergence restraints removed before reporting fitness. The two states showed different preferences for residues highlighted in red. B. Residues highlighted in panel A were applied to the opposing state to analyze the energetic barrier of forced sequence convergence. The energy of these three residues was analyzed when the sequence favored by state 1 (TSY) was applied to state 2, and vice versa with the sequence QQW (intermediate sequence, black/red lines). This was compared to the three-residue fitness when each state was allowed to adopt its own preferred sequence (intermediate sequence, blue line). Energies were compared to the final, "compromised" sequence (QQY). These three amino acids occurred at positions 28, 30, and 53, respectively.

Table II.3. Comparison of CPU runtimes for multi-specificity design using different algorithms.

	CPU Hours		
Benchmark case	**RECON FBB**	**RECON BBM**	**MPI_MSD**
CheY	12.0	24.0	61.1
CR6261	20.8	66.0	137.5
Elastase	24.2	47.7	198.9
FI6v3	21.2	80.2	46.1
FYN	0.8	12.8	21.2
PAPD	37.9	99.5	129.1
Ran	23.1	153.3	276.9
V_H1-69	48.7	171.1	487.1
V_H3-23	43.7	98.1	167.5
V_H5-51	19.5	71.7	95.7
Average	**25.2**	**82.4**	**162.1**

Runtimes in CPU hours for generation of 100 designs using RECON, both by fixed backbone (FBB) and backbone minimized (BBM) methods, and MPI_MSD algorithms.

Comparison of designs to observed sequence profiles

We found that RECON was able to create sequences that more closely mirrored natural sequence variation than MPI_MSD (Table II.4, Figure II.5B). Averaging over the benchmark cases, we observed 69, 73, and 57% similarity to evolutionary sequence profiles using RECON fixed backbone, backbone minimized, and MPI_MSD, respectively. This pattern was especially strong in benchmark cases with large numbers of designable residues, as the number of designed residues correlated positively with the improvement of RECON over MPI_MSD in recapitulating evolutionary sequence profiles (Figure II.5C). When comparing the four largest benchmark cases by number of designable residues (three common germline-derived antibodies and the PAPD complex), RECON shows a marked improvement over MPI_MSD in recovery of evolutionary sequence profile (Figure II.5D). Although this result is not significant due to a small sample size, it is suggestive of the additional benefit provided by RECON when applied to large, computationally intensive design problems. We hypothesize that this is due to compressed sequence space explored by RECON. When design problems are relatively small, the genetic algorithm employed by MPI_MSD is able to efficiently search through sequence space for a low-energy solution. However, when the sequence space increases in a large design problem the compressed sequence search is more advantageous.

RECON searches a compressed, more relevant sequence space

We have shown that designs generated by RECON tend to more closely represent the evolutionary sequence profiles of our benchmark proteins when compared with MPI_MSD. We propose that this is accomplished via a more focused sequence search within the biologically relevant space. To further support this claim, we have analyzed the sequence space searched by RECON and MPI_MSD and compared it to the final output sequences of the top ten designs for

the V_H5-51 benchmark set (Supplementary Figure II.1). We generated the sequence space profile by including any residue that was sampled at any step of the design protocol at each position, and then compared this profile to the final sequences among the top ten designs. Presumably the most efficient algorithm would only sample the sequences that are eventually selected as low energy solutions, resulting in a similarity of 100% between sequence space explored and output designed sequences. Therefore, we used this similarity as an indicator of the degree of "wasted" sequence space, which is explored but never part of a low energy solution. Comparison of the profiles generated by RECON on a fixed backbone and MPI_MSD show that RECON explores space much more closely constrained to the final low energy sequences, with a similarity score of 92.3%, as compared to 79.5% for MPI_MSD. This further supports the claim that RECON searches a compressed search space to encounter a low energy multi-specific solution.

Table II.4. Comparison of design-generated sequences to evolutionary sequence profiles of input proteins.

	Evolutionary sequence similarity (%)[a]		
Benchmark case	**RECON FBB**	**RECON BBM**	**MPI_MSD**
CheY	56.3	70.5	57.5
Elastase	60.3	70.7	65.9
FYN	87.0	87.0	96.0
PAPD	61.7	65.3	52.4
Ran	76.6	79.3	82.5
V_H1-69	90.6	91.7	32.0
V_H3-23	50.7	50.7	36.4
V_H5-51	69.0	67.0	30.4
Average	**69.0**	**72.8**	**56.6**

Designs produced by MPI_MSD or fixed backbone (FBB) or backbone minimized (BBM) RECON algorithms were compared to sequence profiles of evolutionarily related proteins at designed positions.

[a]Sequence similarity is computed as the Sandelin-Wasserman similarity, normalized as a percentage. See methods for details.

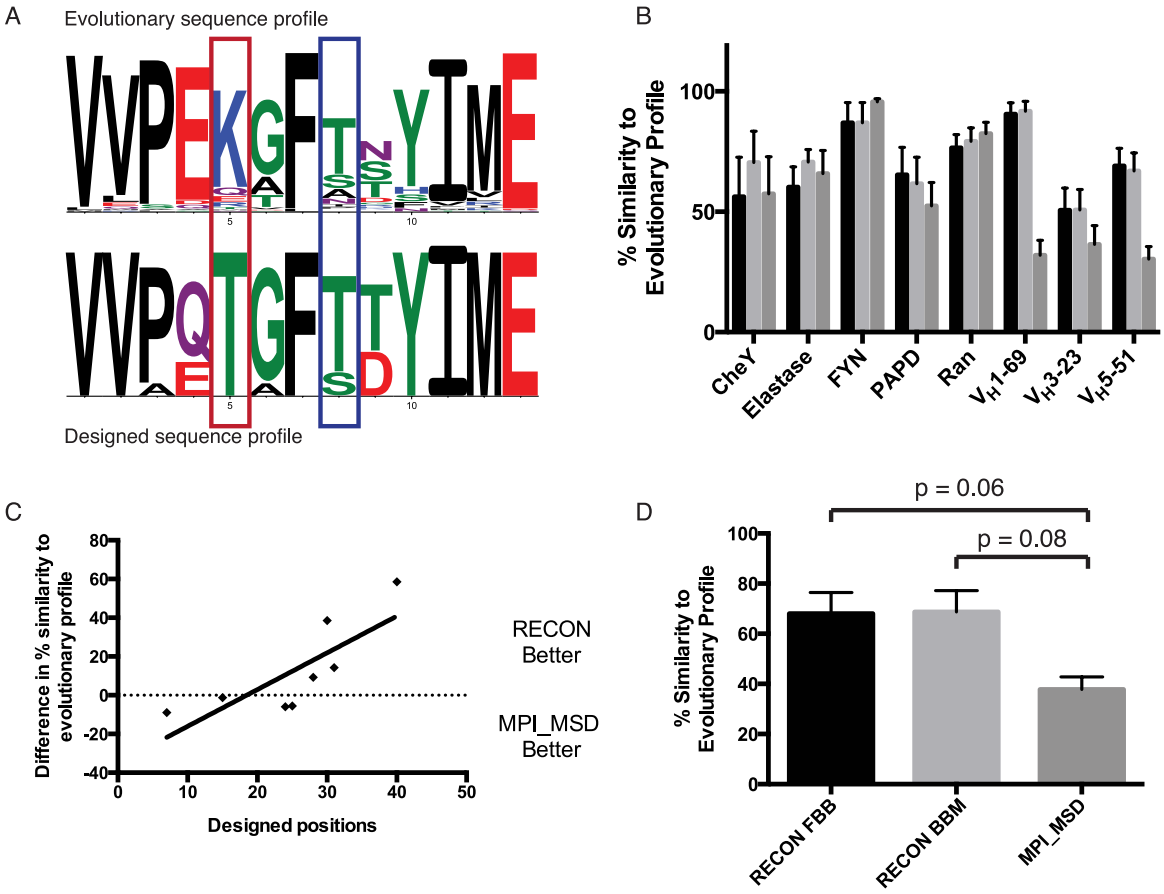


Figure II.5. Recapitulation of evolutionary sequence profiles by multi-specificity design. A. For each protein in the benchmark set, an evolutionary sequence profile (top) was calculated and compared to the sequences generated by MSD (bottom). A similarity score was calculated for each position and averaged over designed positions to measure how well design searches biologically relevant sequence space. Highlighted are example positions where designed sequences either agreed (blue) or disagreed (red) with naturally occurring sequences. The figure displays the designed amino acid profile for a subset of positions in the V_H5-51 benchmark set. See methods for details on percent similarity calculation. Amino acids are colored according to chemical properties. B. RECON-generated designs were more similar to observed evolutionary sequence profiles than those produced by MPI_MSD. Percent similarity was averaged over designed positions that had been mutated by any design method. Plotted are mean and SEM values. Design protocols are colored as in panel D. C. Improvement in recapitulating evolutionary sequence profiles of RECON increases with the number of designed positions. For each benchmark set, the number of designed positions is plotted against the difference in evolutionary sequence similarity between RECON backbone minimized and MPI_MSD. Least-squares linear fit is shown, with an R-value of 0.61 and p value of 0.02. D. Difference in recapitulation of evolutionary sequence profile for the four largest benchmark sets by designs generated by RECON using fixed backbone (FBB) or backbone minimization (BBM) protocols, or MPI MSD. P values were calculated using a paired two-tailed t test.

Structural differences in residues preferred by different algorithms

The algorithms RECON and MPI_MSD feature substantial differences in sequence and structure at many positions of the output design models, particularly in the common germline antibody benchmark sets. We hypothesized that this difference in preference may be due to a failure by MPI_MSD to exhaustively search through sequence space in a large design problem. Concurrently we expect that the sequences selected for by RECON are actually lower in overall fitness. We present structural analysis of three positions, residues 32, 33, and 74 in the V_H3-23 benchmark, to support this claim. Position 32 showed a preference for tyrosine in RECON-generated designs, whereas MPI_MSD prefers glycine. Tyrosine is able to fill a cross-interface gap in the 1S78 complex, and can establish hydrogen bonding to an amide nitrogen across the interface (Figure II.6A). This additional hydrogen bonding produces a large drop in fitness for this residue across all states (-1.85 versus -5.97 REU). Interestingly, tyrosine is the germline residue at this position, and was only recovered using RECON with backbone minimization - both RECON fixed backbone and MPI_MSD favor glycine at this position. Position 33 also showed difference preferences between design methods - alanine was favored by MPI_MSD, whereas RECON favored serine. Serine results in a lower overall fitness due to additional hydrogen bonding with a glutamine residue on the heavy chain CDR3 loop of the antibody (Figure II.6B). At this position, alanine is the germline residue - however the per-residue fitness values indicate that serine is able to stabilize this loop in the 3BN9 complex without compromising stability of the other states (Figure II.6B, fitness shown in parenthesis). Lastly, position 74 showed a preference for threonine in RECON-generated designs, as opposed to serine in MPI_MSD-generated designs. Threonine is able to establish cross-interface hydrogen bonding in the 1S78 complex without causing clashes in other states, whereas serine is somewhat surprisingly not positioned to create this interaction

(Figure II.6C). This is partially due to backbone movements in the RECON-generated structure that position the hydroxyl group for optimal hydrogen bond geometry. In addition to hydrogen

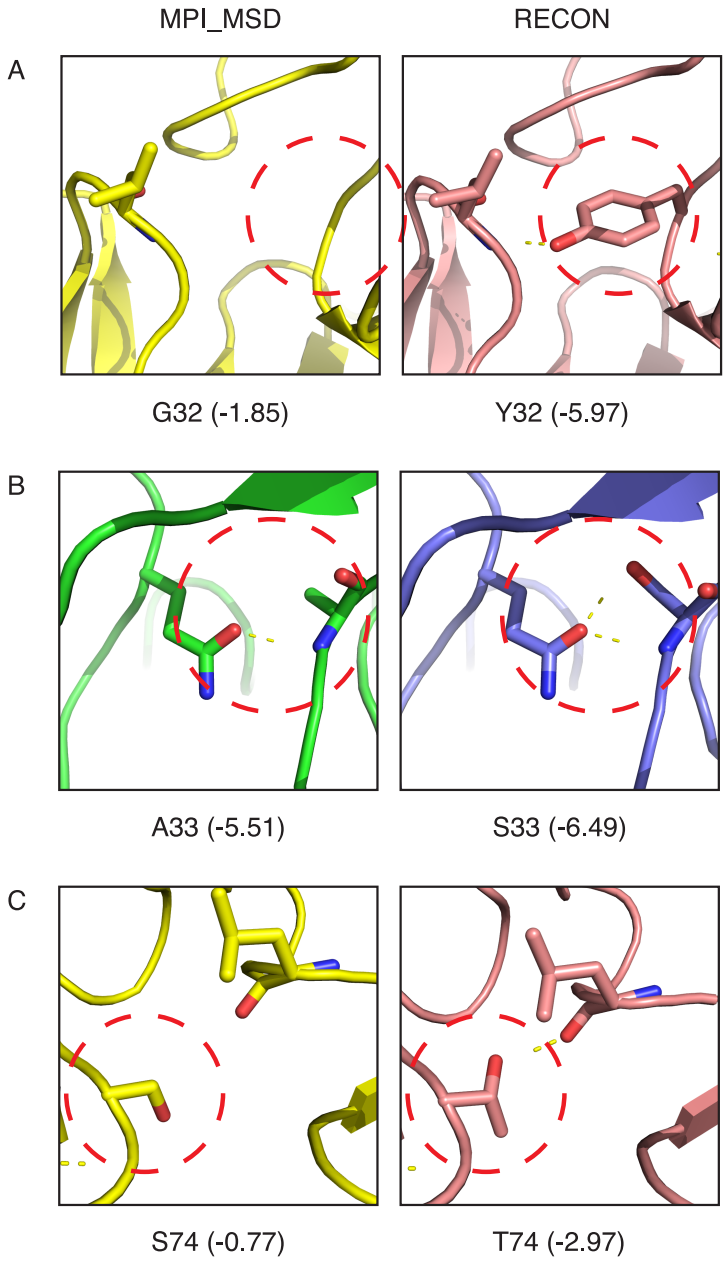


Figure II.6. Structural analysis of sequence preferences of RECON and MPI_MSD. At positions 32 (A), 33 (B), and 74 (C), RECON and MPI_MSD showed consistent difference in sequence preference in the V_H3-23 benchmark. Circled in red are positions that differ between the two structures. Shown in parenthesis are per-residue energy scores in REU summed across all post-minimization states. Shown above are post-minimization structures from designs generated by RECON and MPI_MSD. Structures shown in panels A and C are from the 1S78 complex, and those in panel B are from the 3BN9 complex.

bonding, threonine scores more favorably on the basis of increased van der Waals attractive forces of the additional methyl group with surrounding atoms. At this position, asparagine is the germline amino acid, which was recovered by neither RECON nor MPI_MSD.

Incorporating backbone motion results in increased recapitulation of evolutionary sequence profile

From our initial benchmark results, we did not observe a difference in evolutionary sequence similarity for designs created with fixed backbone versus backbone minimization protocols (Figure II.5D). However, as previous reports have shown the utility of incorporating backbone motion into a design protocol (Harbury et al., 1998; Hu et al., 2007; Humphris and Kortemme, 2008; Mandell and Kortemme, 2009), we hypothesized that the initial minimization of structures before entering them into multi-specificity design reduced the impact of alternating backbone minimization with design. We hypothesize that backbone movement should have a larger impact on design of structures that have not been pre-minimized. To test this hypothesis, we repeated the benchmark with structures that had not been pre-minimized, and performed multi-specificity design with three protocols: 1) fixed backbone design, 2) alternating design with minimization of φ, ψ, and χ angles, and 3) alternating design with backrub movements. The backrub motion involves rotation of a rigid backbone around axes between nearby Cα atoms, and has been shown to recapitulate alternative backbone conformations in high-resolution crystal structures (Davis et al., 2006) as well as improving prediction of the conformation of point mutant side chains (Smith and Kortemme, 2008). We predicted that a design protocol including backrub motions between design rounds should result in the highest agreement to evolutionary sequence profiles, given the sampling of more biologically relevant conformational space than simple minimization. We therefore analyzed the similarity to evolutionary sequence profiles for the top

ten designs produced by the three methods and compared to evaluate whether backbone motion in this context confers any additional benefit. As expected, incorporating backrub movements results in a statistically significant increase in similarity to evolutionary profiles as compared to a fixed backbone protocol or one involving minimization (Figure II.7). This agrees with previous studies indicated that backrub motions are able to sample biologically relevant conformational space, and shows that backrub motions can be incorporated in a multi-specificity context to provide more robust results in terms of evolutionary sequence recovery.

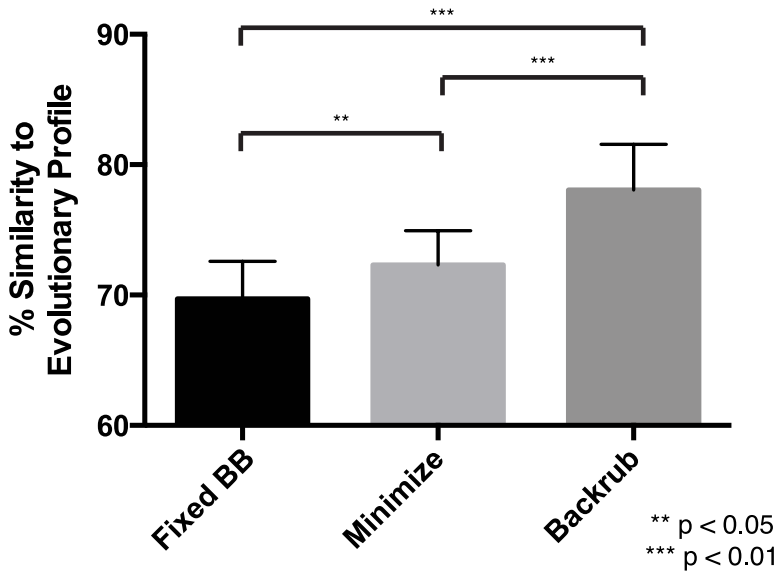


Figure II.7. Incorporation of backbone motion into RECON recapitulates evolutionary sequence profiles in un-minimized structures. Multi-specificity design using RECON was repeated on structures that had not been previously energy minimized to evaluate the benefit of incorporating backbone movements. Designs were generated using either a fixed backbone protocol (Fixed BB), alternating rounds of φ, ψ, and χ angle minimization (Minimize), or using backrub motions (Backrub). P values were generated by a paired two-tailed t test.

Advantage of multi-specificity vs. single-state design

In previous works involving both germline antibodies and promiscuous proteins, the difference in sequence recovery between sequences generated by single-state and multi-specificity design has been analyzed (Humphris and Kortemme, 2007; Willis et al., 2013). Multi-specificity design in both cases was shown to recover the native or germline sequence at a higher rate than single-state design, supporting the proposition that the increased performance of multistate design justifies the increased computational complexity. Given the increased performance of RECON in native sequence recovery, we hypothesized that multi-specificity design performed by RECON would result in a larger difference in germline vs. mature sequence recovery in the germline antibody dataset. We therefore performed fixed backbone single-state design for each complex in this dataset and calculated recovery of the germline sequence and the mature antibody sequences. We can recover the difference in germline and mature sequence recovery as observed in (Willis et al., 2013), and show that design performed by RECON results in a larger difference between germline and mature sequence recovery compared to MPI_MSD (Supplementary Figure II.2). We can therefore conclude that in these cases RECON is more robust at generating germline-like, multi-specific sequences compared to MPI_MSD.

Discussion

Summary of results

We have developed and benchmarked a new method for multi-specificity design, REstrained CONvergence in multi-specificity design (RECON). This algorithm operates by allowing each state to search sequence space independently with a restraint system that gradually encourages convergence between different states on a common sequence. Allowing each state to

adopt a unique sequence reduces the number of sequences required to search in order to find a native-like low energy solution. In two separate benchmark sets consisting of ten total cases, we were able to show that RECON, both with and without iterative backbone minimization cycles, was able to more accurately recapitulate the native, multi-specific sequence of input proteins than the existing MSD application in ROSETTA, MPI_MSD. In addition, we analyzed agreement of designed sequences with observed evolutionary sequence profiles to measure how well MSD simulates natural sequence tolerance. In large design problems with many residues being optimized simultaneously, RECON was able to create sequences that more closely mirrored the natural distribution of sequences seen in evolutionary profiles.

Diversity in predicted sequence tolerance

In this study we analyzed the degree of convergence of a designed protein sequence profiles with the natural sequence variation seen in evolutionary homologs. It is well known that many proteins tolerate a wider variety of sequences than simply the native sequence (Allen et al., 2010; Howell et al., 2014; Humphris and Kortemme, 2008; Smith and Kortemme, 2011) - therefore a major goal of multi-specificity design is to recover not only the native sequence of a protein, but also sequence variations that are tolerated by all binding partners. We found that RECON is able to recover evolutionary sequence profiles more effectively in large complexes - however, it is clear from analysis of sequences sampled by each method (Supplementary Figure II.1) that MPI_MSD is exploring a much larger sequence space. In certain cases, this diversity of sampling may be desired, especially in cases where the interface with both binding partners is compatible with a large number of sequence polymorphisms. Our benchmark cases suggest that sampling near the energy minimum for each individual state is sufficient to recovery the sequence space compatible with all states. However, in cases where generating sequence diversity is at a premium, for example

to explore the tolerated sequence space of a given backbone, it may be advantageous to use RECON and MPI_MSD as complementary approaches.

Comparison of rotamer packing algorithms

RECON in the current study was used with the standard ROSETTA simulated annealing rotamer optimization protocol (Kuhlman and Baker, 2000) – however, other rotamer optimization methods have shown superior performance in certain instances. For example, MPI_MSD uses a modified form of the FASTER algorithm (Allen and Mayo, 2010), referred to as backbone-minimum-energy conformation followed by single-residue perturbation/relaxation (BMEC-sPR) (Leaver-Fay et al., 2011a). Leaver-Fay *et al.* compared the effectiveness of these two algorithms and found that BMEC-sPR consistently reached the global minimum solution in a higher proportion of cases (Leaver-Fay et al., 2011a). Additional rotamer optimization algorithms have been adapted for use in MSD, such as dead-end elimination (Yanover et al., 2007), probabilistic graphical models (Fromer et al., 2010), and iterative batch relaxation/single perturbation and relaxation (Allen and Mayo, 2010). The benefit of RECON is that it can be adapted to work with any single-state-compatible rotamer optimization method, as communication between different states is conducted solely by the restraint system. This opens up the possibility of adapting many more optimization methods for MSD.

Discussion of fixed backbone versus backbone flexibility in restrained multi-state design

One important benefit of RECON is the ability to incorporate backbone motion into an MSD protocol. Traditionally protein flexibility in MSD has been modeled by including multiple backbone conformations as input states (Allen et al., 2010; Ambroggio and Kuhlman, 2006; Davey and Chica, 2014; Howell et al., 2014; Kapp et al., 2012). This is a reasonable strategy for running MSD using RECON. However, RECON offers the benefit that each state can be subject to

additional backbone minimization between design rounds. When incorporating backbone motion into design the conformational and sequence space explodes, making it difficult to reach a global minimum. However the fact that RECON reduces the sampling needed to reach the optimal sequence allows for more search space to be explored. We have shown that incorporating backbone flexibility in the form of backrub motions can improve accuracy of sequences when applied to un-minimized structures. Single-state design protocols have successfully incorporated backbone movement, allowing the introduction of mutations that would have been unfavorable on the original backbone (Harbury et al., 1998; Kuhlman et al., 2003; Mandell and Kortemme, 2009). The ideal protocol for flexible backbone design remains elusive, considering the different methods of backbone perturbation (Davis et al., 2006; Hu et al., 2007). In addition it remains unclear how to best alternate fixed backbone sequence optimization with backbone motion (Fung et al., 2008). RECON opens up the possibility of incorporating these backbone design methods into an MSD context.

Negative design capabilities

One of the most challenging aspects of MSD is the inclusion of unfavorable states to destabilize. The current implementation of RECON is limited in scope compared to approaches such as MPI_MSD due to the inability to perform negative design to disfavor certain states. This limitation is not fundamental as in principle unfavorable states could be designed against with an energy penalty. However, it is outside the scope of the current work to benchmark such an approach. MSD has been successful in engineering proteins when both including (Ashworth et al., 2010; Grigoryan et al., 2009; Havranek and Harbury, 2003; Kortemme et al., 2004) and ignoring (Allen et al., 2010; Ambroggio and Kuhlman, 2006; Kapp et al., 2012) these negative design states. Bolon *et al.* have shown that including negative states produces designs that exhibit better

specificity between competing states (Bolon et al., 2005) – however, this comes at the cost of specificity for the target protein (Bolon et al., 2005; Fromer and Shifman, 2009), and therefore may not be ideal for every design problem. In addition, negative design states result in a significantly more complicated computational protocol – differences between backbone conformations can cause failures in rotamer placement that lead to artificially high energies (Leaver-Fay et al., 2011a). This complicates the inclusion of multiple backbone states in an MSD problem, which mimics the natural flexibility of a protein in solution and results in higher quality designs (Allen et al., 2010; Ambroggio and Kuhlman, 2006; Davey and Chica, 2014; Kapp et al., 2012). Explicit negative design is not currently supported using RECON – the lack of an explicit fitness function makes it difficult to reconcile energies of positive states with negative ones. Grigoryan *et al*. used an intriguing "specificity sweep" protocol that alternates design rounds optimizing stability of positive states with specificity rounds, accepting mutations that destabilize the negative states without a negative effect on the positive ones (Grigoryan et al., 2009). A similar strategy could incorporate RECON to optimize stability and specificity without explicitly designing against a negative state.

Integration of restrained multi-state design into ROSETTA code base

RECON was designed with the intent to be easily integrated into the ROSETTASCRIPTS computational framework (Fleishman et al., 2011a). To this end we emphasize that RECON is compatible with any other protocol that is available within ROSETTASCRIPTS, which is not available for MPI_MSD. This makes it easier for users with experience running SSD protocols in ROSETTASCRIPTS to expand their capabilities by including RECON. This can be used to include additional conformational states, explicitly model bound and unbound conformations, or simultaneously design against multiple partners.

Methods

Selection of databases for benchmarks

Common germline gene-derived antibody complexes were selected and processed as in (Willis et al., 2013) – briefly, candidate complexes were selected by querying the Immunogenetics Information System (IMGT) 3D structural query system for antibodies derived from either V_H1-69, V_H3-23, or V_H5-51 germline genes (Kaas et al., 2004). Only complexes containing protein or peptide ligands were considered. Common germline antibodies were only considered for multi-specificity design when derived from the same allele. Promiscuous proteins used were derived from the multi-specificity design study described in Humphris *et al.* (Humphris and Kortemme, 2007). Complexes were selected to maximize diversity of structure and function, as well as to select proteins with diverse ligands.

Preparation of structures for design simulations

Structures were downloaded from the Protein Data Bank (PDB; www.rcsb.org), and manually processed to remove water and non-proteinogenic molecules. Any chain breaks were closed using kinematic loop closure (Stein and Kortemme, 2013). Due to extensive chain breaks in CDR loops, chains H and L in structure 3GBM were replaced by the same chains in 3GBN. Structures were subject to energy minimization in ROSETTA using the talaris 2013 score function (Leaver-Fay et al., 2011b). The lowest energy model of 50 energy-minimized models for each complex was selected for design.

Multi-specificity design

For common germline antibody multi-specificity design, amino acid sequences deriving from the V_H gene were aligned using ClustalW sequence alignment (Larkin et al., 2007), and positions that varied in any one of the antibodies were specified for design. Germline sequences

were inferred from IMGT/3D Structure DB (Kaas et al., 2004). Designable residues in promiscuous proteins were selected as those present in the interface of all binding partners. To define interface residues, a set of filters was applied to select residues that are likely to engage in interactions with the opposing chain. The first filter eliminates any residue with a Cβ distance larger than 10 Å from the closest residue in the opposing chain. Residues were then selected that either had a heavy atom within 5.5 Å of a heavy atom across the interface, or those with an angle of less than 75° between two vectors, Cα-Cβ of the residue and Cβ-Cβ to the closest residue Cβ on the opposing chain. This vector angle filter allows inclusion of residues where the sidechain is oriented to face the opposing chain. In addition, any residues at the interface on the side of the binding partner were specified for repacking. Identical residues for design and repacking were used for both RECON and MPI_MSD. For RECON benchmarking, fixed backbone design was used with 4 rounds of rotamer packing. RECON constraints were ramped through the 4 rounds of design using convergence restraints of 0.5, 1.0, 1.5, and 2.0 REU. Sequence convergence was enforced at the end of the protocol using a greedy selection algorithm. RECON was also benchmarked with backbone minimization – this protocol was identical to the fixed backbone protocol with the addition of a cycle of minimization of φ, ψ, and χ angles after each design round. At the end of the backbone minimization protocol we performed one full round of a ROSETTA relax protocol, which involves rotamer packing and minimization using a gradually increasing repulsive force (Combs et al., 2013). In designs performed with backrub motions, all backbone atoms on the protein chain being designed were specified as pivot residues - the backrub motion as implemented in ROSETTA is described in detail in (Smith and Kortemme, 2008). MPI_MSD was run with default parameters, with the number of rounds defined as 15 times the number of designable residues (Leaver-Fay et al., 2011a). For MPI_MSD, the fitness function was defined as the sum of energy

of the complexes. Single-state design was run as four rounds of fixed backbone rotamer optimization, using the same designable and repackable residues as previously specified. The talaris 2013 scoring function was used for all methods of design.

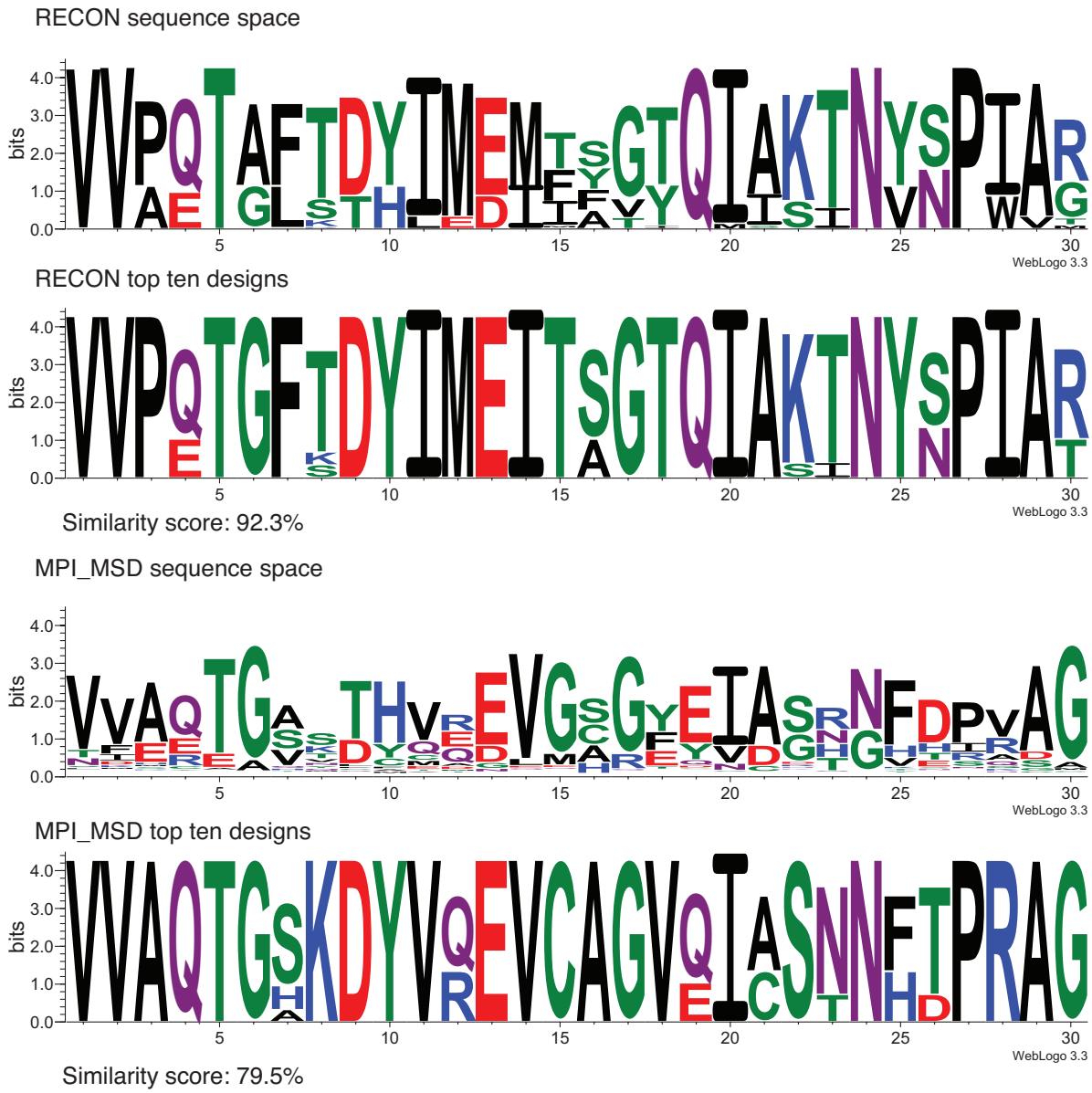
Quantitative measures for analysis of resulting sequences

For each complex, 100 designs were created as described using both RECON and MPI_MSD applications. Sequence logos were created using the Berkeley web logo server (http://weblogo.berkeley.edu). Bitscore was computed for each design trajectory, defined as the Shannon entropy of each amino acid occurring at each designed position, described in (Schneider and Stephens, 1990; Willis et al., 2013). Bitscore was calculated using the following equation: $I_i = p_i \times \log_2(20 \times p_i)$, where i represents the amino acid and p_i is the frequency in the top ten designs. When p_i is 100% the bitscore becomes 4.32, which was used as the maximum possible bitscore in our calculations. To calculate native sequence recovery, the summed bitscore of the native amino acid at each position was divided by the sum of the bitscore of all amino acids at all positions. Designs were analyzed on the basis of the fitness of the top ten designs, with fitness defined as the sum of ROSETTA energy of all states, and native sequence recovery. ROSETTA energy was reported in all cases with convergence restraints subtracted from the total.

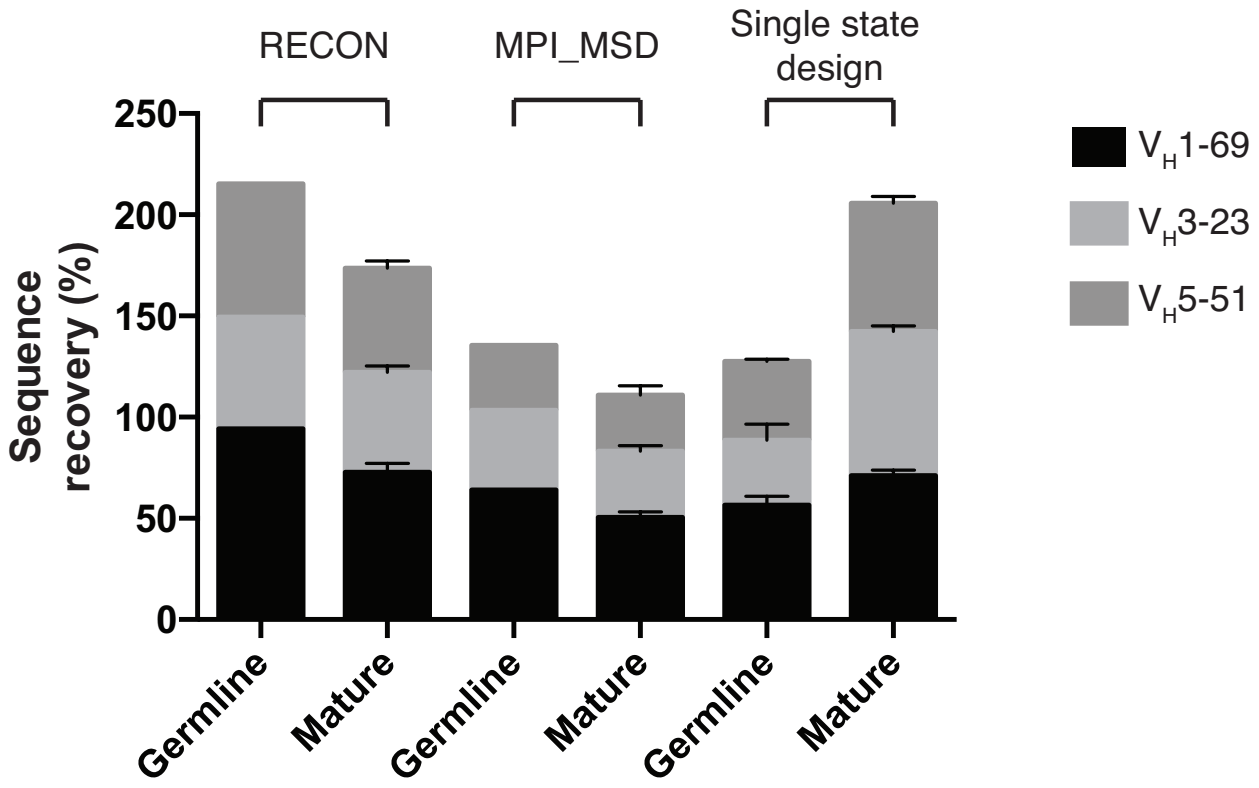
To generate an evolutionary sequence profile we used PSI-Blast with default parameters, querying a non-redundant protein database (Altschul et al., 1997; Pruitt et al., 2005). This position-specific scoring matrix (PSSM) of amino acid frequencies was then compared to a PSSM constructed from observed frequencies in the top ten designs by fitness resulting from RECON or MPI_MSD. To compare PSSMs we used a modified Sandelin-Wasserman similarity score (Sandelin and Wasserman, 2004). This score was calculated by computing the squared difference for each amino acid frequency at each position. The squared difference was summed for all amino

acids at a given position(Willis et al., 2013) and subtracted from two to yield a similarity score from 0 (no similarity) to 2 (identical). This value was then normalized by a factor of two to yield a percent similarity for each position and summed over all designed positions to give an overall similarity score. To reduce background noise when comparing PSSMs we only compared positions that had any observed mutations in the top ten designs produced by any design method. Inclusion of positions where no mutations are observed would inflate evolutionary similarity values for all methods. This reduced the total number of positions considered from 200 among eight benchmark sets to 97. In the benchmark cases of un-minimized structures, eliminating positions with no variation by any method left 151 of 200 possible positions.

Supplemental Information



Supplementary Figure II.1. Sequence space explored by RECON and MPI_MSD. The sequence space explored by RECON and MPI_MSD is compared to the sequence profiles of the top ten designs for the V_H5-51 benchmark case. Sequence space was determined as any amino acid that was sampled at any point throughout the design protocol. A similarity score was calculated between the sequence space explored by an algorithm and the top ten designs produced by the same algorithm.



Supplementary Figure II.2. Germline and mature sequence recovery from multi-specificity and single-state design. Sequence recovery compared to mature and germline sequences is compared for designs generated by RECON and MPI_MSD multi-specificity design, compared to those generated by single-state design.

Supplementary Table II.1. Post-minimization fitnesses of benchmark sets.

	Post-minimization fitness (REU)		
Protein/Germline gene	**RECON FBB**	**RECON BBM**	**MPI_MSD**
CheY	-1113.5	**-1119.7**	-1119.2
CR6261	-2532.6	**-2537.7**	-2532.0
Elastase	-1445.4	-1445.1	**-1447.9**
FI6v3	-2506.0	**-2515.2**	-2506.2
FYN	-777.2	**-780.3**	-778.3
PapD	-1903.5	-1891.4	**-1908.8**
Ran	-3675.2	-3716.4	**-3755.8**
V_H1-69	-5299.1	-5306.7	**-5343.5**
V_H3-23	-3410.0	-3427.1	**-3479.9**
V_H5-51	-2329.4	-2348.3	**-2360.5**

Structures generated by design were energy minimized to relieve small clashes. Fitnesses reported are the sum of energy of all states. Best values in each row are shown in bold.

Supplementary Table II.2. Performance of a control greedy selection algorithm.

	Native sequence recovery (%)		Fitness (REU)	
Protein/Germline gene	**RECON FBB**	**Greedy selection**	**RECON FBB**	**Greedy selection**
CheY	80.6	60.0	-1093.1	-825.1
CR6261	79.0	69.2	-2499.5	-1531.7
Elastase	84.8	80.0	-1383.8	-767.5
FI6v3	57.8	44.1	-2459.1	-1912.4
FYN	100.0	100.0	-758.3	-712.3
PapD	92.5	65.7	-1685.5	-1241.4
Ran	87.1	73.9	-2682.3	-2637.4
V_H1-69	94.2	59.0	-3015.9	-2314.6
V_H3-23	55.3	27.2	-911.7	939.2
V_H5-51	65.9	40.1	-840.5	50.6
Average	**79.7**	**61.9**	**-1733.0**	**-1095.3**

Design of benchmark cases was repeated for a greedy selection algorithm, which lacks the ramping convergence restraints of RECON. This algorithm performs a single round of unrestrained design followed by a greedy selection of amino acids that maximize fitness over all states.

Supplementary Table II.3. Non-converging positions in the V_H5-51 benchmark set.

Position	Converging count	Non-converging count
2	100	0
5	100	0
14	0	100
16	100	0
23	100	0
24	0	100
29	0	100
30	100	0
31	100	0
32	52	48
34	100	0
40	100	0
46	100	0
48	0	100
51	1	99
52	0	100
54	100	0
58	1	99
65	100	0
70	100	0
72	100	0
74	100	0
76	96	4
77	100	0
80	0	100
84	100	0
88	100	0
93	100	0
97	100	0
98	0	100
Number of positions	**21**	**9**
Germline sequence recovery (%)	**66.1**	**74.0**

Failure to converge by the end of the RECON convergence restraint protocol was counted for each designed residue in the V_H5-51 benchmark set, over 100 design trajectories.

CHAPTER III.

Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses

Sevy, A. M., Wu N.C., Gilchuk I.M., Parrish E.H., Burger S., Yousif D., Nagel M.B.M., Schey K.L., Wilson I.A., Crowe J.E. Jr., Meiler J. Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses. Manuscript submitted.

Author contributions: I wrote the algorithm described in this chapter and ran all computational experiments, under the mentorship of Jens Meiler and James Crowe. I also created genes for expression of all mutant antibodies, performed binding experiments, and measured thermostability. I was responsible for experimental design, analyzed data with my co-mentors, and created all figures in this chapter. I collaborated with N.C.W. and I.A.W. for crystallography and I.M.G. for hemagglutination inhibition assays. E.H.P. and D.Y. assisted in protein expression and purification. S.B. assisted with homology modeling experiments. M.B.M.N. and K.L.S. performed hydrogen-deuterium exchange experiments.

Abstract

Influenza is a yearly threat to global public health. Rapid changes in influenza surface proteins resulting from antigenic drift and shift events make it difficult to readily identify antibodies with broadly neutralizing activity against different influenza subtypes with high frequency, specifically antibodies targeting the receptor binding domain on influenza

hemagglutinin (HA) protein. We developed a new computational design method that is able to optimize an antibody for recognition of large panels of antigens. To demonstrate the utility of this multistate design method, we used it to redesign an anti-influenza antibody against a large panel of over 500 seasonal hemagglutinin antigens of the H1 subtype. As a proof of concept, we tested this method on a variety of known anti-influenza antibodies and identified those which could be improved computationally. We generated redesigned variants of antibody C05 to the HA receptor binding site and experimentally characterized variants that exhibited improved breadth and affinity against our panel. C05 mutants exhibited improved affinity across the entire H1 subtype HA panel by stabilizing the CDRH3 loop and creating favorable electrostatic interactions with the antigen. These mutants possess increased breadth and affinity of binding without sacrificing potency against existing targets, surpassing a major limitation up to this point.

Introduction

Influenza is a yearly threat to global public health. As many as 56,000 deaths and 710,000 hospitalizations annually can be attributed to influenza infection in the U.S. (Rolfes et al., 2016). In addition, commonly used influenza therapeutics have been only modestly effective (Jefferson et al., 2014), and vaccine efficacy has been variable depending on the year (Belongia et al., 2016). The inability to formulate a consistently effective vaccine has been a major hindrance to developing sustained, effective influenza immunity on the population level.

A major factor that limits influenza vaccine efficacy is the fact that pre-existing antibodies frequently lack the ability to react with current circulating strains. Of particular interest are antibodies that target the receptor-binding site (RBS) of the HA protein, the site at which the viral protein interacts with the host cell receptor, sialic acid. These antibodies typically neutralize virus

very potently (Ekiert et al., 2012; Krause et al., 2012; 2011; Lee et al., 2014; Schmidt et al., 2015; Whittle et al., 2011), as they directly inhibit binding of virus to the host cell receptor, and are very prevalent in the immune response (Schmidt et al., 2015). However, as the region of HA around the RBS is highly variable, antibodies to this domain tend to have restricted specificity to strains within a single subtype (Krause et al., 2011; 2012; Lee et al., 2014; Whittle et al., 2011).

Recently, many antibodies have been described that mimic the chemical interactions of the sialic acid receptor with HA (Ekiert et al., 2012; Krause and Crowe, 2014; Krause et al., 2011; 2012; Lee et al., 2014; Schmidt et al., 2015; Whittle et al., 2011). These antibodies tend to be broader in their recognition of influenza than others targeting the RBS since they primarily interact with conserved residues required for viral infectivity. The existence of such antibodies has suggested that broad, RBS-specific antibodies elicited by vaccination may be sufficient to protect against future strains, and could become one of the primary components of a proposed "universal flu vaccine" (Krause and Crowe, 2014; Wu and Wilson, 2017). One such antibody, C05, has remarkable breadth of recognition of HAs from certain strains within both group 1 and group 2 viruses, and interacts with the HA molecule using a single antibody hypervariable loop (Ekiert et al., 2011). However, this antibody still has incomplete breadth against HAs within a particular subtype – for example, it is unable to recognize H1 strains circulating in humans after the 2009 H1N1 pandemic, primarily due to a lysine insertion at position 133a (Ekiert et al., 2012; Wu et al., 2017).

Given the limitations of naturally occurring human antibodies, we sought in this study to use computational design to increase the breadth of existing anti-influenza antibodies. Computational design has been successful in redesigning a single antibody-antigen interaction (Lapidoth et al., 2015; Lippow et al., 2007; Willis et al., 2015); however, until recently, it has been

challenging to include multiple antibody-antigen interactions in a single design simulation to optimize the antibody sequence for recognition of multiple antigens simultaneously. We developed a method that significantly improves the computational efficiency of such multi-specificity design (Sevy et al., 2015). To further improve the utility of this method, we re-configured the method to run in parallel on multiple computing nodes, enabling much larger scale simulations, and validated this method on redesign of anti-influenza antibodies.

As a proof of principle of the utility of this computational method, we applied the method to the redesign of existing human antibodies against viruses of the influenza A H1 subtype. We expressed and tested a panel of computationally generated variants of antibody C05 and identified mutant antibodies that bound one influenza strain not recognized by C05 and increased affinity against a strain that is recognized with low affinity by C05.

Results

Experimental workflow

We sought to use RECON multistate design to increase the breadth of certain anti-influenza monoclonal antibodies. The RECON multistate design method was written originally to process states serially, which limited both the number of states that could be included and the number of designed residues in each state (Sevy et al., 2015). To address this limitation, we refactored the RECON algorithm to run in parallel, by handling each state on a separate processor and implementing Message Passing Interface (MPI) communication between the different processors (Supplementary Figure III.1). The improved parallel RECON protocol therefore is able to handle much larger ensembles of input states. We decided to test application of the parallel RECON protocol on redesign of influenza antibodies against a set of seasonal virus variants (Figure III.1).

As a proof of principle for this method, we computationally redesigned existing antibodies against an antigenic panel, then expressed and tested antibody variants for improvement in both breadth and affinity across the panel.

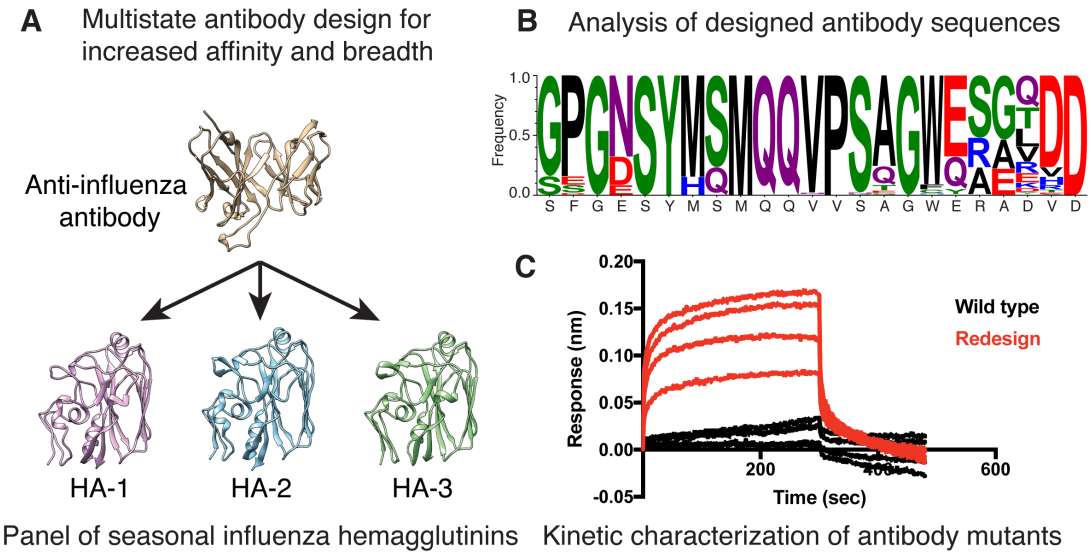


Figure III.1. Experimental workflow of multistate design experiment. Influenza antibodies were modeled against a panel of seasonal influenza hemagglutinin (HA) targets and designed for affinity and breadth (A). The optimized sequences for each antibody were analyzed (B), and mutants with favorable properties were expressed and the binding kinetics were measured using biolayer interferometry (C). Shown are binding kinetics to the HA of the A/Puerto Rico/8/1934 strain.

Benchmark of large-scale design

To test the utility of the parallel RECON protocol, we first sought to redesign an anti-influenza antibody for increased breadth of binding to diverse hemagglutinin (HA) antigens from a large panel of seasonal influenza virus strains. We created homology models of HA proteins from the sequences of 524 viruses in the Influenza Research Database (FluDB) using the RosettaCM multi-template comparative modeling protocol (Bender et al., 2016; Song et al., 2013; Zhang et al., 2017). We paired each of the 524 modeled viral proteins with antibody C05 (Ekiert et al., 2012) and redesigned the antibody sequence for broad recognition of antigens in the viral

panel. We successfully ran multistate design against this large seasonal virus HA panel overnight on a computing cluster, running 50 independent design simulations in 13.2 hours, distributed over 524 processors (Supplementary Figure III.2). The design simulations scaled well with additional states, and the only limitation on number of states was the number of available processors on our computing cluster. The designed models showed significant variability in sequence, specifically in the antibody CDRH3 region, suggesting that this antibody could be improved for breadth of binding to diverse HA antigens in a larger panel.

Design of H1 subtype breadth in influenza HA antibodies

Next, we sought to design antibodies with increased breadth among the H1 subtype of influenza. We first identified all H1 subtype HA proteins with crystal structures in the Protein Data Bank (PDB) with a resolution better than 3.5 Å (Cho et al., 2013; DuBois et al., 2011; Gamblin et al., 2004; Lin et al., 2009; Liu et al., 2009; Schmidt et al., 2013; Xu et al., 2011; 2010; 2012; Yang et al., 2010; 2014b) (Supplementary Table III.1). This search yielded 13 unique antigens to include in the design panel. Next, we identified seven antibodies that are known to bind at least one H1 HA protein in the panel, and that have high-resolution (better than 3.5 Å) co-crystal structures in the PDB (Ekiert et al., 2012; Hong et al., 2013; Lee et al., 2014; Schmidt et al., 2013; 2015; Whittle et al., 2011). We then created complexes of each antibody with all 13 viral proteins in the panel and ran RECON multistate design to generate antibody variants with increased breadth among the panel (Figure III.2).

Some antibodies, such as mAb 5J8, showed a modest improvement across all targets in the panel, but not a drastic improvement for any target (Figure III.2A). Other antibodies, such as mAbs CH65, CH67, and 641 I9, showed a strong energetic improvement for some targets, with a deleterious effect on other targets. These designs were considered unsuitable for testing due to the

fact that they were predicted to have a decrease in breadth, albeit with an increase in affinity for certain targets. The antibody in this study with the strongest improvement for several targets

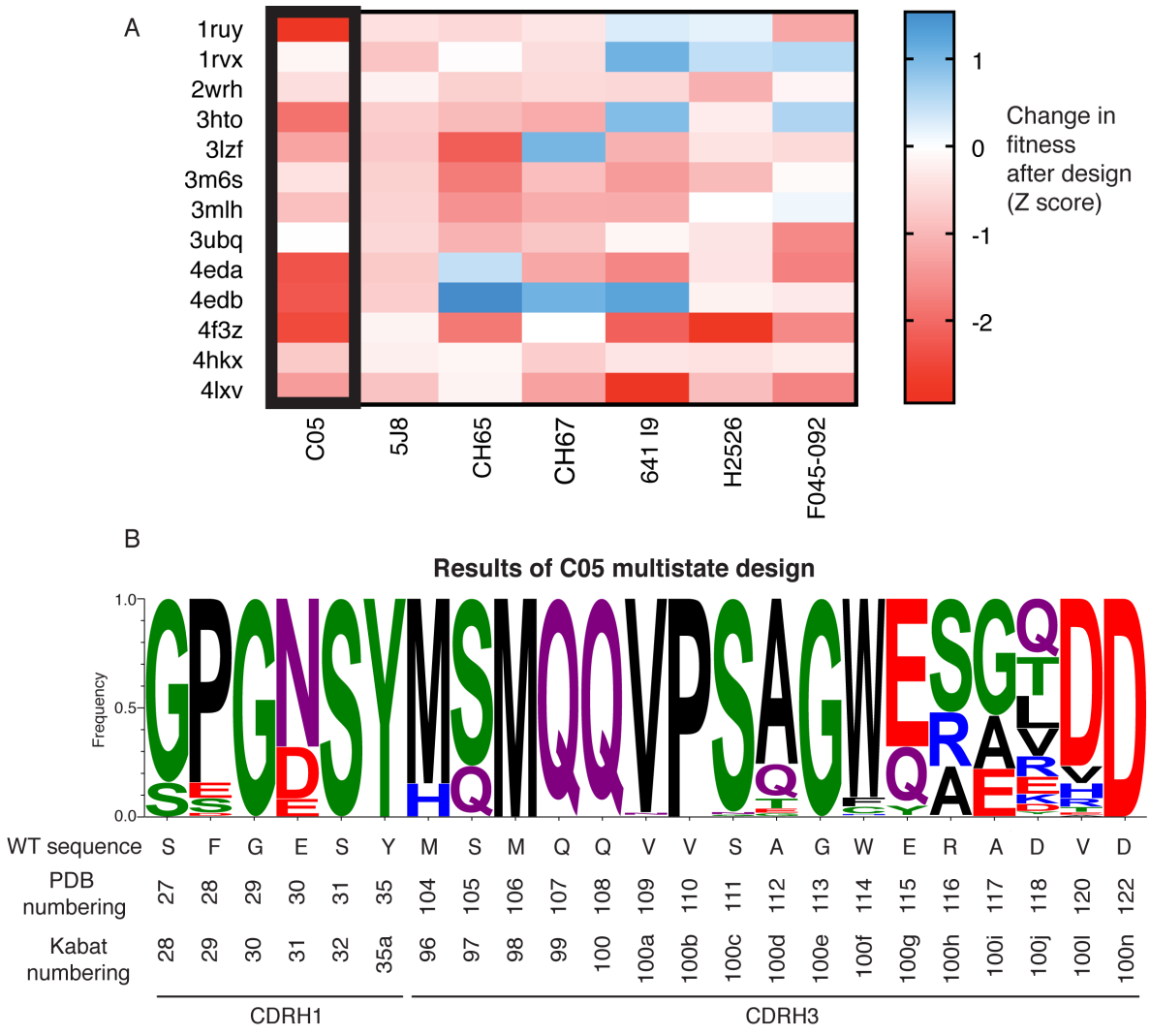


Figure III.2. Fitness and optimized sequences of influenza antibody multistate designs. A. Seven anti-influenza antibodies (x-axis) were designed against the 13 H1 targets in the panel (right legend). 100 designs were generated and the change in fitness from the wild-type to the best design is shown in a heat map. For each design, the fitness was calculated as a normalized sum of the ROSETTA score of the antibody-antigen complex and binding energy and expressed as a Z score. B. The optimized sequences from multistate design of antibody C05 are shown as a sequence logo. Amino acids are colored based on chemical properties. The sequence of wild-type C05 is shown below.

without sacrificing affinity for others was C05. We decided to validate C05 variants experimentally in order to test the effects of multistate design mutations on affinity and breadth.

Experimental validation of C05 mutants

We next sought to validate the predicted increases in breadth and affinity of binding for C05 variant antibodies. We observed many suggested mutations in the CDRH1 and CDRH3 loops of C05 (Figure III.2B). The majority of these mutations were focused in the distal end of the CDRH3 loop, which is in close contact with the antigen. We modeled the effects of each suggested mutation as a single or double amino-acid substitution and measured the effect on the energy of the antibody-antigen complex (Supplementary Figure III.3). Of the mutations introduced, only a small number appeared to contribute the majority of the energetic improvement. We focused our subsequent experimental efforts on mutations that were predicted to have the largest impact on energy of binding. Out of 71 single or double amino acid mutants introduced by multistate design, 27 passed a quantitative and qualitative evaluation for experimental characterization (Supplementary Table III.2). The quantitative filter allowed mutations with an improvement in fitness of greater than 0.5 standard deviations. The qualitative filter consisted of visual inspection and accounting for known pathologies in the ROSETTA score function.

We next synthesized a group of cDNAs for antibody variable genes encoding 33 C05 variant antibodies, comprising 27 single or double amino acid mutants that passed the previously discussed filters and 6 combinations of mutations that were predicted to result in the greatest improvement in stability and binding affinity (Supplementary Table III.2). We expressed and purified the variant antibodies as IgG molecules and measured their activity and binding kinetics using the FortéBio Octet system.

We observed two mutants that exhibited increased affinity and breadth across the panel

(Figure III.3). The majority of the effect on affinity was focused on strains that were recognized by C05 with low affinity, namely the avian strain A/mallard/Alberta/35/1976 and the human strain A/Puerto Rico/8/1934. Full binding data of C05 mutants over all strains tested in this study are shown in Figure III.3 and Supplementary Figure III.4. Mutations V110P and A117E in the CDRH3 loop both increased affinity for A/mallard/Alberta/35/1976 by roughly 4- and 3-fold respectively, with the combination of both mutations improving affinity by roughly a factor of 4-5. Interestingly, the single mutations each increased the on-rate of the antibody-antigen interaction, with a slight decrease in off-rate. The double mutant showed a great increase in on-rate with a concomitant increase in off-rate which limited the total effect on affinity. The V110P mutation also increased breadth by facilitating binding to a new strain that was not recognized by wild-type C05, A/Puerto Rico/8/1934. C05 V110P recognized A/Puerto Rico/8/1934 with a modest but observable affinity

Binding to A/PuertoRico/8/1934

	K_D (μM)	K_D fold change	K_{on} (1/Ms) x 10^4	K_{off} (1/s)x10^{-1}
WT	ND		ND	ND
V110P	42	>4.8	1.2	5.1
A117E	ND		ND	ND
V110P-A117E	ND		ND	ND

ND: binding not detected
WT K_D is estimated to be >200 μM

Binding to A/mallard/Alberta/35/1976

	K_D (nM)	K_D fold change	K_{on} (1/Ms) x 10^3	K_{off} (1/s) x 10^{-3}
WT	511		4.1	2.1
V110P	120	4.3	14.5	1.7
A117E	199	2.6	5.1	1.0
V110P-A117E	106	4.8	202	21.5

Figure III.3. C05 mutants show increased affinity against low affinity strains. Binding kinetics were measured on a FortéBio Octet Red system with four dilutions of antibody. Data were fit to a 2:1 binding model.

in the μM range, whereas the wild-type antibody did not show any binding activity, even when tested at high concentrations. We estimate that this mutation contributed to an increase of affinity by a factor of at least 4.8 for this antigen. These mutants were also tested for binding to the remaining members of the panel; however, apart from the two previously discussed strains and two high affinity strains (A/Solomon Islands/03/2006 and A/Thailand/CU44/2006), no binding was observed for either the wild-type or variant antibodies (data not shown).

Therefore, C05 variant antibodies possessed increased affinity for two weakly bound strains; however, we were interested in whether these variant antibodies lost affinity for strains that were recognized previously. Several groups have reported a tradeoff between affinity and breadth, in which mutated antibodies that have gained affinity for several targets lose affinity for other targets (Babor and Kortemme, 2009; Willis et al., 2013; Wu et al., 2017). This pattern has been observed for antibody C05 in experiments designed to improve affinity for H1 and H3 viruses (Wu et al., 2017). We observed that the mutants in this study maintained high affinity for strains in the panel that were previously recognized by C05 (Figure III.4 and Supplementary Figure III.4). We compared the binding activity of the mutant from this study, V110P-A117E, to an experimentally derived mutant from a study done by Wu *et al*., referred to as VVSSGW (Figure III.4) (Wu et al., 2017). While the VVSSGW variant possessed increased H1 affinity by a greater magnitude, this improvement came at the cost of H3 affinity, which was reduced. The V110P-A117E mutation increased affinity by a more moderate factor, but did not reduce affinity for either of these two strains (Figure III.4). In addition, we compared the binding activity to a panel of strains of different subtypes (Supplementary Figure III.4). In general, the V110P-A117E mutation is able to maintain high affinity binding to these heterosubtypic strains. Notably, all strains which were accounted for in the computational panel displayed an increase or no effect on binding. In

the case of the H2N2 strain A/Japan/305+/1957, we do see a roughly 2 fold decrease in binding; however, it is worthwhile to note that the VVSSGW variant has no detectable binding to this strain (Wu et al., 2017). Therefore, the computational approach appears to have an advantage in preserving high affinity binding across a panel, at least in this analysis.

To investigate whether the newly observed binding activity translated to an increase in biological activity of this antibody variant, we performed a hemagglutination inhibition (HAI) assay with A/Puerto Rico/8/1934 and A/Solomon Islands/03/2006 viruses (Table III.1). Surprisingly, both the wild-type and redesigned antibody were highly potent in inhibiting A/Puerto Rico/8/1934 virus, despite undetectable affinity for the wild-type antibody and low affinity for the variant. This finding is consistent with previously reported data on antibody C05, where neutralizing activity was observed even for viruses with undetectable binding (Ekiert et al., 2012).

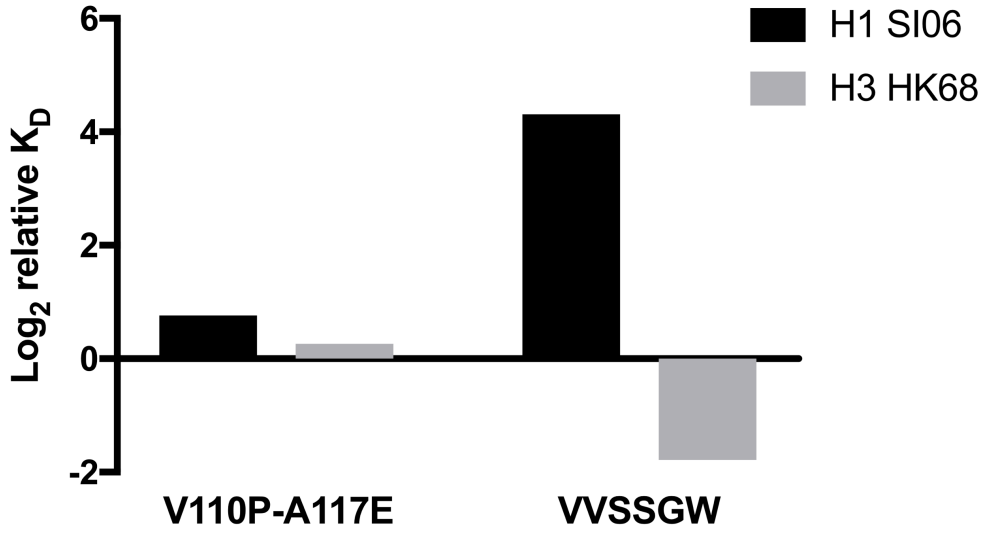


Figure III.4. C05 double mutant does not lose affinity for high affinity strains. Affinity is shown for two high affinity strains, A/Solomon Islands/03/2006 (H1 SI06) and A/Hong Kong/1/68 (H3 HK68). Affinities are compared to an experimentally derived mutant from Wu et al., referred to as VVSSGW. Relative K_D was determined by ELISA for V110P-A117E, and by Octet for VVSSGW.

Table III.1. Hemagglutination inhibition activity of both wild-type and mutant C05.

Virus	C05 WT	C05 V110P
A/Puerto Rico/8/1934	1.5	0.8
A/Solomon Islands/03/2006	0.2	0.4

Shown is the end titer point in μg/mL.

Structural characterization of a redesigned C05 variant

We next sought to confirm the accuracy of our models of the C05 double mutant V110P-A117E. Therefore, we determined the crystal structure of the HA1 subunit from A/Hong Kong/1/1968 (H3N2) in complex with the V110P-A117E C05 variant at a resolution of 3.25 Å (Supplementary Table III.3). Four antibody-antigen complexes were observed in the asymmetric unit. Overall the CDRH3 loop was predicted well by the ROSETTA models, with an RMSD of 1.09 Å over all atoms and 0.43 Å over Cα atoms. The mutation V110P points towards the framework of the antibody and has few contacts with the antigen, similar to the positioning of the wild-type valine (Figure III.5A). This residue has a φ angle of -57.9° in the mutant structure and -61.5° in the wild-type structure, both of which agree with the preferred φ of proline of -65° that limits its conformational freedom. This explains why a proline at this turn in the CDRH3 loop stabilizes the active conformation (Morris et al., 1992). This finding is consistent with observations made by Wu *et al.* in their study of *in vitro* C05 mutagenesis (Wu et al., 2017). We predicted that the mutation A117E improves electrostatic interactions between the antibody and the antigen, interacting with a lysine at position 125a (H3 numbering) of the antigen (Figure III.5B). The crystal structure was obtained in complex with an HA protein (A/Hong Kong/1/1968) that does not have a lysine at this position (Figure III.5A), so the presence of this interaction could not be confirmed. However, E117 appears to be in position to make the electrostatic contact and occupy a similar space as in the model (Figure III.5). This hypothesized mechanism is also consistent with the observation that A117E is not universally favorable for all antigens – it confers an increase in

affinity for A/mallard/Alberta/35/1976 with a slight decrease in affinity for A/Puerto Rico/8/1934 (Figure III.3), as evidenced by the fact that the double mutant is unable to recognize A/Puerto Rico/8/1934 whereas V110P can.

As a complementary approach, we characterized the interaction of C05 V110P-A117E with the head domain of A/Solomon Islands/03/2006 HA using hydrogen-deuterium exchange mass spectrometry. We mapped the perturbation of hydrogen-deuterium exchange upon antibody-antigen binding on both the epitope and paratope (Supplementary Figure III.5). We observed peptides originating from the CDRH3 loop to be most solvent occluded in the antibody-antigen complex, most notably at short time points (<1 min). This is in accordance with the predicted binding mode from the ROSETTA models. In addition, the epitope peptides shielded upon binding are located along the rim of the receptor-binding domain, which agrees with the models and crystal structure.

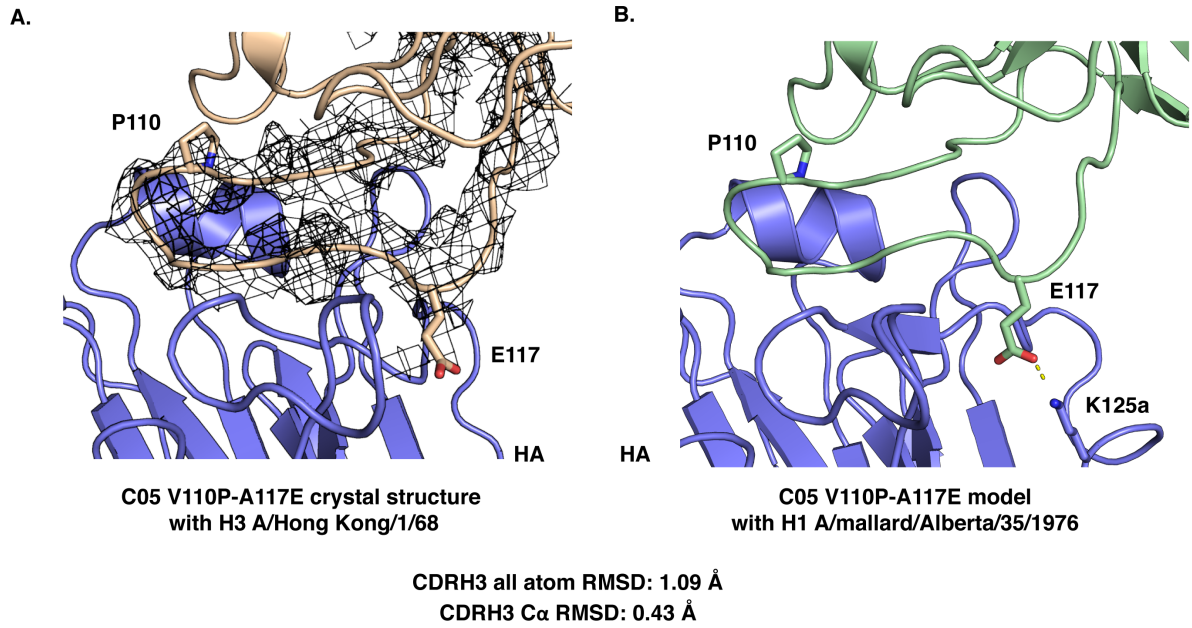


Figure III.5. Crystal structure of the C05 V110P-A117E double mutant in complex with A/Hong Kong/1/68 head domain confirms the accuracy of the computational models. A. Structure of V110P-A117E is shown in complex with A/Hong Kong/1/68, with the 2Fo-Fc electron density contoured at 1.0 σ. B. Model of C05 V110P-A117E in complex with A/mallard/Alberta/35/1976, with predicted hydrogen bonding shown in dashed lines. RMSDs in Å over all atoms and Cα atoms are shown below.

To test the effect of mutations on the thermodynamic stability of the antibody, we measured the melting temperature of variants using differential scanning fluorimetry (DSF). The variants mostly exhibited two melting transitions, one at approximately 62 °C and another at approximately 69 °C, corresponding to the Fc and Fab domains, respectively (Supplementary Figure III.6). This finding agrees with previous data on IgG melting transitions (Ionescu et al., 2008; Vermeer and Norde, 2000). To confirm these domain assignments, we repeated the experiment with a cleaved Fab protein and observed the transition at ~70 °C (Supplementary Table III.4). As predicted, mutation V110P did increase the antibody stability by roughly 0.5 °C, although addition of A117E reduced stability by 0.6 °C (Table III.2). Notably, several other mutations also increased the melting temperature by a significant margin, including mutations that had a neutral or negative impact on binding affinity (Table III.2). Previous studies suggested that an increase of 1 °C is sufficient to increase affinity 10-fold (Willis et al., 2015). However, these results suggest that an increase in stability does not necessarily confer an increase in binding affinity.

Table III.2. Thermodynamic stability of C05 mutants as measured by differential scanning fluorimetry (DSF).

Variant	**Transition 2 (°C)**	**Change from WT (°C)**	**Significance**
C05 WT	69.8		
S27G F28P V110P	70.8	1.0	**
F28E	70.4	0.6	*
Y35H V110P	70.8	0.9	**
V110P	70.3	0.5	*
V110P A117E	69.2	-0.6	*
A117E	69.2	-0.6	*
D118R	70.6	0.8	**
D120R	70.3	0.5	*
6 aa mutant	71.9	2.1	**

Statistical significance was assessed using a two-tailed T test. **p<0.005, *p<0.05

Tradeoff in breadth and affinity

A common theme in design of antibody breadth is a tradeoff between breadth and affinity. This relationship has been shown both theoretically (Babor and Kortemme, 2009; Willis et al., 2013) and in practice (Wu et al., 2017), and is a major motivation for computational methods that can account for hundreds of antigens during design, such as the RECON algorithm. To test if an antibody must sacrifice affinity for an individual target in order to acquire breadth, we repeated design of C05 against each of the antigens in the panel, using single-state design against each of the targets individually, instead of RECON multistate design. The results showed a tradeoff between breadth and affinity, as the single-state designed antibody was consistently better against each target than the multistate solution (Supplementary Figure III.7). For each of the antigens, the redesigned C05 for single-state optimized binding had lower total score than the optimal multistate C05, and 10 out of 13 had lower predicted binding affinity for single-state design than multistate design. This finding supports the idea that multistate design achieves a compromise between the optimal sequence for each individual target.

Discussion

Summary of results

In this study, we report a novel protocol for multistate design of large, parallelized panels of influenza virus strains. We adapted a previously reported protocol for multistate design to run in parallel on a computing cluster and showed that this protocol can scale to very large (>500) panels of antigens. As a proof of principle, we applied this methodology to designing anti-influenza antibodies for increased breadth against panels of seasonal H1 viruses. We report redesigned antibodies that have roughly a 5-fold increased affinity against one strain and now

detectable binding to another strain in the panel, without sacrificing affinity for any members of the panel.

Large-scale panels in multistate design

Multistate design has been successful in a number of different applications; however, it is generally applied to modulating specificity in protein-protein binding partners (Grigoryan et al., 2009; Havranek and Harbury, 2003; Lewis et al., 2014), or modeling conformational ensembles of a single protein (Davey and Chica, 2014). We instead focused here on design of an antibody against a large ensemble of targets. In typical computational antibody design approaches, a single antigen or a small panel of representative antigens is modeled and assumed to represent the scope of antigenic variability (Fleishman et al., 2011b; Willis et al., 2015). However, using the protocol reported here, it should be possible to include a much larger panel of targets, easily making an antibody robust to antigenic variation. In this work, the affinity increases that we report (~5-fold) are modest compared to the increases reported in other reports that use experimental approaches (Wu et al., 2017) or computational approaches with single antigens (Willis et al., 2015). We expect that, as the size of the target panel increases, it will become increasingly difficult to find mutations that can improve affinity for some targets in the panel without sacrificing affinity for any other targets. Therefore, a modest increase in affinity may be more realistic when designing against large and diverse antigenic panels, especially when considering already high affinity complexes. However, the advantage of this approach is that the affinity-enhancing mutations can be selected to be compatible with all targets, which is often not the case with experimental approaches that do not account for multiple states (Wu et al., 2017).

Mechanisms of action

We hypothesize that the V110P mutant reported in this work enhances affinity by increasing the thermodynamic stability of the antibody. It is common that affinity-enhancing mutations are located in positions that do not directly contact antigen and function by increasing antibody stability and rigidifying CDR loops (Wang et al., 2013; Xu et al., 2015). This phenomenon is seen in *in silico* engineered antibodies, mutations introduced by directed evolution, and naturally occurring somatic mutations from mature antibodies. We also generated several mutants that increased thermodynamic stability but failed to improve binding affinity, showing that increased stability is not sufficient by itself for increase of binding to a target. A mutant with increased affinity but decreased thermodynamic stability, A117E, is predicted to do so by establishing electrostatic contacts on the antigen, which has traditionally been a difficult task in ROSETTA protein interface design (Stranges and Kuhlman, 2013). This mutation is also more selective than V110P, only improving affinity for strains that have the correct electrostatic partner in position to make contact. The difference in mechanism between these two mutations illustrates the balance between breadth and affinity in antibody evolution. Mutations that improve only antibody stability without directly contacting the antigen are more likely to be beneficial across a panel of targets, whereas mutations that require specific electrostatic partners are likely to be more selective.

Implications for influenza studies

The antibody highlighted in this work, C05, is a clinically relevant antibody of interest for therapeutic and vaccine development. Since it targets a very small epitope on the receptor binding domain of influenza, it potently neutralizes certain viruses from both H1 and H3 subtypes (Ekiert et al., 2012). Using ROSETTA design and a trimeric linker, Strauch *et al*. were able to engineer a

protein binder targeting the influenza receptor binding domain that was based on the C05 epitope (Strauch et al., 2017). Our work suggests that C05 can be optimized further for affinity and breadth. One limitation to C05 binding is that it is susceptible to the 133a insertion that emerged after the 2009 H1 pandemic (Ekiert et al., 2012). Interestingly, one of the strains that was bound more tightly by C05 variants, A/mallard/Alberta/35/1976, is an avian virus that contains the lysine insertion at 133a characteristic of C05 escape (Wu et al., 2017). This finding suggests that C05 recognition of K133a strains may be possible with further optimization, improving this already potent antibody further.

Methods

Structure preparation

To generate templates for multistate design, we downloaded structures of the influenza hemagglutinin (HA) proteins and co-complex structures of influenza-binding antibodies from the Protein Data Bank (PDB). The structures were processed manually to remove waters and non-protein residues. The heavy chain constant region 1 (C_H1) and light chain constant region (C_L) domains of antibody structures were removed from the structure manually, and the structure was renumbered starting from residue 1. The HA structures were truncated to the head domain based on the start and end residues of the structure in PDB ID 4yk4. To generate mock complexes of antibody and antigen, the antigens in the panel were aligned to the antigen in the co-crystal structure of each antibody using the structural alignment feature in PyMOL (Schrodinger, LLC, 2015). To increase diversity of designed sequences, we performed replicates of multistate design using all copies of antibodies included in the asymmetric unit of the co-crystal structure and included all output models in our analysis. All HA sequences are denoted in H3 numbering.

Homology modeling

For modeling of large panels of HA structures, we first downloaded all unique H1 HA sequences from the Influenza Research Database (Zhang et al., 2017), which yielded 8,725 sequences. To reduce this panel to a size that could be processed on our computing hardware, we clustered these sequences at 95% homology using the CD-HIT software (Fu et al., 2012) to yield 524 sequences for homology modeling. We used RosettaCM to generate homology models based on 13 H1 HA template structures (Song et al., 2013). The top 5 templates in sequence homology were used with the multi-template RosettaCM protocol. 250 models were generated for each HA target, and the lowest energy model was moved forward for multistate design.

RECON multistate design

For inclusion in design, we considered any residue on the antibody with a heavy atom within 7 Å of a heavy atom on the HA. Residues on the HA that fulfilled the same distance cutoff were included as residues available for repacking. We ran RECON multistate design with four rounds of a ramping sequence constraint (Sevy et al., 2015). Backrub movements were performed on the backbone of the designable region of the antibody in between rounds of sequence design, to increase diversity (Smith and Kortemme, 2008). Designed models were refined by ROSETTA relax, with constraints to the starting coordinates to prevent the backbone from making substantial movements. Constraints were placed on all Cα atoms with a standard deviation of 1.0 Å. Sequences generated by multistate design were visualized using the WebLogo tool (Crooks et al., 2004). For RECON multistate design benchmarking, we included an ensemble of 524 antibody-antigen pairs, where the antigens were homology models made by RosettaCM, each paired with antibody C05. For production runs of multistate design, 7 different antibodies were paired with a panel of 13 H1 HAs with available structures. Each multistate design run was distributed on a

computing cluster such that each state (*i.e.,* each antibody-antigen pair) was handled on its own processor. This approach resulted in distribution over 524 processors for benchmarking and 13 processors for production.

Design validation

To validate the designs introduced by multistate design, we remodeled the mutations as isolated point mutations, either with each mutation separately, or with two or three mutant combinations in the case that the mutations appeared to be complementary. Multistate design models were evaluated visually to determine which mutations were complementary. The point mutants were refined with the same protocol as previously described, using ROSETTA relax with a 1.0 Å backbone constraint to the starting coordinates. We evaluated the effect of these mutations by normalizing the total score of the antibody-antigen complex and the binding energy (DDG) to a single metric of fitness, expressed as a Z score. DDG was defined as below:

$$DDG = E_{complex} - (E_{Ab} + E_{Ag})$$

where E_{Ab} and E_{Ag} are the energies of the antibody and antigen alone, respectively. A cutoff of 0.5 standard deviations was applied to identify candidates for expression and testing.

Recombinant antibody expression

33 variants of antibody C05 were identified from the computational screen and prioritized for experimental characterization. Point mutants were generated using site-directed mutagenesis with the QuikChange II kit (Agilent Technologies), using the recommended protocol. Variants that incorporated multiple mutants were synthesized (Synthetic Genomics) and cloned into an Ig expression vector (McLean et al., 2000) using the Gibson Assembly Master Mix reagent (New England BioLabs). Antibody variants were expressed by transient transfection of small scale (3 mL) cultures of Expi293F human embryonic kidney cells in serum-free medium (ThermoFisher

Scientific). The supernatants were harvested after 7 days, filter-sterilized with a 0.2-μm filter, and IgG concentration in supernatant was measured using the FortéBio Octet system for quantitation using anti-IgG AHQ sensors. Variants were screened for activity by loading variant IgG onto Octet anti-IgG biosensors and testing for binding to recombinant HA protein. Variants that showed increased association in screening were expressed in Expi293F cells in a larger scale. Supernatant was harvested as described above and purified using a 5 mL HiTrap MabSelectSure protein A column (GE Healthcare).

Recombinant HA expression

Sequences encoding the HA genes of interest were optimized for expression in human cells and synthesized (Genscript). Genes were constructed as soluble trimer constructs by replacing the transmembrane and cytoplasmic domain sequences with a GCN4 trimerization domain and a 6x-His tag at the C-terminus. Synthesized genes were cloned into the pcDNA3.1(+) mammalian expression vector (Invitrogen). HA protein was expressed by transient transfection of Expi293F cells (ThermoFisher Scientific). Supernatants were harvested after 7 days, filter-sterilized with a 0.2-μm filter, and purified using affinity chromatography with a 5 mL HisTrap excel column (GE Healthcare). HA head domain used for hydrogen-deuterium exchange was synthesized as a maltose-binding protein (MBP) fusion in pMAL-c5x vector (New England BioLabs). Head domain was expressed in SHuffle T7 Express competent *E. coli* (New England BioLabs) to enable disulfide formation in the cytoplasm, induced by the addition of 1 mM IPTG overnight at 18 °C, and purified using amylose resin (New England BioLabs).

Biolayer interferometry assay

Binding kinetics were determined using biolayer interferometry (BLI) with an Octet Red instrument (FortéBio, Menlo Park, CA). Recombinant HA proteins were labeled with biotin using

an EZ-Link Sulfo-NHS-Biotin labeling kit (ThermoFisher Scientific), at a molar ratio of 1:100 protein to biotin. HAs were loaded onto streptavidin biosensors at 10 μg/mL in kinetics buffer (PBS + 1% BSA, 0.05% Tween 20). The binding experiments were performed with the following steps: 1) baseline in kinetics buffer for 60 s, 2) loading of HA for 120-150 s, in order to achieve a response of 0.5 – 1.0 nm, 3) baseline for 60 s, 4) association of antibody for 300 s, and 5) dissociation of antibody into kinetics buffer for either 3 or 20 min. A reference well with antigen loaded onto the biosensor but no antibody was run in all experiments subtracted from sample wells to correct for drift and buffer evaporation. Four dilutions of antibody were used for each binding assay. Curves were fit to a 2:1 binding model using the FortéBio software and accepted if they fulfilled an R^2 of > 0.9.

ELISA binding assay

Recombinant HA was coated onto an ELISA plate (Nunc MaxiSorp flat-bottom plate, ThermoFisher Scientific) at 1 μg/mL and incubated overnight at 4 °C or for 1 hour at 37 °C. To reduce nonspecific binding, uncoated sites on wells were blocked with 5% milk powder (Bio-Rad) in PBS for 2 hours at room temperature. Antibodies were diluted serially 2-fold in blocking buffer starting at 1-20 μg/mL, for a total of 12 dilutions. Antibody dilutions were incubated with the coated plate for 1 hour at 37 °C. To detect binding, plates were incubated with mouse anti-human IgG Fc-HRP secondary antibodies (Southern Biotech) for 1 hour at 37° C. Binding was detected by addition of 100 μL of TMB substrate (ThermoFisher Scientific) and incubated for 5-10 min before quenching the reaction with 100 μL of 1 N HCl. Plates were read at 450 nm using a BioTek plate reader. After plate coating and primary and secondary antibody incubation, plates were washed 3x with wash buffer (PBS +0.05% Tween 20, Cell Signaling Technologies). EC_{50} values were calculated in GraphPad Prism using robust nonlinear regression.

Viruses and hemagglutination inhibition assay

Influenza virus strain A/Puerto Rico/8/1934 H1N1 was obtained from BEI Resources. A/Solomon Island/3/2006 H1N1 strain was provided by Influenza Reagent Resource of US CDC. The working stocks used for hemagglutination inhibition assay (HAI) were made in MDCK cell culture. For HAI, 25 µL of four hemagglutination units of virus were incubated for 1 hour at room temperature with 2 µL two-fold serial dilutions of antibodies starting at 100 µg/mL in PBS. The 50 µL of antibody-virus mixture was incubated for 45 minutes at 4 °C with 50 µL of turkey red blood cells (Rockland) diluted in PBS. The HAI titer was defined as the highest dilution of antibody that inhibited hemagglutination of red blood cells.

Differential scanning fluorimetry

Differential scanning fluorimetry was performed with 50 µg/mL IgG and SYPRO orange dye (ThermoFisher Scientific) at a 1:5,000 dilution, in a total reaction volume of 25 µL. Temperature cycling was done in a BioRad CFX96 real-time PCR system (BioRad), with a temperature gradient from 25 °C to 95 °C. The temperature was ramped in increments of 0.1 °C with a hold time of 3 s, and a two-minute hold at the first and last steps. Fluorescence was detected in FRET mode. The data were imported to Prism (GraphPad Software) and, to determine the apparent melting temperatures, the peaks in the first derivative plot were calculated. All melting curves were performed with four replicates and the average value is reported.

Crystallization and structural determination

HA1 (H3 numbering: residues 43–309) from A/Hong Kong/1/1968 (HK68/H3) was expressed in insect cells as described (Ekiert et al., 2012) and purified by Ni-NTA Superflow (Qiagen) and subsequently by size exclusion chromatography on a Hiload 16/90 Superdex 200 column (GE Healthcare) in 20 mM Tris pH 8.0, 150 mM NaCl, and 0.02% NaN_3. The C05

V110P/A117E mutant was incubated with HK68/H3 HA1 in a molar ratio of 1.5:1 overnight at 4 °C. The C05 V110P/A117E -HK68/H3 HA1 complex was purified by size exclusion chromatography on a Hiload 16/90 Superdex 200 column (GE Healthcare) in 20 mM Tris pH 8.0, 150 mM NaCl, and 0.02% NaN_3 and concentrated to 10 mg ml^{-1} in 10 mM Tris pH 8.0, 50 mM NaCl, and 0.02% NaN_3. Crystal screening was carried out using our high-throughput, robotic CrystalMation system (Rigaku, Carlsbad, CA) at The Scripps Research Institute, which was based on sitting drop vapor diffusion method with 35 µL reservoir solution and each drop consisting 0.1 µL protein + 0.1 µL precipitant. Diffraction quality crystals were obtained with reservoir solution containing 20% PEG 3000 and 0.1 M sodium citrate pH 5.5. The resulting crystals were cryoprotected by soaking in well solution supplemented with 20% ethylene glycol, flash cooled, and stored in liquid nitrogen until data collection.

Diffraction data were collected at the Stanford Synchrotron Radiation Lightsource beamline 12–2. The data were indexed, integrated and scaled using HKL2000 (HKL Research, Charlottesville, VA) (Otwinowski and Minor, 1997). The structure was solved by molecular replacement using Phaser (McCoy et al., 2007) with PDB 4FP8 (Ekiert et al., 2012) as the molecular replacement model, modeled using Coot (Emsley et al., 2010), and refined using Refmac5 (Murshudov et al., 2011). Ramachandran statistics were calculated using MolProbity (Chen et al., 2010).

Hydrogen-deuterium exchange mass spectrometry

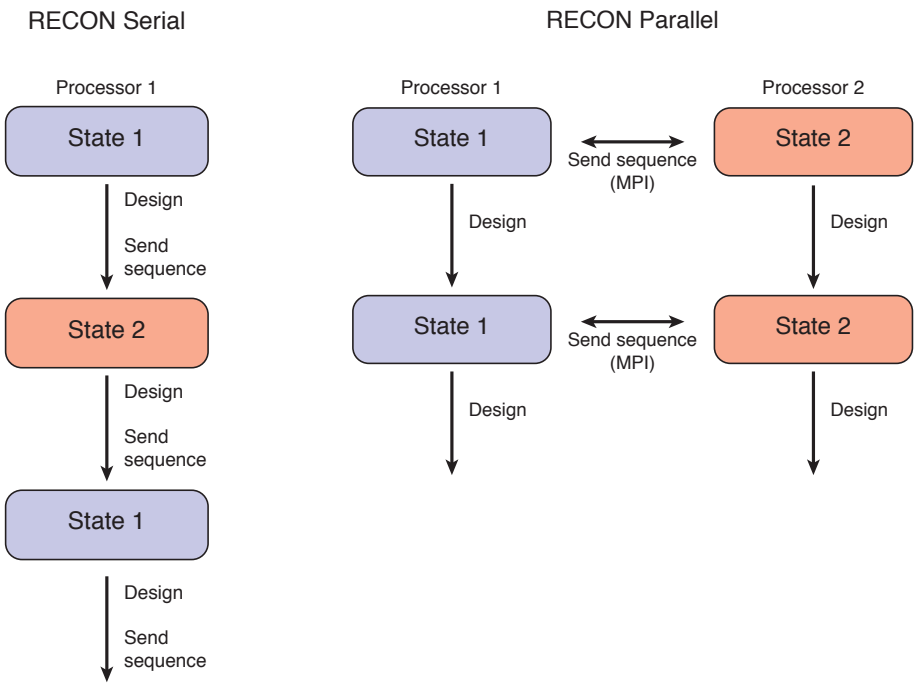
C05 variant V110P-A117E and A/Solomon Islands/03/2006 head domain were prepared at a protein concentration of 3 mg/mL individually and in complex before incubation for 2 h at 0 °C. A 20-fold dilution in 10 mM phosphate buffered saline, pH 7.5 in H_2O (no labelling) or D_2O (labelling), was performed. Diluted samples were incubated for 0 s, 5 s, 60 s, 30 min or 12 h at 20

°C. The labelling reaction was quenched by pH reduction to 2.5, by addition of 50 µL 4 M guanidinium/HCl, 100 mM tris(2-carboxyethyl)phosphine in 10 mM phosphate buffer saline, pH 2, at 0 °C. Samples were immediately injected into a nano-ACQUITY UPLC system with HDX technology (Waters Corporation, Milford, MA, USA). Online digest was performed at 20 °C and 4700 psi at a flow of 100 µL/min of 0.1 % formic acid in H_2O, using an immobilised-pepsin column. Peptides were trapped for 6 min at 0 °C, using a Waters VanGuard™ BEH C18 1.7 µm guard column, followed by separation using a 5-35 % acetonitrile gradient over 6 min, a flow of 40 µL/min at 0 °C on a Waters ACQUITY UPLC BEH C18 1.7 µm, 1 mm × 100 mm column. Online-coupled MS^E was performed with a Waters Xevo G2-XS with electrospray ionization and lock-mass acquisition (Leucine enkephalin, m/z=556.2771) of 3 scans every 60 s. The capillary was set to 2.8 kV, source-temperature to 80 °C, desolvation temperature to 175 °C, desolvation gas to 400 L/h and the instrument was set to scan over a m/z-range of 50-2000. A blank injection was performed between samples to avoid carry-over and all experiments were carried out in quadruplicate.

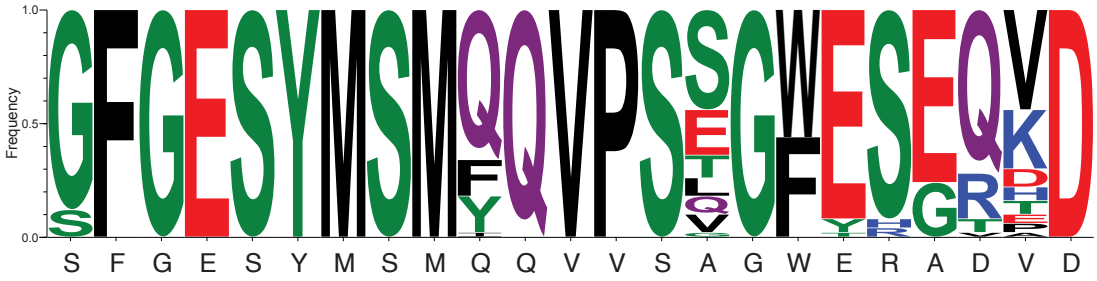
Peptides were identified in un-deuterated samples using Waters ProteinLynx Global Server 3.0.3 software with non-specific protease, min fragment ion matches per peptide of three, FDR 4% and oxidation of methionine as a variable modification. Deuterium uptake was calculated and compared to the non-deuterated sample using DynamX 3.0 software. Criteria were set to minimum intensity of 500, minimum products 3, minimum products per amino acid 0.2, mass error < 20 ppm and file threshold 3. The deuterium incorporation result is reported as the difference of the centroid values across the backbone amide population compared to the 0 s time point. Results were averaged across replicate analyses at a given time point and the standard deviation determined. For

this series of experiments, the average error for a single data point was ±0.8 Da or less within a single replicate.

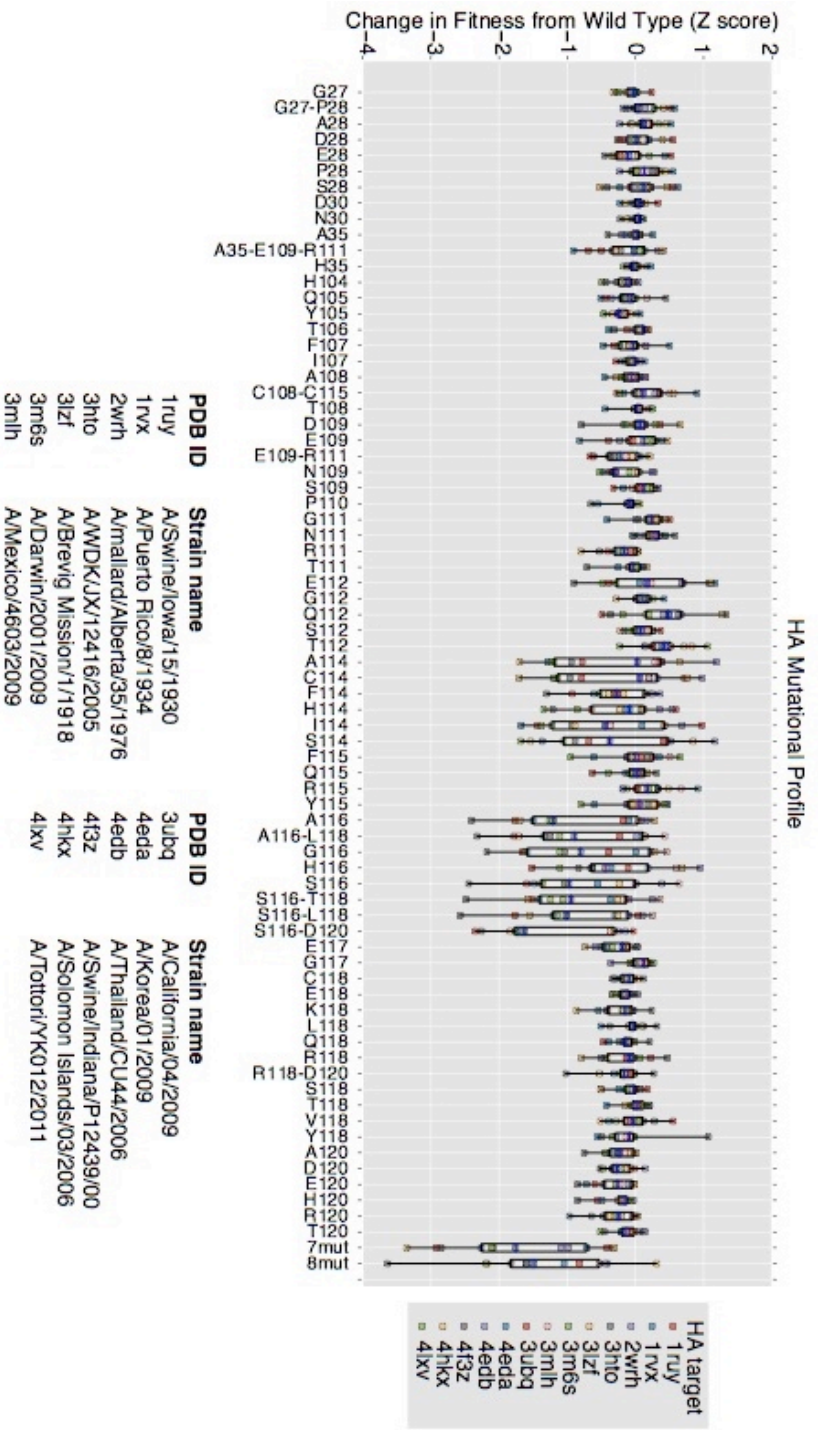
Supplemental Information



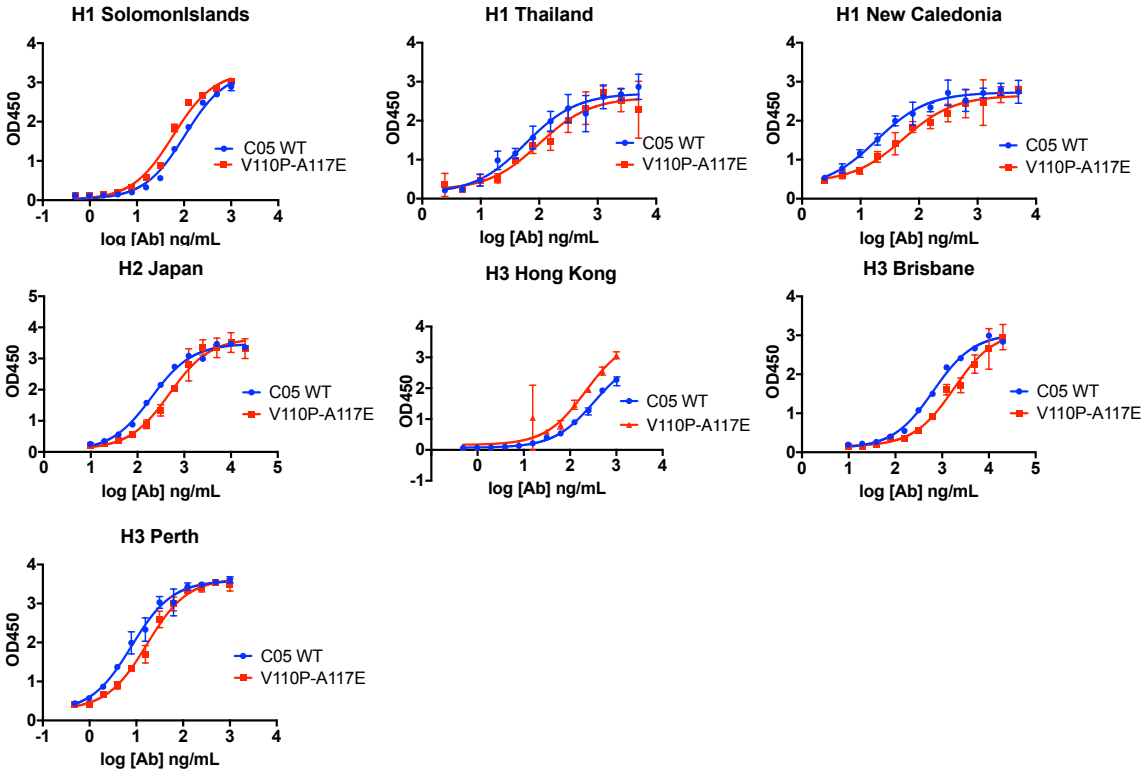
Supplementary Figure III.1. Schematic of RECON parallelization protocol.



Supplementary Figure III.2. Results of multistate design of anti-influenza antibody C05 against a panel of 524 viral proteins. The parallel RECON protocol was used to generate 50 independent design simulations of antibodies with predicted increased breadth against viruses in the panel.

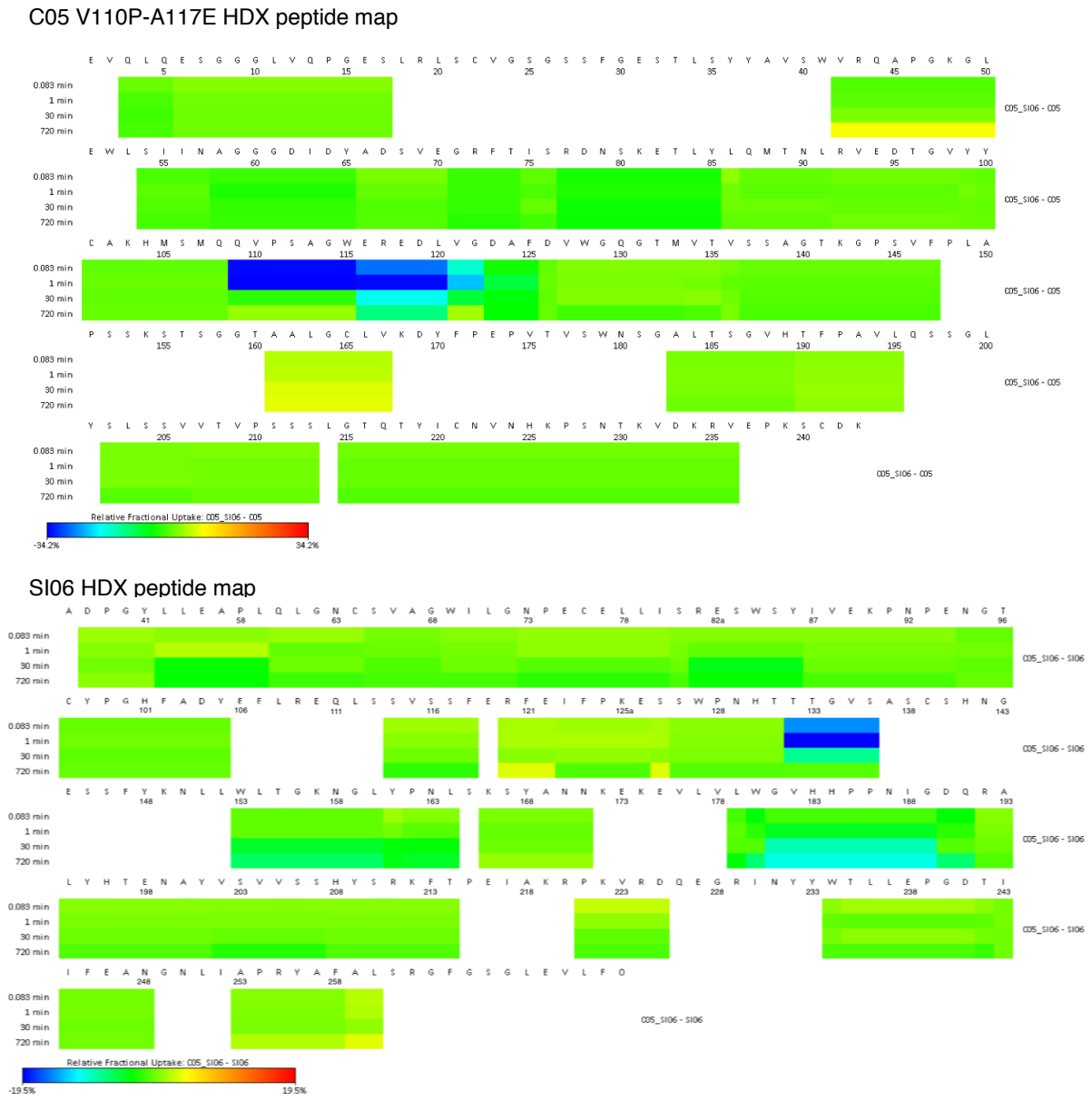


Supplementary Figure III.3. Breakdown of single and double amino acid mutations in antibody C05. All mutations introduced by multistate design of antibody C05 were modeled as single point mutants, or double mutants in the case that there were complementary mutations. In addition, two sets of mutation combinations were modeled (7mut and 8mut). Mutations are shown on the X axis. Y axis shows mutant fitness subtracted from wild-type fitness. Fitness is calculated as a normalized sum of antibody stability and antigen binding, expressed as a Z score.

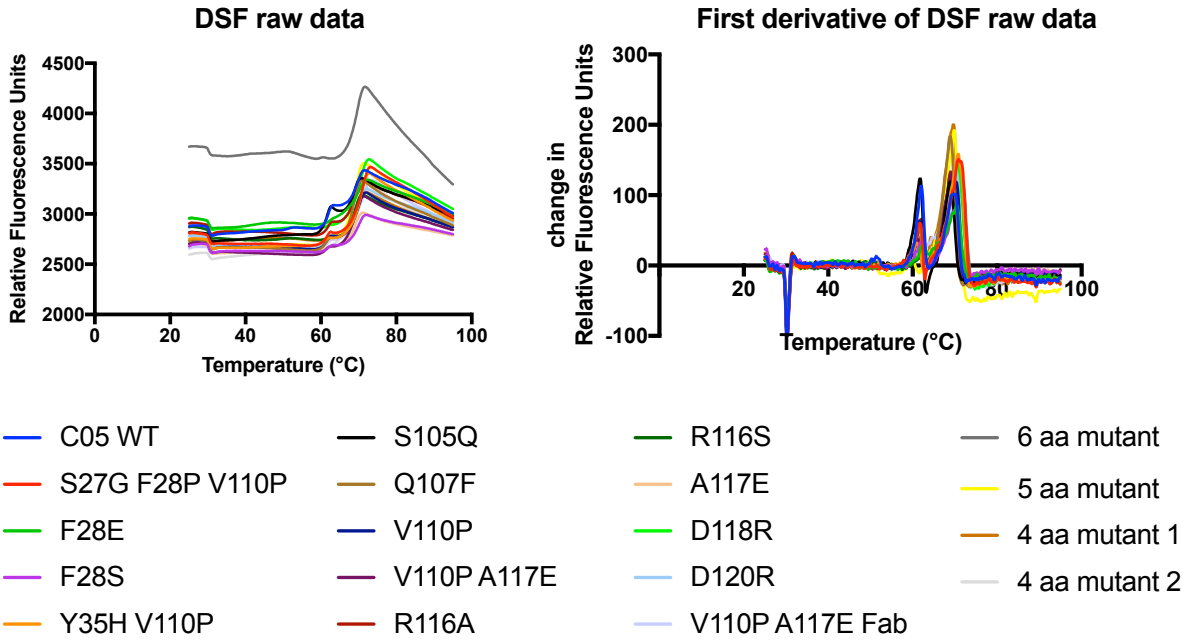


Strain	Subtype	Included in design?	EC_{50} (ng/mL)		EC_{50} 95% CI (ng/mL)	
			C05 WT	V110P-A117E	C05 WT	V110P-A117E
A/Solomon Islands/03/2006	H1	Yes	105	58	87-126	47-73
A/Thailand/CU44/2006	H1	Yes	59	99	37-94	58-170
A/New Caledonia/20/1999	H1	No	19	54	12-28	33-88
A/Japan/305+/1957	H2	No	199	483	159-250	369-632
A/HongKong/1/1968	H3	No	327	214	273-390	107-431
A/Brisbane/10/2007	H3	No	647	1771	492-851	1274-2482
A/Perth/16/2009	H3	No	8	17	6-10	14-20

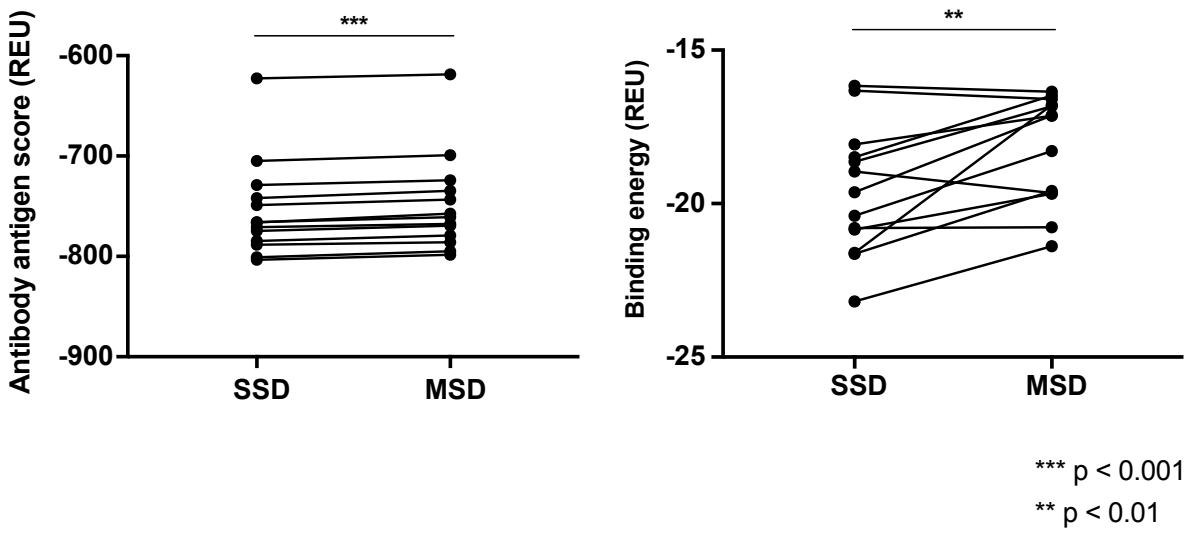
Supplementary Figure III.4. ELISA binding data of C05 mutants. ELISA binding curves are shown for 7 strains previously bound by C05 with high affinity (top). EC_{50} values were calculated for each binding curve and the EC_{50} and 95% confidence intervals are shown below. Also shown is the strain name, subtype, and whether the strain was included in the computational design panel.



Supplementary Figure III.5. Hydrogen-deuterium exchange (HDX) data of C05 V110P-A117E binding to the head domain of A/Solomon Islands/03/2006 (SI06). We measured the difference in deuterium uptake in the bound vs unbound states of both the antibody and antigen to map peptides comprising the paratope and epitope of the interaction. Peptides that were blocked from deuterium uptake upon binding mapped primarily to the CDRH3 loop of the antibody and the rim of the receptor binding domain on the antigen, in agreement with the modeling data and the co-crystal structure. Color code: yellow-green indicates no difference in deuterium uptake in bound vs. unbound state. Blue indicates less deuterium uptake in bound state compared to unbound state. Blank spaces are peptides not observed in experiment.



Supplementary Figure III.6. Melting curves from differential scanning fluorimetry (DSF). Curves are shown for all C05 mutants measured. The first derivative in the melting curve was calculated (right) to assign melting transitions for domain 1 (Fc) and domain 2 (Fab).



Supplementary Figure III.7. Tradeoff of affinity and breadth in design against H1 strains. For each of the 13 targets in the panel, C05 was redesigned using single state design (SSD) to increase affinity, as well as multi-state design (MSD) to increase breadth. We calculated the average score and binding energy for the ten best models resulting from both MSD and SSD. The points are connected by antigen, i.e. the SSD models against target PDB ID 1rvx are connected to the MSD models against target PDB ID 1rvx. The SSD models had significantly lower score (left) and binding energy (right) than the MSD models. Statistical significance was calculated using a Wilcoxon matched signed rank test in GraphPad Prism.

Supplementary Table III.1. H1 antigens used for multistate design. These antigens comprised all of the H1 structures in the Protein Data Bank (PDB) at high resolution (< 3.5 Å).

Strain	PDB ID	Resolution	Species
A/mallard/Alberta/35/1976	2wrh	3.0	Avian
A/WDK/JX/12416/2005	3hto	2.95	Avian
A/Thailand/CU44/2006	4edb	2.5	Human
A/Swine/Indiana/P12439/00	4f3z	3.2	Swine
A/Tottori/YK012/2011	4lxv	3.0	Human
A/Swine/Iowa/15/1930	1ruy	2.7	Swine
A/Puerto Rico/8/1934	1rvx	2.2	Human
A/Brevig Mission/1/1918	3lzf	2.8	Human
A/Darwin/2001/2009	3m6s	2.8	Swine/Human
A/Mexico/4603/2009	3mlh	2.1	Swine/Human
A/California/04/2009	3ubq	2.0	Swine/Human
A/Korea/01/2009	4eda	2.7	Swine/Human
A/Solomon Islands/03/2006	4hkx	2.5	Human

Supplementary Table III.2. Explanation of the energetic contributions of mutated residues.

Mutant	HAs with positive effect	Description
G27 P28	4edb,4f3z	Stabilizes the CDRH1 loop
E28	4f3z	Makes an H bond with R56 on the light chain, also with N159 on antigen
S28	3lzf,4f3z	S28 can make a H bond with Q192 on the antigen
H35	3m6s,4edb	Makes an H bond with T32 on the CDRH1
Y105	3m6s,4lxv	Y packs well against the antigen, makes an H bond to D131 on the antigen
Q105	1ruy, 1rvx	Makes an H bond with E107 MC on the antigen
F107	1rvx	Favorable packing against the CDRH2 up into the antibody - minimal effect though
T108	1rvx	Makes an H bond with W114 main chain in the CDRH3
E109	1rvx	Makes an H bond with N193 on the antigen
D109	4f3z	D109 makes an H bond with S193 on the antigen. May be redundant with N109.
E109-R111	1ruy,1rvx,3ubq	In some cases E and R can interact with each other in addition to the 190 helix - in others E interacts with helix and R interacts with E219 further down the antigen
P110	1rvx, 4f3z	Increased VDW interactions, phi-psi angle stabilization and pi-pi stacking with Y35 on the CDRH1
R111	3hto,3lzf	Makes H bonds with S189 and E190 side chains on the antigen
E112	4f3z	Makes favorable H bonds with H183 and S186
F114	1ruy,3hto,4eda,4lxv	F can stack slightly better with W153 at RBD base. Lets some neighboring residues adopt more favorable rotamers
H114	1ruy,3hto,3m6s,4lxv	H can stack slightly better with W153 at RBD base. Lets some neighboring residues adopt more favorable rotamers
Y115	1rvx,3m6s	Y can make nice VDW interactions with K145 on the antigen. May be redundant with F115
A116	1ruy, 3hto, 3lzf, 3m6s, 4eda, 4f3z, 4lxv	Relieves clash with K133a insertion on RBD rim
S116	1ruy, 1rvx, 3hto, 3lzf, 3m6s, 4eda, 4f3z, 4lxv	Relieves clash with K133a insertion on RBD rim
S116-D120	1ruy, 3hto, 3lzf, 3m6s, 4eda, 4f3z, 4lxv	D120 can interact with K133a once S116 relieves the clash
E117	1rvx, 2wrh, 3hto, 3lzf, 4edb	E can interact with K125a on the antigen
R118	1rvx, 3hto, 3lzf, 4hkx	R can interact with E131 on the antigen surface
S118	3lzf	Makes an H bond with T133 on the antigen
D120	3hto, 4f3z, 4lxv	D can interact with S105 and Q107 via H bonding
T120	4f3z, 4lxv	T can interact with S105 and Q107 via H bonding
H120	3ubq,4f3z	Three way H bond between H120, S105 and D122 on the CDRH3
R120	1rvx,4f3z	Electrostatic interactions and H bonds between R120 and D50 (LC), Y91 (LC), and E158 (Ag)

Multiple mutants expressed:

8mut	E28,Y105,F107,E109,P110,R111,A116,D120	5mut	E28,P110,R111,A116,E117
7mut	E28,Q105,P110,S116,E117,R118,D120	4mut1	P110,R111,A116,E117
6mut	E28,P110,R111,A116,E117,R118	4mut2	E28,P110,R111,A116

Supplementary Table III.3. X-ray collection statistics of C05 V110P-A117E double amino acid mutant in complex with HA of H3 A/Hong Kong/1/68.

X-ray data collection and refinement statistics

Data collection	
Beamline	SSRL 12-2
Wavelength (Å)	1.0332
Space group	$P2_1$
Unit cell parameters (Å and °)	a=91.6, b=258.4, c=91.9, β=90.5
Resolution (Å)	50-3.25 (3.35-3.25)[a]
Unique Reflections	64,646 (5,630)[a]
Redundancy	4.3 (3.9)[a]
Completeness (%)	96.0 (93.0)[a]
$\langle I/\sigma_I \rangle$	8.9 (1.1)[a]
R_{sym}[b]	0.16 (0.78)[a]
R_{pim}[b]	0.09 (0.43)[a]
$CC_{1/2}$[c]	0.99 (0.50)[a]
Z_a[d]	4
Refinement statistics	
Resolution (Å)	50-3.25
Reflections (work)	61,178
Reflections (test)	3,065
R_{cryst}(%)[e] / R_{free}(%)[f]	25.4 / 26.8
No. of atoms	
HA1	8,292
Fab	13,644
Glycan	28
Average *B*-value (Å^2)	
HA1	112
Fab	119
Glycan	95
Wilson *B*-value (Å^2)	63
RMSD from ideal geometry	
Bond length (Å)	0.012
Bond angle (°)	1.39
Ramachandran statistics (%)	
Favored	96
Outliers	0.3
PDB code	Pending

[a] Numbers in parentheses refer to the highest resolution shell.

[b] $R_{sym} = \Sigma_{hkl} \Sigma_i | I_{hkl,i} - \langle I_{hkl} \rangle | / \Sigma_{hkl} \Sigma_i I_{hkl,i}$ and $R_{pim} = \Sigma_{hkl} (1/(n-1))^{1/2} \Sigma_i | I_{hkl,i} - \langle I_{hkl} \rangle | / \Sigma_{hkl} \Sigma_i I_{hkl,i}$, where $I_{hkl,i}$ is the scaled intensity of the ith measurement of reflection h, k, l, $\langle I_{hkl} \rangle$ is the average intensity for that reflection, and *n* is the redundancy.

[c] $CC_{1/2}$ = Pearson correlation coefficient between two random half datasets.

[d] Z_a is the number of HA1-Fab complexes per crystallographic asymmetric unit.

[e] $R_{cryst} = \Sigma_{hkl} | F_o - F_c | / \Sigma_{hkl} | F_o | \times 100$, where F_o and F_c are the observed and calculated structure factors, respectively.

[f] R_{free} was calculated as for R_{cryst}, but on a test set comprising 5% of the data excluded from refinement.

Supplementary Table III.4. Melting temperatures of all C05 mutants measured.

Variant	**Transition 1 (°C)**	**Transition 2 (°C)**	**Significance**
C05 WT	62.0	69.8	
S27G F28P V110P	61.8	70.8	**
F28E	62.7	70.4	*
F28S	62.1	70.1	
Y35H V110P	61.3	70.8	**
S105Q	61.8	68.8	**
Q107F	61.6	68.9	**
V110P	61.6	70.3	*
V110P A117E	61.5	69.2	*
V110P A117E Fab		70.7	
R116A	62.0	69.1	*
R116S	61.8	68.9	**
A117E	61.7	69.2	*
D118R		70.6	**
D120R	61.8	70.3	*
6 aa mutant	61.8	71.9	**
5 aa mutant		69.9	*
4 aa mutant 1		69.7	
4 aa mutant 2		69.7	

**$p<0.005$

*$p<0.05$

CHAPTER IV.

Integrating linear optimization with structural modeling to increase HIV neutralization breadth

Adapted from Sevy, A. M., Panda, S., Crowe, J. E., Meiler, J. & Vorobeychik, Y. Integrating linear optimization with structural modeling to increase HIV neutralization breadth. PLoS Comput. Biol. 14, e1005999 (2018).

Author contributions: I was the co-first author of this manuscript, with collaboration from S.P. I co-wrote this manuscript with S.P. under the mentorship of Jens Meiler, James Crowe, and Eugene Vorobeychik. I came up with hypotheses, designed experiments, and conducted all ROSETTA modeling experiments described in this work. All figures are reprinted with permission from the publisher.

Abstract

Computational protein design has been successful in modeling fixed backbone proteins in a single conformation. However, when modeling large ensembles of flexible proteins, current methods in protein design have been insufficient. Large barriers in the energy landscape are difficult to traverse while redesigning a protein sequence, and as a result, current design methods only sample a fraction of available sequence space. We propose a new computational approach that combines traditional structure-based modeling using the ROSETTA software suite with machine

learning and integer linear programming to overcome limitations in the ROSETTA sampling methods. We demonstrate the effectiveness of this method, which we call BROAD, by benchmarking the performance on increasing predicted breadth of anti-HIV antibodies. We use this novel method to increase predicted breadth of naturally-occurring antibody VRC23 against a panel of 180 divergent HIV viral strains and achieve 100% predicted binding against the panel. In addition, we compare the performance of this method to state-of-the-art multistate design in ROSETTA and show that we can outperform the existing method significantly. We further demonstrate that sequences recovered by this method recover known binding motifs of broadly neutralizing anti-HIV antibodies. Finally, our approach is general and can be extended easily to other protein systems. Although our modeled antibodies were not tested *in vitro*, we predict that these variants would have greatly increased breadth compared to the wild-type antibody.

Introduction

Computational design has been used successfully by protein engineers for many years to alter the physicochemical properties of proteins (Dahiyat and Mayo, 1997; Kuhlman et al., 2003). In the simplest case, protein design involves optimizing the amino acid sequence of a protein to accommodate a desired 3-D conformation. This approach has been extended to related tasks such as protein-protein interface design, de novo design of protein binding molecules, design of self-assembling protein nano-cages, etc. (Fleishman et al., 2011b; King et al., 2012; Strauch et al., 2017; Willis et al., 2015). Each of these examples involves the straightforward application of design methodologies to a single, static protein conformation. However, there is a need to extend protein design to apply to several conformations simultaneously. These approaches, referred to as multistate design (MSD), can be used to modulate protein specificity, model protein flexibility,

and engineer proteins to undergo conformational changes (Davey and Chica, 2014; Guntas et al., 2015; Havranek and Harbury, 2003; Howell et al., 2014; Lewis et al., 2014; Shifman and Mayo, 2002; Willis et al., 2013). Several methods have been developed to enable computationally expensive multistate design (Leaver-Fay et al., 2011a; Sevy et al., 2015). However, these methods all suffer from large energetic barriers that limit sampling in sequence space, resulting in sub-optimal designs (Sevy et al., 2015). In addition, these methods are severely limited in scale by the size and number of states that can be included. To address these limitations, we have developed a method that integrates structural modeling with integer linear programming to enable a fast global search through large ensembles of target states.

Results

Experimental workflow

Our design algorithm, which we call BROAD (BReadth Optimization for Antibody Design) incorporates ROSETTA-based structural modeling with integer linear programming to more easily traverse boundaries in the energy function (Figure IV.1). The experimental workflow involves generating a large training set of randomly mutated proteins, fitting a linear model (described below) to predict binding, and using integer linear programming to find an optimal antibody sequence balancing stability and binding with respect to a collection of target virus epitopes. We applied this method to the problem of designing broadly binding anti-HIV antibodies. We modeled anti-HIV antibody VRC23 (Georgiev et al., 2013) against a set of 180 diverse viral proteins, creating antibody variants that were mutated randomly in the paratope region. The viral panel used was derived from Chuang *et al.* (Chuang et al., 2013). Based on known binding patterns of VRC23 we calculated the predicted binding energy that corresponds to observable binding, and

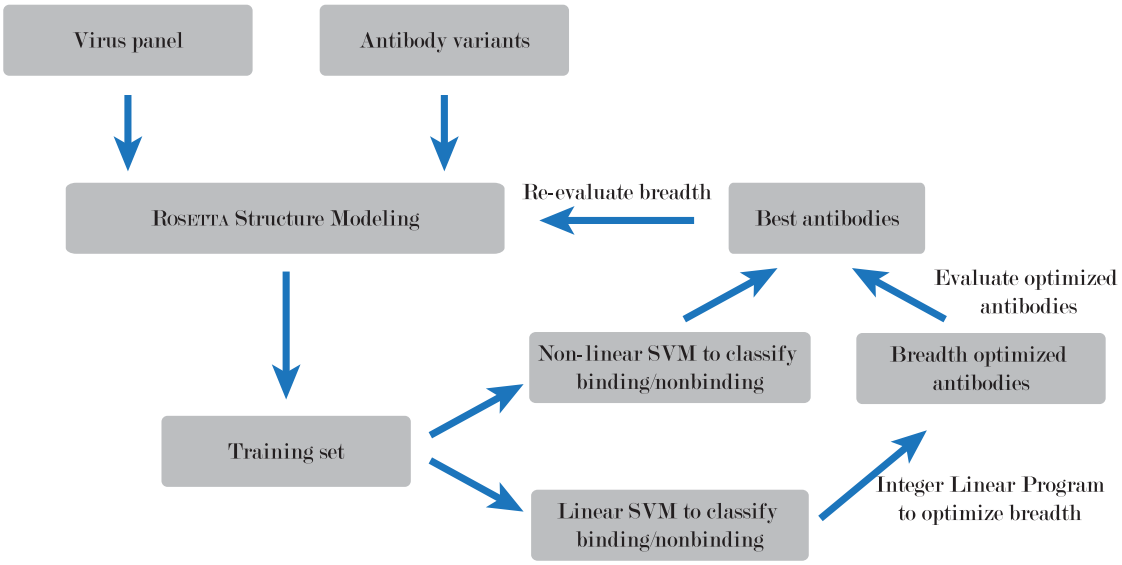


Figure IV.1. Experimental workflow of the BROAD design method. The method uses ROSETTA structural modeling to generate a large set of mutated antibodies, support vector machines (SVM) to predict ROSETTA energy from amino acid sequence, and integer linear programming to optimize breadth of binding across a set of viral proteins.

searched antibody space using integer linear programming to optimize stability of the unbound antibody while achieving predicted 100% binding breadth to the 180 target viral proteins. We then used a non-linear Support Vector Machine classifier, trained on the entire dataset produced by ROSETTA, to identify top sequences. Finally, we entered the top scoring sequences back into ROSETTA structural modeling to measure the predicted breadth of antibody variants.

Sequence-based Linear Classification and Regression Models to Predict Binding and Stability

Our end goal is to design broadly binding and stable antibodies by searching the sequence space, i.e., to optimize the amino acids at each binding position of the antibody. The key challenge for this approach is that an exhaustive search in the combinatorial sequence space is intractable.

To address this issue, we first propose to learn sequence-based linear classification and regression models to predict binding and stability from data. Building on these models, we formulate an integer program to accomplish global search in the antibody sequence space.

To generate our training set, we determined three contiguous stretches on the antibody that are in contact with the viral protein. These positions were determined to be residues 46-62, spanning FR2-CDR2-FR3; residues 71-74 in FR3; and residues 98-100b in CDR3 (Supplementary Figure IV.1). We then created randomly mutated antibody variants, modeled their binding poses using ROSETTA, and used this data to train a binding classifier to predict ROSETTA score and binding energy from amino acid composition.

The binding classifier is based on the assumption that the amino acids at the binding positions of the antibody interact with those on the binding positions of the virus. In particular, this model assumes that binding between an antibody and a viral protein is determined by two factors: a) the individual amino acids in each binding position of the antibody and the virus respectively and b) the effects of the pairwise amino acid interactions between the antibody and the virus respectively. To capture these, we construct a sequence-based binary feature vector from the input antibody and virus pair, which explicitly represents the individual and pairwise amino acid contributions. Let the input antibody-virus pair represented as vectors of amino acids, be denoted by $(\mathbf{a}, \mathbf{v})$. Let $b(\mathbf{a}, \mathbf{v})$ denote the ROSETTA predicted binding energy for $(\mathbf{a}, \mathbf{v})$ and let $\Phi(\mathbf{a}, \mathbf{v})$ denote the binary binding decision. We chose a threshold θ such that $\Phi(\mathbf{a}, \mathbf{v}) = +1$ if $b(\mathbf{a}, \mathbf{v}) \leq \theta$ (i.e., **a and v** bind) and $\Phi(\mathbf{a}, \mathbf{v}) = -1$ otherwise. For evaluation of our approach, we choose the value of θ based on experimental neutralization data. This data is available as the experimental neutralization IC50 (in units of μg/ml) of VRC23 with the 180 virus sequences in the panel (Chuang et al., 2013). Lower values represent better neutralization potency and values that

have '>50' concentration represent a virus that is not neutralized by VRC23. Accordingly, VRC23 has a neutralization breadth of 63.5% on this panel. We set θ = -28.5 such that the VRC23 breadth of binding computed on the ROSETTA generated data (sequences and the corresponding ROSETTA binding scores) is consistent with the above experimental neutralization data.

We learn the classifier $\Phi(\mathbf{a}, \mathbf{v})$ as a linear Support Vector Machine (SVM) (Cortes and Vapnik, 1995) using the binary feature set comprised of actual antibody and virus sequences along the corresponding binding sites, as well as all pairwise interactions of antibody and virus amino acids. The SVM classifier uses the ROSETTA binding energy as the ground truth, and allows more efficient sampling by approximating the ROSETTA score function by sequence alone. To optimize the L2 regularization parameter of the SVM, we performed 10-fold cross-validation on the full dataset, using 80% of the data for training and 20% for testing. Smaller parameter values enforce higher regularization and higher values lead to overfitting. The average prediction accuracy is shown in Figure IV.2A for different values of the L2 regularization parameter. We also plot the prediction error on the two classes: binders (+1) and non-binders (-1). The prediction accuracy is 67% on the test set using the optimized parameter (a random predictor would be at 50%). We observe that even if the prediction accuracy is relatively low, it provides reasonable signal within the subsequent breadth optimization step (discussed in the results section). Since the final decision is determined by solving the breadth optimizing integer linear program, our approach does not rely on a highly accurate classification model. In previous research (Kamisetty et al., 2015), a similar model was introduced to predict ΔG values for interaction between PDZ domains and peptide ligands. The result was a 0.69 correlation coefficient in 10-fold cross validation. This model can also be interpreted to identify the important binding position pairs that contribute significantly to the final prediction. We plot this interaction strength for each pairwise interaction in Figure IV.2C

(please refer to the methods section for details).

Next, we learned a linear regression model to predict the thermodynamic stability, using only the antibody amino acids as features. The prediction of thermodynamic stability is necessary to ensure that our designed antibodies can be expressed stably. To simplify the approach, we predicted the stability of the antibody-virus complex as a function of the antibody sequence only (note that we do not make this assumption during evaluation). Specifically, we constructed a binary feature vector restricted to amino acids in the antibody binding positions. Let $s(\mathbf{a}, \mathbf{v})$ denote the ROSETTA stability for the pair $(\mathbf{a}, \mathbf{v})$. We learn a linear model $\Psi(\mathbf{a})$ to predict $s(\mathbf{a}, \mathbf{v})$ for an antibody $\mathbf{a}$ (i.e., independent of the virus). To measure the accuracy of prediction, we computed the correlation coefficient between the true scores and the predicted scores. Interestingly, our assumption that stability scores are only weakly dependent on the virus protein sequence is borne out: we found a correlation of 0.85 between the predicted and actual stability energy score on the test set (Figure IV.2B).

Algorithm

Given the classification and regression model learned from data, we formulate an integer linear program (ILP) to optimize the amino acids in the antibody sequence space to achieve both breadth and stability. The variables are the amino acids in the antibody binding positions. The objective function optimizes the predicted stability score (i.e., minimizes $\Psi(\mathbf{a})$). The constraints represent the condition that the designed antibody should bind to all the viruses in the panel, using binding predictions from $\Phi(\mathbf{a}, \mathbf{v})$. This algorithm is outlined in Supplementary Figure IV.2.

Armed with these tools, we used the following protocol to generate a collection of candidate antibodies to be evaluated using ROSETTA. First, we took a random subsample of the full training data corresponding to 100 out of the 180 virus sequences. Using only this subsample,

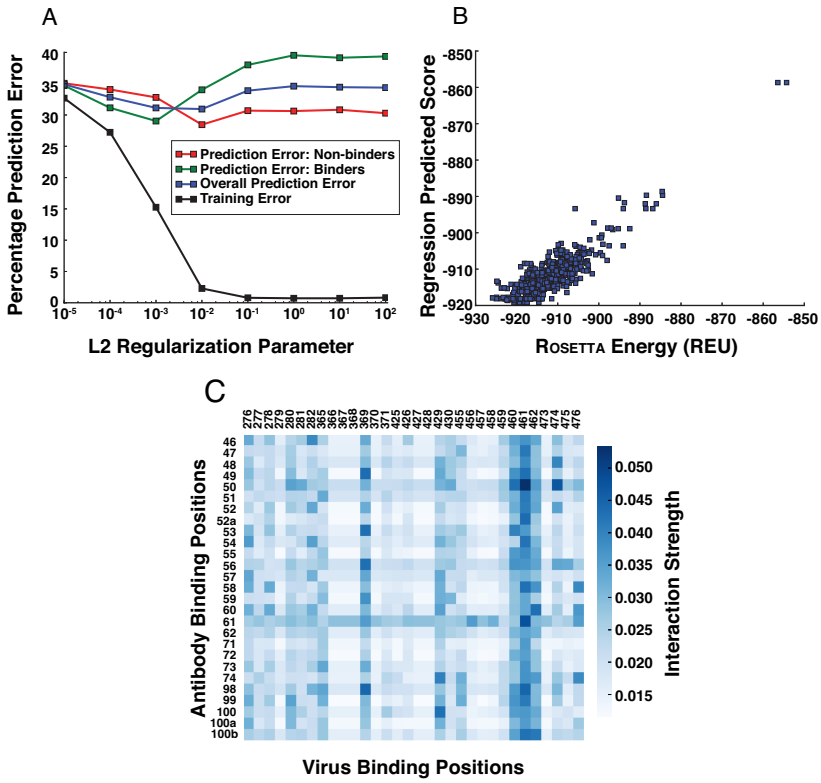


Figure IV.2. Training results for the linear classification. A. 10-fold cross validation results. B. Correlation between predicted score and ROSETTA energy score in linear regression. C. Interaction strength of each pairwise interaction between antibody and virus binding positions are also shown.

we trained the binding and stability models, $\Phi(\mathbf{a}, \mathbf{v})$ and $\Psi(\mathbf{a})$ respectively. We then solved the ILP described above to compute a stable, broadly-binding antibody sequence, considering only the 100 out of 180 selected virus sequences (that is, we only constrain the ILP to bind to these 100 virus proteins, rather than the full set of 180). We repeated this procedure 50 times, to obtain 50 candidate antibody sequences. To validate these optimized antibody candidates, we predicted binding and stability scores using a model trained on all the data. In case of stability prediction, we used a linear model as described above (since the model is reasonably accurate). For binding prediction however, we trained a non-linear (radial basis function kernel) SVM for improved prediction accuracy. Each of the 50 candidate antibodies were scored using these models trained

on all data, in terms of predicted binding breadth and stability, and 10 best candidates were then chosen for ROSETTA evaluation using the full panel of 180 virus proteins. This procedure is outlined in Supplementary Figure IV.3.

Redesign of VRC23 improves predicted breadth

After generating redesigned antibody sequences with predicted increases in breadth, we threaded these sequences onto the VRC23-gp120 complexes and subjected them to structural modeling to measure the change in predicted breadth. We refined the complexes using the ROSETTA relax protocol – to test the accuracy of the ROSETTA relaxed models, we compared the relaxed models to solved structures of gp120 viral variants and computed the root mean squared deviation (RMSD) over Cα atoms on gp120. We observed that the relax protocol recapitulates the gp120 conformations with an average RMSD of 2.2 Å, whereas the pairwise RMSD between gp120 conformations, representing the intrinsic flexibility of these molecules, is 1.8 Å (Supplementary Table IV.1). Considering that we substituted only residues at the binding site of the gp120 variants, and not the entire gp120 sequence, we consider that the variant gp120 conformations are recapitulated with sufficient accuracy for this experiment. As a control, we generated sequences using structure-based multistate design with the RECON method (Sevy et al., 2015). The RECON method uses ROSETTA design combined with coordination between differing states to generate an antibody sequence with increased affinity for all target states. Using RECON to redesign antibody-antigen complexes has been benchmarked and been shown to generate germline-like, broadly binding antibodies (Sevy et al., 2015). We compared the 10 sequences created by BROAD to 10 sequences generated by RECON multistate design to compare the change in breadth to alternate approaches. We found that the BROAD method resulted in a significant increase in predicted breadth over the RECON multistate design method (Figure

IV.3A). The BROAD-designed antibodies were able to achieve predicted breadth ranging from 86.1 – 100% of viruses, whereas multistate designed antibodies reached a predicted breadth of 62.8 – 85.6% of viruses. Notably, both methods were able to increase predicted breadth from the starting value of 53.3% for wild-type VRC23. This finding suggests that the wild-type VRC23 sequence is sub-optimal for breadth, which is supported by the observation that other known broadly neutralizing antibodies bind in a similar mode to VRC23 but with breadths exceeding 85% (Diskin et al., 2011; Klein et al., 2013; Scheid et al., 2011; Zhou et al., 2010). In addition, we observed that the BROAD method samples sequence space that is not sampled in multistate design (Figure IV.3B). We hypothesize that the BROAD method is able to cross energetic barriers that restrict sampling in traditional structure-based design methods, and is thereby able to generate antibodies with greater predicted breadth and lower energy. To support this hypothesis, we analyzed the difference in score and binding energy for antibodies designed by BROAD and multistate design over the panel of viral proteins (Figure IV.4). BROAD was consistently able to generate lower energy antibody-antigen complexes, with a marked decrease in binding energy. This finding supports the hypothesis that BROAD is able to search sequences that are unavailable to multistate design, and that these new sequences have favorable score and binding energy.

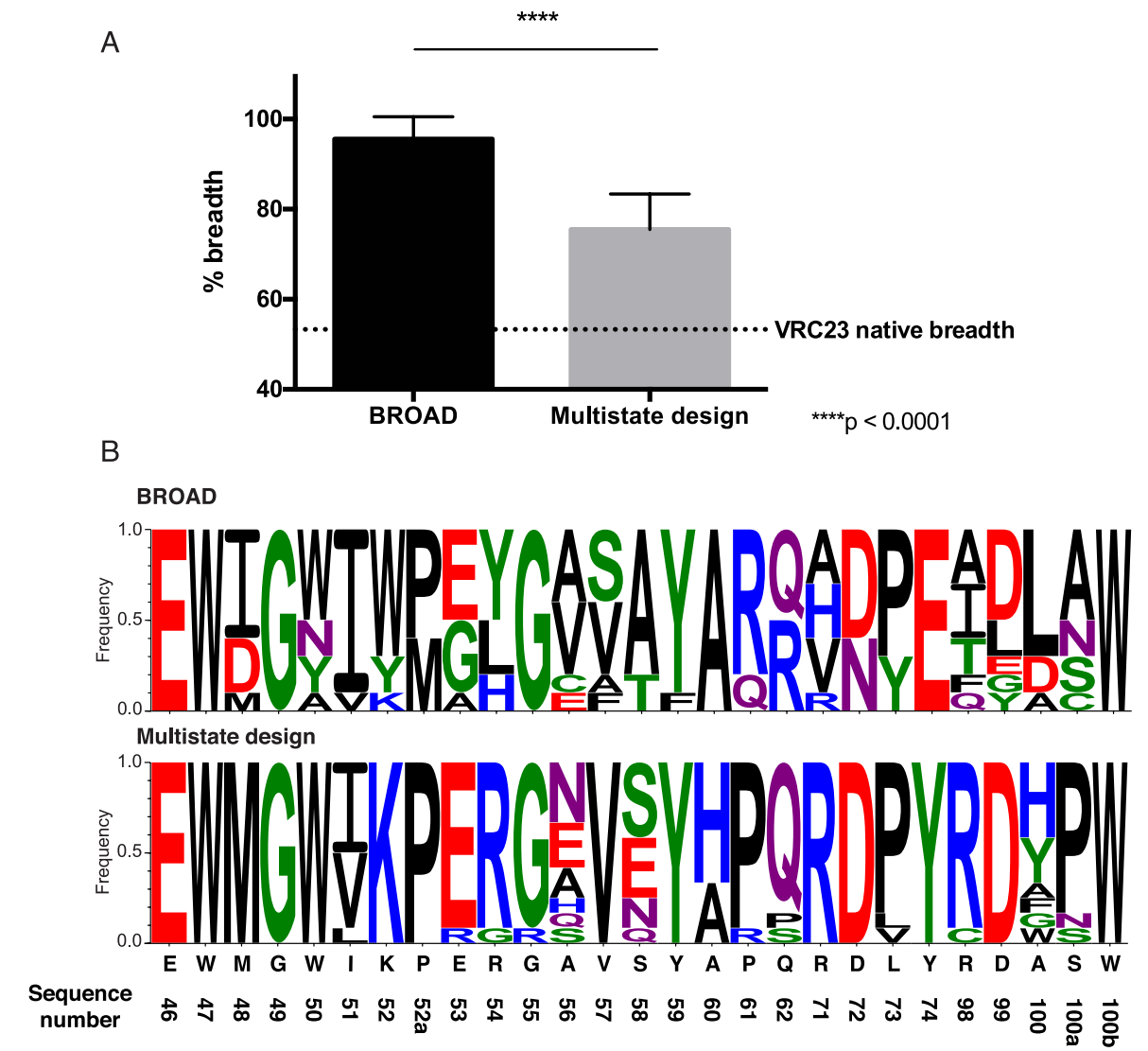


Figure IV.3. Redesign of VRC23 using integer linear programming increases predicted breadth over HIV viral strains. A. Predicted breadth of 10 redesigned antibodies generated either by BROAD or multistate design. Bars show mean and standard deviation of 10 sequences. Dotted line shows the predicted breadth of the native VRC23 antibody. B. Sequence logos of designed antibodies generated by BROAD or multistate design. Amino acids are colored based on chemical properties. The native VRC23 sequence is shown below.

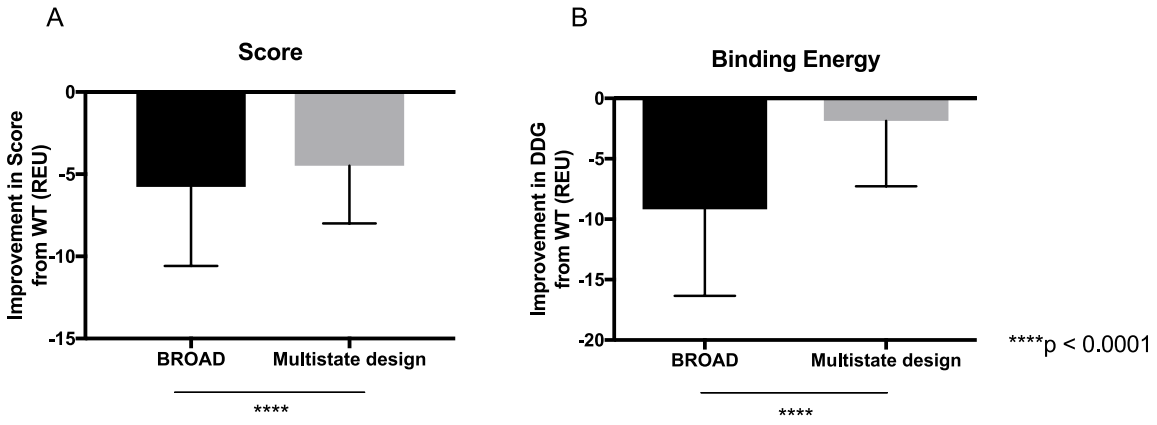


Figure IV.4. Score comparison of redesigned antibodies. The ROSETTA score (A) and binding energy (DDG) (B) are shown for ten redesigned antibodies made either by BROAD or multistate design, paired with 180 viruses. Bar plots shown mean and standard deviation. Shown on the Y axis is difference between score/DDG between the redesigned antibody and wild-type.

Designed residues recapitulate known binding motifs

A frequent problem in computational protein design is false positives – that is, sequences that are predicted to be favorable according to the score function, but are unable to recapitulate that activity *in vitro*. The ROSETTA score function uses many approximations of energetic terms to enable faster simulations, and these approximations can introduce inaccuracies (Bender et al., 2016; Leaver-Fay et al., 2013). To reduce the possibility that the redesigned VRC23 variants are scored favorably due to inaccuracies in the score function, we compared the designed residues introduced by BROAD to structural motifs of known broadly neutralizing antibodies (Figure IV.5). In several cases, the residues introduced by BROAD mimicked a known interaction of an existing antibody. For example, position 61 was mutated from proline in VRC23 to arginine (Figure IV.5, top left). The broadly neutralizing antibody VRC01 has an arginine that occupies similar space to the designed arginine (Zhou et al., 2010). This phenomenon can be observed for

several different broadly neutralizing antibodies, such as VRC-CH31, 3BNC117, and NIH45-46, all of which target the CD4 binding site, but at slightly different orientations (Diskin et al., 2011; Klein et al., 2013; Zhou et al., 2010; 2013). We observed several examples of this type of recapitulation. Mutation Q62R on VRC23 placed an arginine residue to fill space that is occupied by a tyrosine on VRC-CH31 (Figure IV.5, top right) - this mutation fills a void at the interface to improve antibody-antigen packing. Mutation L73Y places an aromatic group overlapping with the position of a tyrosine in antibody 3BNC117, which also improves packing with the antigen (Figure

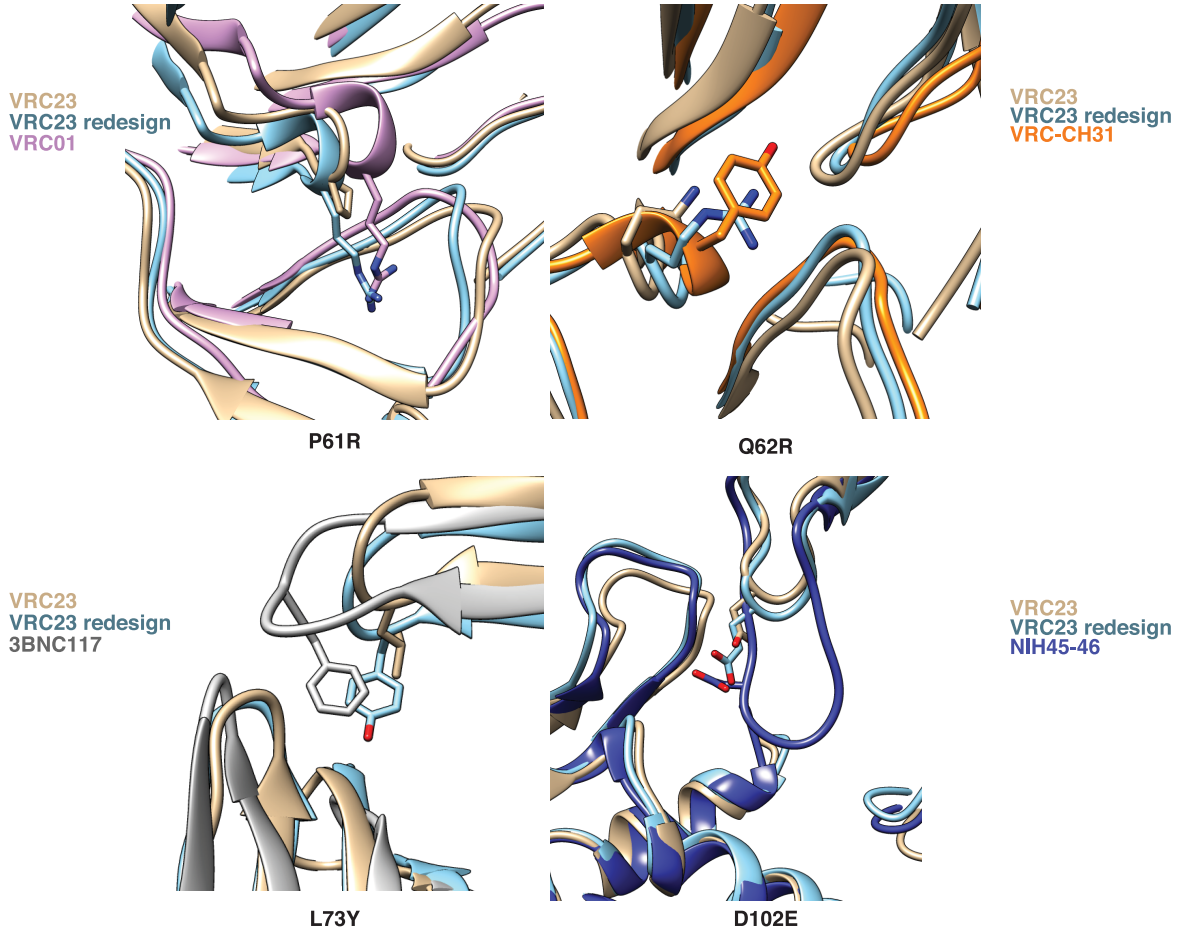


Figure IV.5. BROAD design recapitulates structural motifs of known broadly neutralizing antibodies. Residues that were mutated from the native VRC23 sequence were compared to known antibodies. Proteins shown are VRC23 (PDB ID: 4j6r); VRC01 (3ngb); VRC-CH31 (4lsp); 3BNC117 (4jpv); and NIH45-46 (3u7y).

IV.5, bottom left). Lastly, the D102E mutant on the CDRH3 places a carboxylic acid group in the same position as a glutamic acid on NIH45-46, improving electrostatic interactions with the antigen (Figure IV.5, bottom right). This observation is remarkable due to the fact that the antibody loops occupy different space, but redesigned residues are able to mimic the interactions of the broadly neutralizing antibody side chains. In addition, it is worthwhile to note that out of these four mutants that recapitulate known broad motifs, three were unobserved in the sequences sampled by multistate design (Figure IV.3B).

As an additional comparison, we identified 1,041 sibling sequences of known broadly neutralizing antibody VRC01, that were isolated in a previous study (Wu et al., 2015). These siblings presumably represent the sequence space accessible to VRC01, and are a good test case to compare how well our design algorithms are capturing natural sequence variation in a broad HIV antibody. Since these sequences have CDRH3 loops of different lengths we were not able to include the portion of the binding site corresponding to the CDRH3 loop – however we compared the rest of the binding site to the sequences seen in the VRC01 lineage (Figure 6). We observe that at several positions, BROAD samples sequences that are present in the VRC01 lineage but absent from MSD-sampled sequences (Figure 6, blue boxes). For example, at the third position in the binding site isoleucine is sampled at a high frequency in BROAD and VRC01 lineage sequences, but is never sampled by MSD (Figure 6). We highlight a total of five positions where BROAD is outperforming MSD in sampling sequences that are seen in the VRC01 lineage. To quantify the sequence similarity we computed a sum of squared difference between the two matrices and normalized the values to 100% (Sandelin and Wasserman, 2004; Sevy et al., 2015). According to this metric the sequences sampled by BROAD are 79.5% similar to those from the VRC01 lineage, whereas those sampled by MSD are only 76.3% similar. We conclude that BROAD more

accurately recapitulates motifs known in broadly neutralizing antibodies.

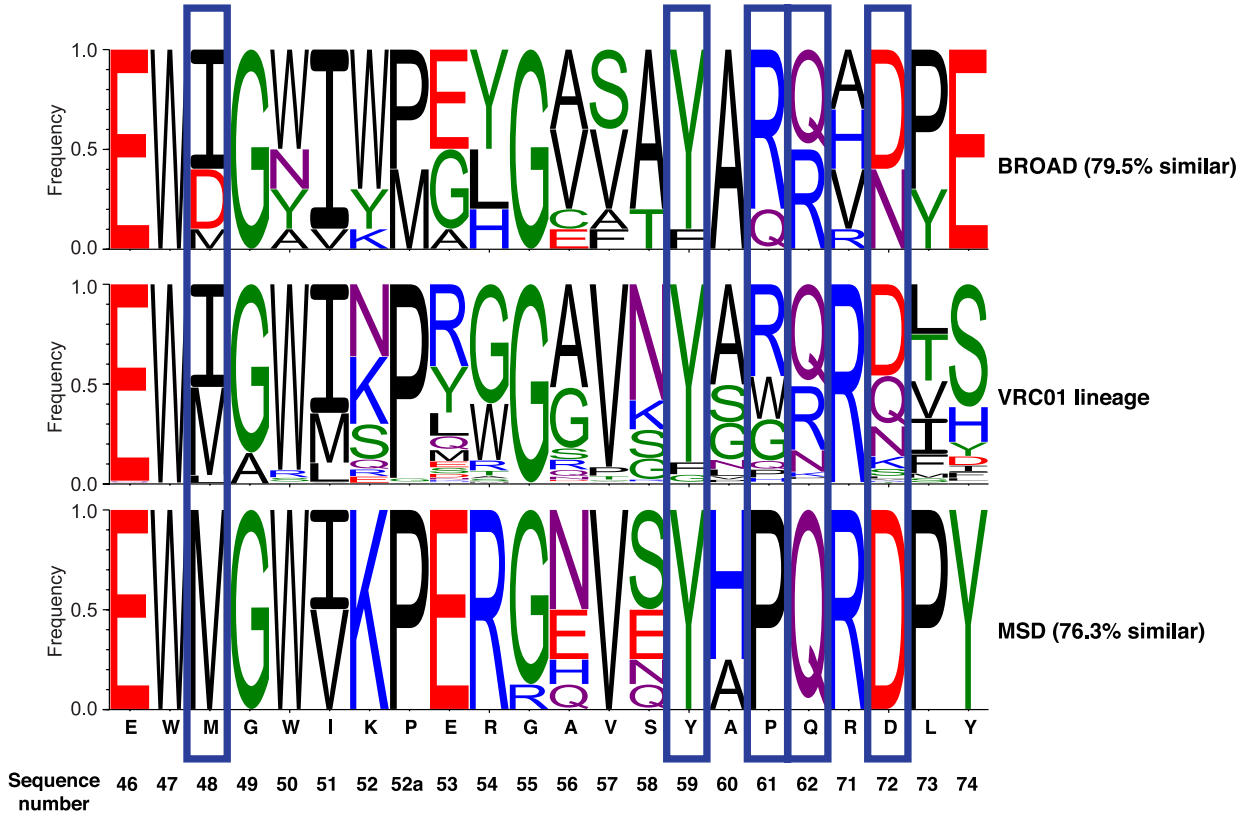


Figure IV.6. Sequences from BROAD design recapitulate sequences observed in the lineage of broadly neutralizing antibody VRC01. For BROAD and MSD sequences a percentage similarity to the VRC01 lineage was computed (similarity shown in parenthesis). Blue boxes highlight positions where BROAD samples an amino acid that is present in the VRC01 lineage but was not sampled by MSD. The VRC23 native sequence is shown below.

Discussion

Summary of results

In this chapter we describe the development of a new protein design method that we call BROAD. This method uses structural modeling with ROSETTA combined with integer linear programming optimization techniques to rapidly search through sequence space for broadly binding antibodies. We validated this method by computationally optimizing the amino acid sequence of the broadly neutralizing anti-HIV antibody VRC23. After modeling VRC23 variants *in silico* we were able to generate VRC23 variants with a predicted breadth of 100% over the simulated viral panel, compared to a predicted 53% breadth for the wild-type antibody. This outcome represents a substantial step forward in protein design, and our methodologies can be used to address a wide variety of protein design problems in which traditional structural models are insufficient.

Although we did not test antibody variants *in vitro* in this study, we predict that the computationally designed variants will have greater breadth against the HIV viral panel. However, we note several caveats with respect to experimental validation of these antibodies. Since this experiment was designed as a computational proof of principle, we modeled only the amino acids at the antibody binding interface of gp120, and not the entire gp120 sequence. This led to gp120 models with ~2 Å accuracy (Supplementary Table IV.1), which we consider sufficient for validating our design principles but not necessarily for experimental validation. Future directions in this work include optimize protocols for gp120 homology modeling to reduce this discrepancy and enable experimental validation.

Backbone optimization in protein design

A distinct advantage of the BROAD method is the ability to truly incorporate backbone movement into protein design. Many protein design methods have been developed that incorporate backbone ensembles to some degree (Allen et al., 2010; Davey and Chica, 2014; Leaver-Fay et al., 2016; Sevy et al., 2015) – however, this work typically involves either pre-generating large backbone ensembles, many of which may be redundant, or introducing backbone movement iteratively after steps of sequence design. In our approach, since we are relaxing the backbone of all mutants before fitting the sequence-based predictor, we were able to design sequences that may be slightly sub-optimal on the starting backbone coordinates, but can be highly favorable when a slight backbone relaxation is applied. This approach allows us to search sequence space that is not accessible to other methods, which are highly constrained to the initial backbone coordinates. We observed that the BROAD-generated sequences are not sampled by ROSETTA design using the RECON method, and indeed are more favorable according to the ROSETTA energy function. Therefore, we conclude that we are searching a "blind spot" in the sequence space that is missed by traditional design.

Application to HIV immunology

This approach to research could be of great utility to the field of HIV immunology. A longstanding goal of the field is discovering broadly neutralizing antibodies as the basis of a rational structure-based vaccine strategy (Huang et al., 2012; Walker et al., 2009; Wu et al., 2010). Much work has gone into redesigning existing antibodies to increase their breadth and potency (Diskin et al., 2011; Willis et al., 2015). However, HIV is known for its variability, and with this variability comes a difficulty in generating a single antibody with potent neutralization against all possible variants. The BROAD method addresses this problem by enabling rapid redesign of

known antibodies against viral panels of arbitrary size. This technology can be used in the future as part of the antibody discovery and characterization process, by rapidly searching sequence space for variants for greater breadth. In addition, protein design also has been used on the reverse side of the vaccination problem, namely, to design a vaccine with high affinity for antibodies of interest (Correia et al., 2015; Jardine et al., 2013; Ofek et al., 2010). We can foresee the application of the BROAD method to this problem as well, by optimizing immunogens for recognition of germline precursors of known broadly neutralizing antibodies.

Methods

Structural modeling

The VRC23-gp120 complex used for modeling was from the Protein Data Bank (PDB ID: 4j6r). The structure was downloaded from the PDB (www.rcsb.org) and processed manually to remove water and non-protein residues. The CH1 and CL1 domains of the antibody structure were removed from the structure manually, and the structure was renumbered starting from residue 1. To select binding sites on the antibody and virus, we applied a distance cutoff of 4 Å from the opposing protein chain, where any residue with a heavy atom within 4 Å of a heavy atom on the opposing protein was considered to be at the binding site. Distance calculations were done using PyMol visualization software (Schrodinger, LLC, 2015). We expanded this binding site to several neighboring residues to include contiguous stretches of at least four residues to constitute a binding site. A total of 27 residues on the antibody were included in the binding site. We similarly determined a viral binding site to use for structural modeling. This site included 5 contiguous stretches that were determined to be in contact with VRC23 (32 positions total). These positions were 276-282; 365-371; 425-430; 455-462; and 473-476 (HXB2 numbering). To model gp120

variants, we performed a multiple sequence alignment using ClustalW (Larkin et al., 2007) of the variant sequences with the gp120 in the crystal structure (Q23.17), and substituted the corresponding amino acids at the binding site using ROSETTA side chain optimization (Leaver-Fay et al., 2013).

Training set

To generate a training set of structural models, we made random antibody substitutions in the previously defined binding site. Each antibody variant had five randomly selected amino acid mutations. Viral variants were taken from a set of 180 known HIV gp120 sequences (Chuang et al., 2013). We chose random combinations of antibody variants and viruses, as well as the native antibody sequence with all 180 viruses, for a total of 2200 antibody-virus pairs to serve as the training set. All antibody-virus pairs were subjected to an energy minimization via the ROSETTA relax protocol, which involves iterative rounds of side chain repacking and backbone minimization with an increasing repulsive force (Combs et al., 2013). 50 models of each antibody-virus pair were generated by ROSETTA relax, and the lowest scoring model was used for further evaluation. The talaris2013 score function was used for all ROSETTA simulations.

Linear classification and regression.

Our data-driven sequence-based model to learn amino acid contributions to binding and stability is similar to the graphical model approach proposed in (Kamisetty et al., 2015). Let N_a and N_v denote the number of binding positions on the antibody and the virus respectively. Let $\mathbf{A} = \{A_1, A_2 \ldots A_{N_a}\}$ be a set of discrete variables representing the amino acids in the binding positions of the antibody. Each A_i takes values in the set of $M = 20$ amino acids. Similarly, let $\mathbf{V} = \{V_1, V_2 \ldots V_{N_v}\}$ represent the variables for the virus-binding positions. The inputs for binding

prediction are the antibody sequence $\mathbf{a} = \{a_1, a_2 \ldots a_{N_a}\}$ and virus sequence $\mathbf{v} = \{v_1, v_2 \ldots v_{N_v}\}$ where a_i and v_j are the amino acid values for the variables A_i and V_j. Amino acid contributions to binding can be modeled as a bipartite graph in which nodes for **A** and **V** represent the amino acids and the edges $\Omega \subseteq \mathbf{A} \times \mathbf{V}$ represent the pairwise amino acid interactions. Each node a_i and v_j has associated weight vector $\mathbf{x}_i$ and $y_j \in \mathbb{R}^M$. The edge (i, j) between nodes a_i and v_j has an associated weight matrix $Q_{ij} \in \mathbb{R}^{M \times M}$ to represent the position-specific contribution to binding for each amino acid pair, where q_{kl}^{um} is the umth entry of matrix Q_{ij}. Consequently, given **a** and **v**, the binding score varies as the sum of individual amino acids and pairwise interaction effects. Given this setting, **a** and **v** are predicted to bind, i.e., $\Phi(a, v) = +1\ (b(a, v) \leq \theta)$, if

$$\sum_{i=1}^{N_a}\sum_{j=1}^{M} x_{ij}a_{ij} + \sum_{i=1}^{N_v}\sum_{j=1}^{M} y_{ij}v_{ij} + \sum_{k=1}^{N_a}\sum_{l=1}^{N_v}\sum_{u=1}^{M}\sum_{m=1}^{M} a_{ku}q_{kl}^{um}\, v_{lm} + c \leq 0 \qquad (1)$$

where c is the intercept term and a_{ij} and v_{ij} are binary indicator variables that take the value 1 if amino acid j is present at position i ($\sum_j a_{ij} = 1,\ \sum_j v_{ij} = 1\ \forall\ i$). The q_{kl}^{um} term represents $Q_{kl}(u, m)$. These weights can be learned efficiently using a linear support vector machine (SVM) classifier. The feature vector **f** consists of $N_a \times M$ binary antibody features, $N_v \times M$ binary virus features and $N_a \times N_v \times M \times M$ binary pairwise interaction features corresponding to $\mathbf{x}, \mathbf{y}$ and Q respectively. Given a set of d training instance-label pairs $(\mathbf{f_i}, l_i)$, $i = 1 \ldots d$, $l_i = \{+1, -1\}$, a L2-regularized linear SVM generates a weight vector **w** by solving the following unconstrained optimization: $\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \lambda \sum_{i=1}^{d}(\max(1 - l_i\mathbf{w}^T\mathbf{f}_i, 0))^2$, where $\lambda > 0$ is the L2 regularization parameter. Smaller λ values enforce higher regularization. The second term is the squared hinge loss function. The decision function is given by sign $(\mathbf{w}^T\mathbf{f})$. We used the LIBLINEAR SVM

implementation (Fan et al., 2008) to learn the classifier. Finally, the weights **x**, **y** and Q are retrieved from the combined weight vector **w**.

On each training set of the viruses, we trained this classifier and saved the weights and the intercepts for future use in optimization. In our example, $N_a = 27$ and $N_v = 32$. To tune the regularization parameter λ of SVM, we performed 10-fold cross-validation on the full dataset, using 80% of the data for training and 20% for testing. The average prediction accuracy is shown in Figure IV.2 for different values of the L2 regularization parameter λ. As expected, higher λ values lead to overfitting. We simultaneously plot the prediction error on the two classes: binders (+1) and non-binders (-1). We chose λ = 0.001 for our experiments based on the bias-variance trade-off (corresponding to 33% test error).

The above model can be interpreted to identify the important binding positions on the antibody and the virus side, i.e., the pairs that contribute significantly to the final prediction. Specifically, we denote the Euclidean norm of the coefficient matrix of interactions Q_{ij}, for each position pair as the strength of interaction between those positions. We plot this interaction strength for each pairwise interaction in Figure IV.2C.

The linear regression model $\Psi(\mathbf{a})$ predicts the stability scores as a function of the antibody sequence features:

$$\Psi(\mathbf{a}) = \sum_{i=1}^{N_a} \sum_{j=1}^{M} x_{ij}^{s} a_{ij} + c^{s} \qquad (2)$$

where $\mathbf{x}^s \in \mathbb{R}^M$ is the weight vector in regression and c^s is the intercept. Given a set of d training instance-score pairs $(\mathbf{a}_i, s_i)\ i = 1 \ldots d$, ($s_i = s(\mathbf{a}_i, \mathbf{v}_i)$, so there are multiple scores for the same antibody feature vector), the regression objective with l_1 (sparse) regularization is given by:

$\min_{\mathbf{x}^s} \frac{1}{2d}(\| (x^s)^T \mathbf{a}_i + c^s - s_i \|_2)^2 + \alpha \| \mathbf{x}^s \|_1$, where the first term is the least squares penalty, α is the regularization parameter and $\| \mathbf{x}^s \|_1$ is the l_1-norm of the weight vector. We used the Lasso implementation in scikit-learn (Pedregosa et al., 2011) to learn this model. To measure the effectiveness of the prediction, we computed the correlation coefficient between the ROSETTA calculated stability scores (in ROSETTA energy units, or REU) and the scores predicted by regression. We performed a 10-fold cross validation experiment similar to linear classification, with 80% of the data for training and 20% for testing. Based on this parameter tuning, we chose $\alpha = 0.01$ with an average correlation of 0.85 between predicted and actual stability energy score. Again, for each training set of viruses, we learn this model and save the weights and the intercept for the optimization in the next step.

Breadth maximization integer program

We leverage the weights in the binding and stability prediction models $\Phi(\mathbf{a}, \mathbf{v})$ and $\Psi(\mathbf{a})$ to formulate an ILP for optimization in the antibody sequence space. The objective is to minimize stability score. The constraints enforce the condition that the designed antibody should bind to each virus sequence in the training set. Finally, we add the constraint that the binary variables at each antibody binding position should sum to 1, i.e., each position admits one amino acid. The ILP is given by the following:

$$\text{minimize} \quad \sum_{k=1}^{N_a} \sum_{u=1}^{M} (x_{ku}^s) a_{ku}$$

subject to

$$\sum_{k=1}^{N_a} \sum_{u=1}^{M} \left(\sum_{l=1}^{N_v} \sum_{m=1}^{M} q_{um}^{kl} \, v_{lm}^n + x_{ku} \right) a_{ku} + \sum_{i=1}^{N_v} \sum_{j=1}^{M} y_{ij} v_{ij}^n + c \leq -\epsilon \,, \qquad \forall \, n \in 1, \ldots, t$$

$$\sum_{u=1}^{M} a_{ku} = 1, \qquad \forall k,\ a_{ku} \in \{0,1\}$$

where $\epsilon = 0.0001$ (which constrains that the antibody binds to all virus variants in the dataset, with a slight margin to ensure that binding is strictly below the 0 threshold). We used CPLEX version 12.51 to solve the above ILP. We solve this optimization problem for each binding and stability model learned for data obtained from randomly chosen 100 virus variants (from the dataset in which all 180 are represented).

Non-linear classification for binding prediction

Our final step is to take 50 antibodies generated using the integer program above from 50 random subsets of data, and choose the top 10 candidates to evaluate with ROSETTA. This decision is based on a non-linear model of binding learned on the full dataset which includes all 180 viral variants, combined with a full-dataset linear model of stability. The top 10 most stable antibodies from all which are predicted to have 100% binding breadth are then chosen for evaluation. The linear model of stability is identical to what we had described above.

For the non-linear model of binding we use a kernel support vector machine with the radial basis function (RBF) kernel. This model uses the same feature set as the linear model. The kernel function enables learning in a high-dimensional, *implicit* feature space without explicitly computing the coordinates of the data in that space. The RBF kernel of two feature vectors $\mathbf{f}$ and $\mathbf{f}'$ is defined as:

$$\boldsymbol{K}(\mathbf{f}, \mathbf{f}') = \exp\left(-\frac{\| \mathbf{f} - \mathbf{f}' \|^2}{2\sigma^2}\right),$$

where $\| \mathbf{f} - \mathbf{f}' \|^2$ is the squared Euclidean distance between the two feature vectors, and σ is a

free tunable parameter. Consequently, we have two free parameters to tune: the regularization parameter λ, and the RBF kernel parameter σ. Similar to the earlier set-up, we used 80% data for training and 20% for testing in a 10-fold cross validation experiment to tune these. We performed a grid-search over all pairwise combinations of σ and λ values in 10^{-2} to 10^{2}. The LIBSVM implementation in scikit-learn was used to train the RBF SVM. We chose the model with σ = 0.01 and λ = 1 corresponding to the prediction accuracy of 68%. All learning and ILP experiments were performed on a 2.4GHz hyper threaded 8-core Ubuntu Linux machine with 16 GB RAM.

RECON multistate design

VRC23 was placed in complex with all 180 viruses and designed via RECON multistate design to increase predicted breadth across the panel. Models of viral variants were created as previously described, by substituting amino acids at the binding site. All VRC23-gp120 pairs were refined by ROSETTA relax with constraints to the starting coordinates to prevent the backbone from making substantial movements. Constraints were placed on all Cα atoms with a standard deviation of 0.5 Å. All residues at the binding site of VRC23 were included in design, for a total of 27 residues. The RECON protocol was run in parallel over 180 processors (manuscript describing parallelization in preparation), with four rounds of design and a ramping convergence constraint (Sevy et al., 2015). The binding sites on both the antibody and gp120 chain was subjected to backrub movements between rounds of design to increase sequence diversity (Smith and Kortemme, 2008). A total of 100 designs were generated. Sequences generated by both BROAD and RECON methods were visualized using the WebLogo tool (Crooks et al., 2004).

Sequence validation

To compare sequences generated by BROAD optimization and RECON multistate design, we threaded the optimized antibody sequences over the unprocessed VRC23-gp120 complexes,

and subjected these complexes to ROSETTA relax to determine the score and binding energy of optimized antibodies vs. wild-type. 50 models were generated for each complex, and the lowest scoring model was used for evaluation. To compare native and optimized VRC23 sequences, we compared the total energy of the VRC23-gp120 complex as well as the binding energy (DDG), defined below:

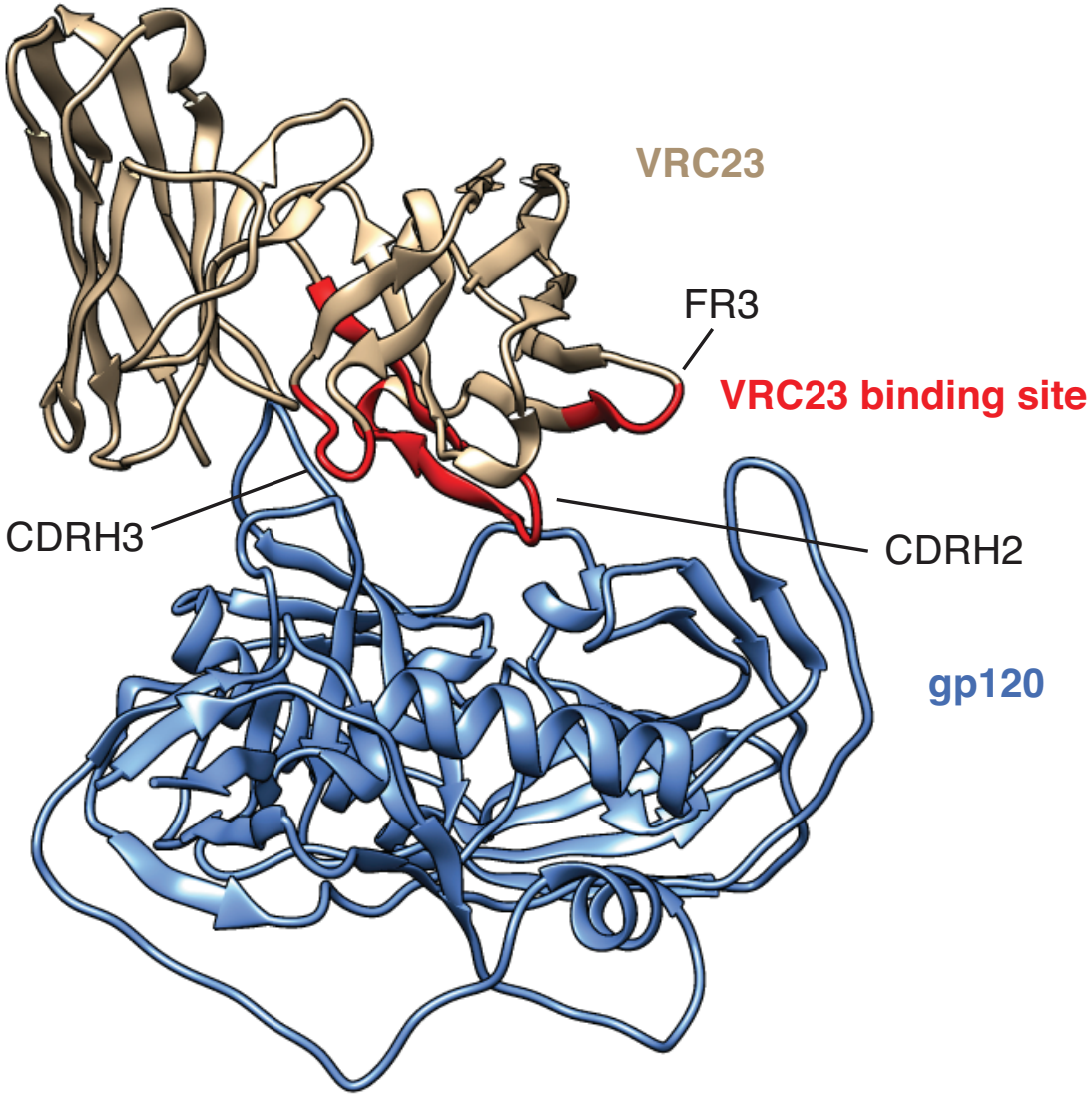
$$DDG = E_{complex} - (E_{Ab} + E_{Ag})$$

where E_{Ab} and E_{Ag} are the energies of the antibody and antigen alone, respectively. Structures of modeled VRC23-gp120 complexes were visualized using Chimera software (Pettersen et al., 2004).

Comparison to VRC01 lineage sequences

VRC01 lineage sequences were derived from a previous study (Wu et al., 2015). The 1,041 curated heavy chain sequences we used in this analysis are available in GenBank with accession numbers KP840719–KP841751. To compare sequence profiles we used a modified Sandelin-Wasserman similarity score, as described in (Sandelin and Wasserman, 2004; Sevy et al., 2015). Briefly, this score was calculated by computing the sum of squared difference for each amino acid frequency at each position, which was then subtracted from two and normalized to yield a percent similarity for each position and summed over all designed positions to give an overall similarity score.

Supplemental Information



Supplementary Figure IV.1. Binding site of VRC23 shown in context of the antibody-antigen complex. The binding site encompasses FR2, CDR2, FR3 and CDR3 regions of the antibody heavy chain.

function SolveILP
Input: linear binding and stability models
Output: optimized antibody sequence
Variables: amino acids at the antibody binding sites
Objective: maximize stability
Constraints: the antibody should bind to each virus sequence in the training set
end function

Supplementary Figure IV.2. Pseudocode describing the Integer Linear Program.

Generate Data: Rosetta(virus panel,antibody variants)
Learn Models: binding Φ and stability Ψ on all data
Choose 50 random subsamples of 100 viruses
for each random subsample of 100 viruses **do**
Learn linear binding and stability models
SolveILP
Evaluate breadth (against full panel) using Φ and Ψ
end for
Choose top optimized antibody candidates
Evaluate: Rosetta structure modeling on full panel

Supplementary Figure IV.3. Pseudocode describing the BROAD algorithm for design of broadly binding antibodies.

Supplementary Table IV.1. Deviation of ROSETTA relaxed gp120 models from the starting crystal structures.

gp120 for indicated HIV strain	**PDB ID**	**RMSD between relaxed model and crystal structure**
Q23-17	4j6r	1.4
YU2-DG	3tgq	2.1
Du172-17	5te7	2.8
RHPA-7	5t33	2.2
X2088-c9	5te4	2.5
ZM109-4	3tih	1.9
JRCSF-JB	4r2g	2.4
HXB2-DG	1g9m	2.5
Q842-d12	4xmp	2.4
Average		**2.2**

ROSETTA relaxed models used in BROAD optimization were compared to solved structures of gp120 viral variants and the root mean squared deviation (RMSD) was computed over Cα atoms on gp120. The relax protocol recapitulates the gp120 conformations with an average RMSD of 2.2 Å.

CHAPTER V.

Engineering cross-reactivity to influenza and HIV antigens

Sevy A.M., Parrish E.H., Chapman N., Bombardi R., Soto C., Meiler J., Crowe J.E. Jr.

Unpublished.

Author contributions: I proposed the model of cross-reactivity discussed in this chapter, and performed all computational modeling experiments, under the mentorship of Jens Meiler and James Crowe. Next-generation sequencing was performed in collaboration with R.B. for data collection and C.S. for data processing. E.H.P. and N.C. assisted with experimental design and testing.

Introduction

Influenza and human immunodeficiency virus (HIV) are two pathogens that create an enormous public health burden worldwide. Although very different in their viral characteristics and replication cycles, they share similarities in that they both undergo rapid cycles of evolution and antigenic shift to evade the host immune response. This rapid shift challenges the immune response by generating many viral variants that evade recognition by circulating antibodies, creating difficulty in effective neutralization of multiple strains simultaneously. In spite of this variability, broadly neutralizing antibodies recognizing diverse viral variants can be generated by adaptation of strain-specific antibodies via somatic hypermutation, conferring broad recognition and neutralization.

Multiple neutralizing human antibodies targeting the receptor-binding site (RBS) of influenza hemagglutinin (HA) and the membrane-proximal external region (MPER) of HIV envelope protein gp41 have been characterized and structurally determined. Antibodies targeting the influenza RBS neutralize their targets by mimicking the binding position of the cellular receptor, sialic acid, preventing host cell recognition and entry (Hong et al., 2013; Schmidt et al., 2013; 2015). MPER-targeting antibodies recognize a conserved region of HIV gp41, in many cases neutralizing large swaths (98%) of viral isolates, and are thought to neutralize by interfering with membrane fusion (Huang et al., 2012; Song et al., 2009). Both of these sites are susceptible to targeting by broadly neutralizing antibodies, as they both perform functions critical to the viral replication cycles (recognition of host cell receptor and membrane fusion, respectively). Antibodies recognizing either of these two antigenic sites have surprisingly homologous conformations of the heavy chain complementarity determining region 3 (CDRH3) loop, with root mean square deviation (RMSD) as low as 3.3 Å between disparate antibodies (Pejchal et al., 2009; Schmidt et al., 2013; Stanfield et al., 2011). Since the CDRH3 loop is the principal mediator of binding to influenza RBS, anti-HIV antibodies with a homologous loop conformation could, in theory, also accommodate binding to influenza RBS.

In light of the structural similarity observed between antibodies targeting influenza HA and HIV gp41, I propose an immune mechanism wherein previously mutated influenza antibodies are stimulated by HIV gp41 upon infection and "repurposed" to potently neutralize HIV. I hypothesize that influenza HA can serve as an intermediate stimulating antigen in the development of an anti-gp41 response, implying that antibodies isolated against influenza HA can be mutated through affinity maturation to recognize HIV gp41, and vice versa. To test this hypothesis, I used computational modeling and design using the ROSETTA software suite to simulate affinity

maturation and reconstruct hypothetical intermediates in this maturation process, using existing mature antibodies as templates and developing variants with altered specificity.

Results

Rationale

Based on structural similarities between mature HIV and influenza antibodies, I hypothesize that naïve B cells undergo limited maturation in response to influenza HA, and are later matured to bind gp41 upon HIV infection (Figure V.1). In this model naïve B cells are stimulated by exposure to influenza early in life, inducing somatic hypermutation which stabilizes the CDRH3 loop into a conformation capable of binding both HIV and influenza antigens. These partially mutated intermediate antibodies can then be recruited to respond to subsequent exposure influenza, or can be mobilized to target HIV antigens upon infection.

To date there are 13 antibodies deposited in the Protein Data Bank (PDB) that target the influenza HA RBS, which function by interfering with recognition of the host cell receptor sialic acid (Ekiert et al., 2009; 2012; Hong et al., 2013; Lee et al., 2014; 2012; Schmidt et al., 2013; 2015; Thornburg et al., 2013; Whittle et al., 2011; Xu et al., 2013). Since binding by many such antibodies is primarily mediated by insertion of the CDRH3 loop, an antibody with a stable CDRH3 conformation similar to that of the known anti-RBS antibodies should be compatible with binding to the RBS. Several anti-HIV MPER antibodies bind their target peptides with similar CDRH3 conformations as anti-influenza RBS antibodies. I identified a cluster consisting of CDRH3 loops from five anti-RBS and three anti-MPER antibodies that assume similar bound conformations (Allcorn and Martin, 2002) (Table V.1). The loop RMSD of length-matched pairs of anti-RBS and anti-MPER antibodies ranged from 3.3-5.2 Å, whereas the same measurement

between pairs of anti-RBS antibodies ranged from 0.5-3.5 Å, suggesting that these deviations are within range for antibodies targeting a common epitope. Additionally, all of the CDRH3 loops are accommodated within the recessed binding pocket of HA without creating excessive clashes with HA residues. The structural similarity between these two CDRH3 conformations coincides with an approximate co-localization of helical motifs on the surface of influenza HA and HIV MPER when superimposing the antibody-antigen co-crystal structures, suggesting that this CDRH3 conformation may be a common structural solution to binding short helical motifs.

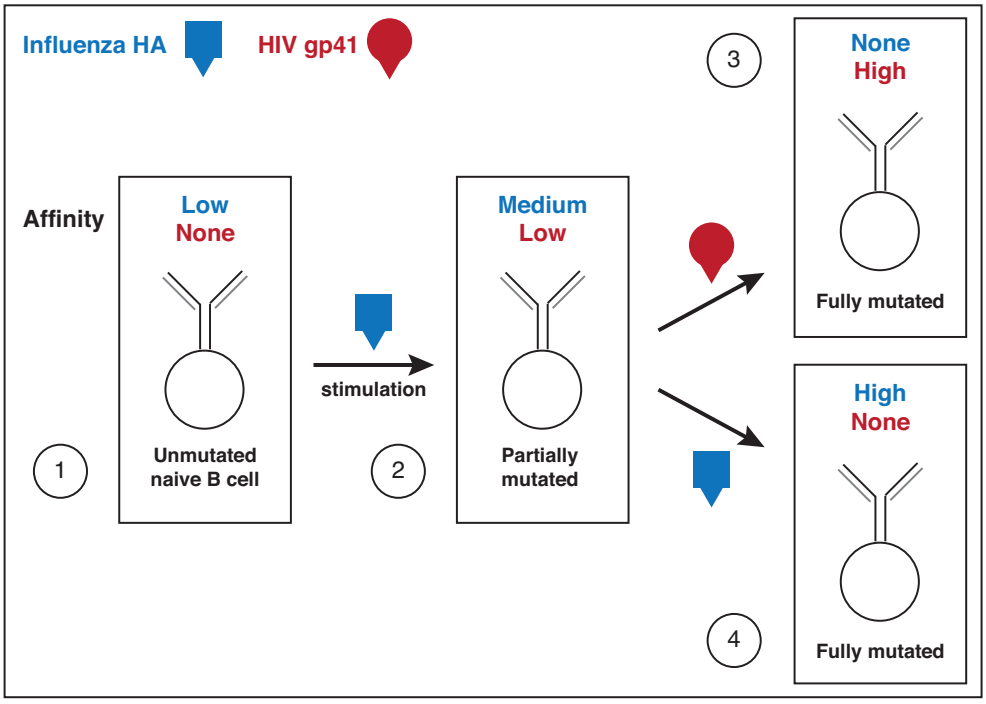


Figure V.1. Proposed model of cross-reactivity in the antibody response to HIV and influenza. Structural similarities between mature HIV (3) and influenza (4) antibodies combined with the lack of recognition of HIV Env by germline-reverted HIV antibody ancestors (1) lead to the hypothesis that an intermediate antibody undergoes limited maturation in response to influenza (2) and is later stimulated in response to HIV infection.

Experimental workflow

To test this hypothesis, we took an approach of next-generation sequencing combined with computational modeling to identify human B cells predicted to cross-react with both HIV and influenza antigens (Figure V.2). We first sequenced antibody heavy chain variable genes from HIV-infected donors after influenza vaccination. After obtaining these sequences, we used a computational protocol in ROSETTA known as a position-specific structural scoring matrix (P3SM) to predict the interaction of an antibody sequence with either influenza HA or HIV MPER peptide. The P3SM protocol uses ROSETTA modeling of a small subset of antibody sequences to extrapolate for rapid scoring of large numbers of sequences (Willis et al., 2016). After predicting the activity of over 140,000 human sequences, we identified those with the highest predicted likelihood of cross-reactivity between influenza and HIV and subjected them to multistate design using the RECON algorithm to increase affinity for both targets.

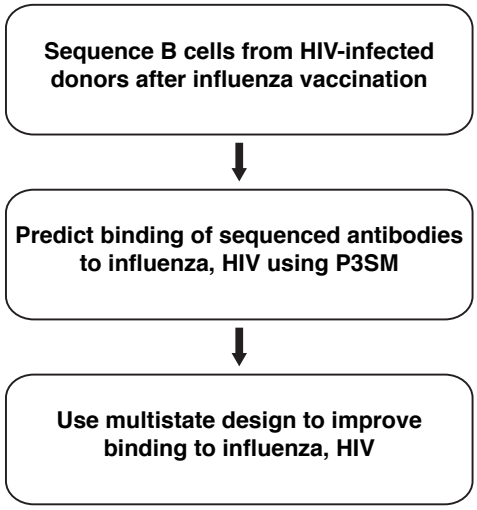


Figure V.2. Experimental workflow of identifying antibodies cross-reactive to influenza and HIV. We first collected B cells from HIV-infected human donors after seasonal influenza vaccination. We then used a computational protocol called a P3SM to predict whether a given sequenced antibody would bind to either influenza or HIV. After predicting the binding activity, we used multistate design to further increase affinity of sequenced antibodies for both influenza and HIV.

Computational modeling using ROSETTA

To perform P3SM prediction of influenza- or HIV-binding activity, I first identified templates for modeling, which would serve as a framework for threading the human sequences. I chose two antibodies with previously determined structures, anti-influenza mAb 641 I-9 and anti-HIV mAb Z13e1, both encoded by IGHV gene V_H4-59 with 19 amino acid CDRH3 loops (Nelson et al., 2007; Schmidt et al., 2015; Zwick et al., 2001). Both of these antibodies neutralize their targets (influenza or HIV, respectively) with high potency and breadth. To predict the likelihood of an antibody sequence interacting with either influenza or HIV, I threaded the CDRH3 sequence of a given antibody sequence onto the CDRH3 conformation of either 641 I-9 or Z13e1 and measured the ROSETTA energy score after structural refinement. This energy score gives a measure of the likelihood of a new antibody sequence adopting the active, antigen-binding conformation as observed in the template structure. The P3SM is a useful heuristic because it uses a training set of ROSETTA-modeled sequences to learn a simplified energy function that can rapidly predict ROSETTA score from sequence alone. As a training set, I generated models of 500 randomly selected CDRH3s from human donors threaded over either the 641 I-9 or Z13e1 co-crystal structure. I then fit a linear regression model to predict the ROSETTA score from sequence with high accuracy.

Next-generation sequencing

We collected peripheral blood mononuclear cells (PBMCs) from five HIV-infected donors seven days after seasonal influenza vaccination to validate our hypothesis. We reasoned that these donors would have the highest frequency of cross-reactive B cells, as they have memory B cells specific for HIV that could potentially be recruited to respond to influenza vaccination. Temporally, this is the opposite of the hypothetical scenario proposed earlier in this chapter, where

individuals with previous exposure to influenza use this antibody response to target HIV. We sequenced the antibody heavy chain of plasmablasts from these donors, obtaining a total of 2,241,196 unique clonotypes (Table V.2). We filtered the sequences to those which use the same IGHV gene (V_H4-59) and have the same CDRH3 length (19 amino acids) as the template mAbs, which yielded 142,716 sequences for P3SM analysis. I scored each of these sequences for likelihood to bind influenza based on the 641 I-9 structure and likelihood to bind HIV based on the Z13e1 structure and plotted the two scores to identify potential cross-reactive sequences (Figure V.3). As a control, I also modeled the wild-type CDRH3 sequences of the template mAbs Z13e1 and 641 I-9. Each of these template sequences scored favorably in their cognate structures, but not the opposing complex. This reflects the known behavior of these antibodies, which is specific to one antigen with no cross-reactivity to the other. The donor sequences show a range of P3SM scores between the native complexes of each of the two antigens. No donor sequence was predicted to bind to both antigens with comparable affinity to the wild-type template mAbs, but several sequences are comparable to the templates in P3SM score for one antigen or the other. The P3SM score distributions suggest that when transitioning from one antigen to the other an affinity tradeoff is necessary, and a cross-reactive antibody capable of recognizing both targets would do so with low affinity.

Structural analysis of models

I next analyzed the distribution of P3SM scores of human sequences against both antigens to identify those which are Pareto-optimal. Pareto analysis is used to find the optimal combination of two or more variables (Nivón et al., 2013). The Pareto-optimal frontier is the set of data points where one variable cannot be improved without a sacrifice in the other variables. In this case, I wanted to identify the optimal set of HIV or influenza binding antibodies, where improvement in

HIV affinity can only be achieved with a loss in influenza affinity, or vice versa. I calculated this set of Pareto-optimal sequences within an error of 5 ROSETTA energy units (REU), which yielded a total of 367 sequences, and analyzed the characteristics of these clones (Figure V.4A). These sequences were very degenerate in their CDRH3 sequences, with the only consistent pattern an abundance of tyrosine at the C-terminal end of the loop, likely due to use of the IGHJ6 gene. I then generated ROSETTA models of these Pareto optimal sequences to test the accuracy of the P3SM prediction (Figure V.5). The ROSETTA score was well predicted by P3SM and most clones remained in the range of affinities between the wild-type antibodies. From analysis of models of individual sequences, I observed recapitulation of the binding modes of wild-type 641 I-9 and Z13e1 (Figure V.4B). Sequence 12827-37812 was able to mimic the 641 I-9 binding mode by placing an aspartic acid next to a hydrophobic residue at the tip of the CDRH3, a common dipeptide motif in RBS which mimics sialic acid (Schmidt et al., 2015). In addition, sequence 3610-63118 is able to recapitulate the binding mode of Z13e1 to its HIV epitope by placing a hydrophobic patch in the groove corresponding to HIV MPER binding.

These computational results suggest that naturally occurring human antibodies are able to recognize HIV and influenza antigens with a gradient of affinities, with some presumably binding both targets with low affinity. I next asked whether these sequences could be improved by computational design for binding to both targets. I applied the RECON multistate design method to these sequences and measured their predicted affinity for HIV and influenza after redesign (Sevy et al., 2015). These sequences are predicted to be improved in affinity both for HIV and influenza antigens after redesign (Figure V.5). The wild-type 641 I-9 antibody was greatly improved for predicted HIV affinity after RECON redesign without any loss in influenza affinity. The wild-type Z13e1 antibody could be improved substantially for influenza affinity with a slight improvement

for HIV affinity. These computational experiments suggest that it is possible to engineer an antibody which binds both HIV and influenza antigens by redesign of existing human antibodies. In addition, this analysis suggests that circulating human antibodies in HIV-infected donors may have low to moderate affinity for influenza as well as HIV.

Table V.1. Anti-influenza HA and HIV MPER antibodies identified with similar CDRH3 conformations.

Antibody	Antigen	V_H gene	J_H gene	CDRH3 length
5J8	HA	4-38-2	4	17
CH65	HA	1-2	6	19
CH67	HA	1-2	1	19
H5.3	HA	4-4	5	16
641 I-9	HA	4-59	3	19
Z13e1	MPER	4-59	6	19
3D6	MPER	3-9	3	19
10e8	MPER	3-15	1	22

Table V.2. Primary blood mononuclear cells (PBMCs) were collected from five HIV-infected donors after influenza vaccination for next-generation sequencing.

Donor ID	Gender	Race	Ethnicity	Age (years)	Site of collection	# unique clonotypes
24407	M	Caucasian	Non-Hispanic	51	Nashville, TN	402,925
24408	M	Caucasian	Non-Hispanic	57		419,370
24409	M	Caucasian	Hispanic	32		652,153
24410	M	African-American	Non-Hispanic	51		450,526
24411	M	Caucasian	Non-Hispanic	49		316,222

Clonotypes are defined as antibody clones that share IGHV and IGHJ genes and 100% amino acid identity in the CDRH3 region. Antibody sequences were filtered for quality and grouped into clonotypes before de-duplication. For more details on quality filtering see Methods.

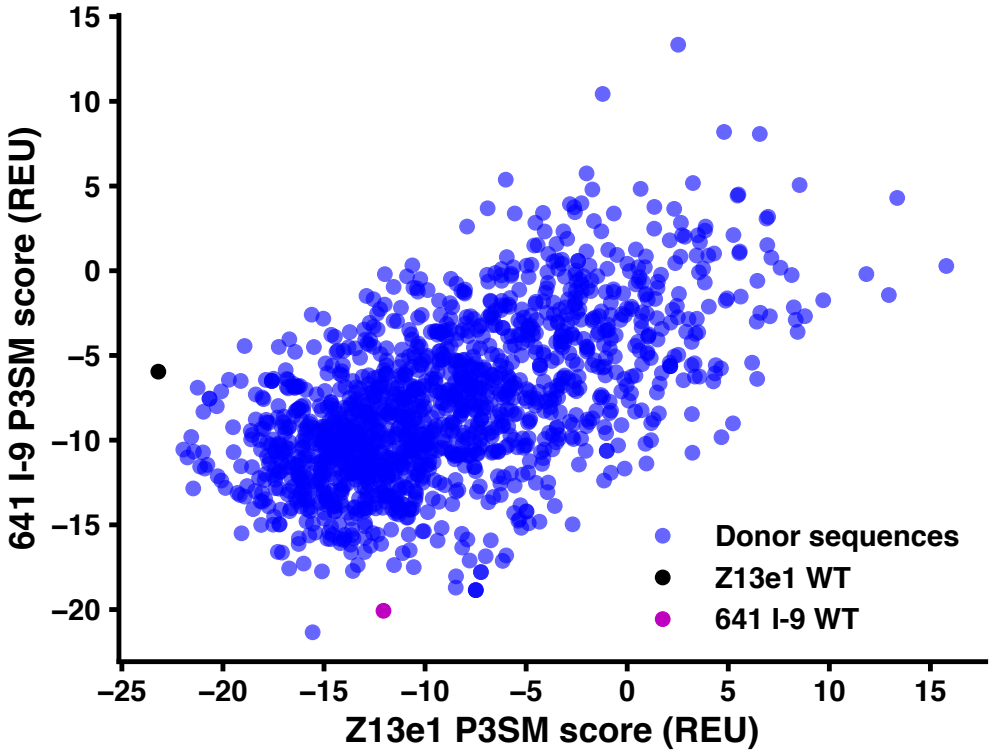


Figure V.3. Cross-reactive sequences evaluated by position-specific structural scoring matrix (P3SM). Human antibody sequences from HIV-infected donors after seasonal influenza vaccination were predicted by P3SM for likelihood to adopt the CDRH3 conformation of anti-HIV mAb Z13e1 (X-axis) or anti-influenza mAb 641 I-9 (Y-axis). P3SM score is expressed in ROSETTA energy units (REU). Also shown are the wild-type (WT) CDRH3 sequences of Z13e1 and 641 I-9 subjected to the same protocol.

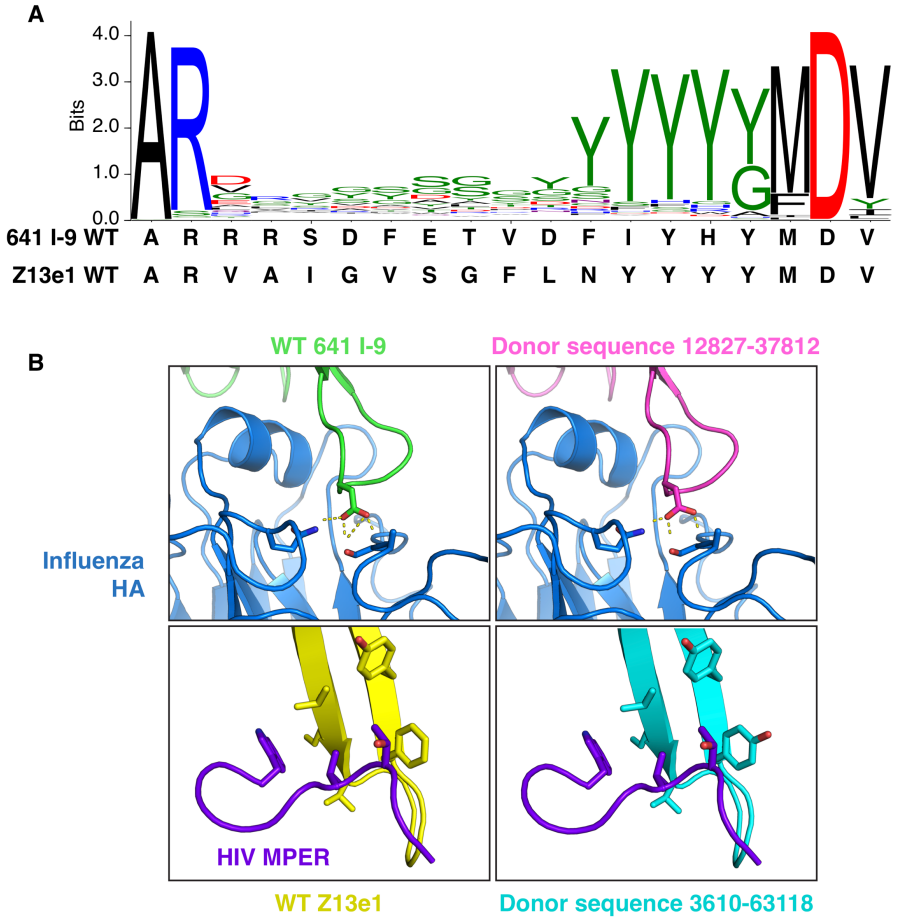


Figure V.4. Properties of putative HIV/influenza reactive antibodies identified by P3SM approach. A. Sequence logo of 367 CDRH3 sequences with Pareto optimality for 641 I-9 and Z13e1-like CDRH3 conformations. Amino acids are colored according to chemical composition. Wild-type sequences of 641 I-9 and Z13e1 CDRH3 loops are shown below. B. Example donor sequences that are predicted to mimic activity of wild-type antibodies.

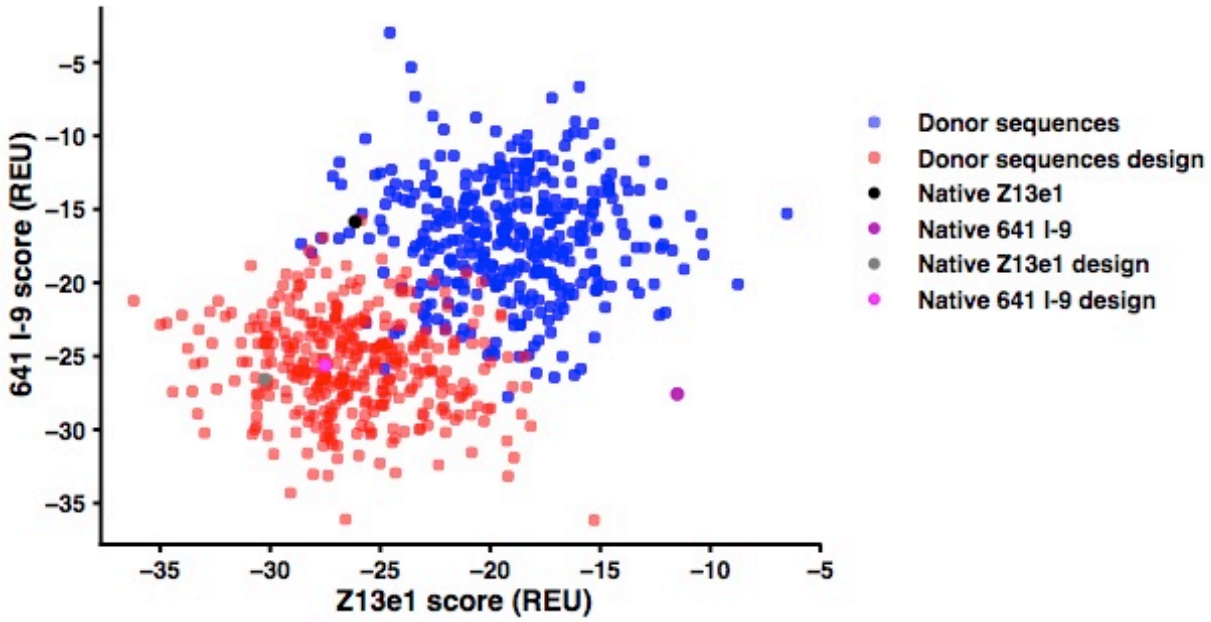


Figure V.5. CDRH3 sequences from infected donors can be improved by multistate design for binding to both HIV and influenza. Donor CDRH3 loops identified by the P3SM protocol were homology modeled in ROSETTA and assessed for favorability in Z13e1 HIV-binding complex (X-axis) or 641 I-9 HA-binding complex (Y-axis). These sequences were redesigned by RECON multistate design and the scores are plotted (red). Also shown are the wild-type Z13e1 and 641 I-9 CDRH3 loops, subjected to the same protocol of homology modeling (black and magenta, respectively) or multistate design (gray and pink, respectively). Scores are expressed in ROSETTA energy units (REU).

Discussion

In this chapter, I proposed an immune mechanism in which antibodies are first raised in response to influenza, undergo limited maturation to rigidify the combining site, and later respond to HIV infection by undergoing somatic hypermutation to gain high affinity for the HIV MPER epitope. I approached this hypothesis from a structural perspective, first identifying clusters of antibodies targeting influenza RBS or HIV MPER with homologous CDRH3 conformations, then using computational modeling to predict their tendency for cross-reactivity between antigens. I used next-generation sequencing to identify heavy chain antibody sequences of memory B cells from HIV-infected donors after seasonal influenza vaccination, and used a modeling heuristic known as a P3SM to identify sequences predicted to have cross-reactive activity against HIV and influenza. Lastly, I used RECON multistate design in ROSETTA to optimize the CDRH3 loops of these human sequences against both antigens simultaneously and engineer antibodies with predicted high affinity against both of these targets.

Although the work presented in this chapter is promising, several caveats should be noted. The antibody sequences described were only evaluated *in silico* and were never tested experimentally to verify their binding activity. This is due to limitations in high-throughput antibody synthesis, expression, and testing necessary to validate a large number of recombinant antibodies. However, as high-throughput gene synthesis and recombinant antibody expression becomes more routine, in the future it will be possible to test a large number of these sequences and verify their binding activity.

Another caveat is based on the format of the sequences modeled in this work. Using the P3SM method, only the predicted binding of variant CDRH3s presented on the backbone of known wild-type antibodies (641 I-9 and Z13e1) was measured. However, it would be more biologically

relevant to test whether a CDRH3 loop in the context of its native antibody framework sequence is capable of binding a given antigen. Since the CDRH3 is the primary mediator of antigen recognition in the case of influenza HA, and a key contributor in the case of HIV MPER, I argue that modeling the CDRH3 is an adequate proxy for predicting activity of a given sequence. In addition, the fact that the antibodies are both encoded by V_H4-59 implies that the framework regions will be mostly similar in sequence, making the CDRH3 the main differentiator between influenza and HA binding. However, the difference between modeling chimeric antibodies with a mutated CDRH3 and full-length sequences is an important distinction and something that should be addressed in future work. In addition, all donor sequences were modeled with the light chain from one of the two wild-type antibodies previously mentioned. As the next-generation sequencing only gathered heavy chain sequences, we are unable to model native heavy-light chain pairs. However, recent work has shown it possible to perform high-throughput next-generation sequencing with heavy-light chain pairs (Wang et al., 2018), and we expect these advances to greatly benefit antibody modeling process.

In the immunological model presented here, I hypothesized that an intermediate antibody would have low affinity for both HIV and influenza, and after undergoing further maturation it would gain affinity for one antigen at the cost of the other. The P3SM screening of donor sequences fits this model by showing that, among circulating, 641 I-9 and Z13e1 are optimized for binding to their antigens, and cannot be substantially improved for the other antigen without sacrificing affinity to its own antigen (Figure V.3). The implications of this hypothesis are that such low affinity intermediates could be targeted in an HIV vaccination strategy. If the intermediates are induced by exposure to influenza, they should be common in the population and should be present at an early age. They could be targeted by an HIV vaccine that spurs their development towards

HIV recognition, such that upon HIV infection there is already a potent response in memory. This would extend on work done by other groups targeting precursors of broadly neutralizing HIV antibodies for a more effective vaccine response (Jardine et al., 2013; 2015). However, a key detail of the mechanism proposed in this chapter is whether or not these intermediate antibodies have potent neutralizing activity. If targeting these intermediates leads to a high affinity but nonneutralizing response, targeting them would be counter-productive since it would drive the immune response towards nonneutralizing antibodies. In this case it would be desirable to avoid these intermediates to avoid the nonneutralizing response. However, since Z13e1 neutralizes HIV potently and with relatively high breadth (Nelson et al., 2007; Zwick et al., 2001), I hypothesize that the intermediate antibodies induced by this mechanism would also be neutralizing after affinity maturation.

Methods

Sample preparation and sequencing

Peripheral blood was obtained from healthy adult donors following informed consent, under a protocol approved by the Vanderbilt Institutional Review Board. B cells from approximately 1 x 10^7 PBMCs per donor sample were enriched using EasySep Human Pan-B Cell Enrichment Kit on the RoboSepTM-S according to the manufacturer's protocol (Stemcell Technologies). After the enrichment, cells were washed and pelleted for total RNA extraction using the RNeasy Mini Kit (Qiagen). First-strand cDNA synthesis was performed by using PrimeScript Reverse Transcriptase (Clontech), following the manufacturer's instructions (with optional steps), using 20 pmol of J gene-specific primers (van Dongen et al., 2003)with unique molecular identifiers incorporated into the 5' end of the primers (Khan et al., 2016). After cDNA

synthesis, samples were purified using the AmpureXP Size Select Bead Kit (Beckman Coulter). Immediately following bead clean up, 30 μL of PCR mixture containing 2.5 pmol of each V gene-specific region primer (van Dongen et al., 2003) and 2X Kapa Hifi Hotstart Ready Mix (Kapa Biosystems) was added directly to the 20 μL purified first-strand synthesis product. PCR reaction conditions were 95°C for 3 min, 9 cycles of 98°C for 20 s, 65°C for 15 s, and 72°C for 30 s, and a final extension step of 72°C for 5 min. The first-round PCR reaction was purified using the Ampure Size Select Bead Kit (Beckman Coulter). Second-round PCR mixture containing 25 pmols of each Illumina adapter extension primer and 2X Kapa Hifi Hotstart Ready Mix (Kapa Biosystems) was added directly to 20 μL of the purified first-round PCR reaction product. PCR reaction conditions were 95°C for 3 min, 23 cycles of 98°C for 20 s, 65°C for 15 s, and 72°C for 20 s, and a final extension step of 72°C for 5 min. The second-round PCR products were purified using the Ampure Size Select Bead Kit (Beckman Coulter). Illumina-ready amplicon libraries were quantified using the Real-time Library Amplification Kit (Kapa Biosystems) and pooled at equimolar amounts. Samples were loaded onto 2X flow cells for sequencing on the HiSeq 2500 next-generation sequencer with PE-250 V2 chemistry (Illumina).

Data processing and analysis

The processing pipeline consisted of the following steps. First, the FASTQC toolkit was used to make a visual inspection of the quality of the run (Andrews). Next full-length reads were generated from Illumina paired end reads using the software package USEARCHv9.1 (Edgar and Flyvbjerg, 2015); 3) The BIOMEDII primers, as in (Soto et al., 2018a), were removed using the software package FLEXBARv3.0 (Roehr et al., 2017). Data was then processed using PyIR software to assign germline V and J genes (Soto et al., 2018b). Sequences were filtered by the following criteria: 1) productive sequences, 2) V and J gene E value less than 10^{-6}, 3) in-frame

sequence, and 4) no unknown amino acids in protein sequence. Sequences were then reduced to clonotypes, where a clonotype is defined as a set of sequences with a common V and J gene and an identical CDRH3 amino acid sequence. Clonotypes were de-duplicated and the frequency of each V-J gene pair was computed for each individual donor.

ROSETTA modeling

The co-crystal structures of 641 I-9 and Z13e1 in complex with their antigens were downloaded from the Protein Data Bank (PDB IDs 4yk4 and 3fn0, respectively). The structures were processed manually to remove waters and non-protein residues. The heavy chain constant region 1 (C_H1) and light chain constant region (C_L) domains of the antibodies were removed from the structures manually, and the structures were renumbered starting from residue 1. To generate a P3SM, 500 CDRH3 sequences from the previously described donors were randomly selected and threaded over the CDRH3 loop of either the Z13e1 or 641 I-9 complex, along with the wild-type sequence of each antibody. The chimeric structure was then refined with ROSETTA relax using constraints to the starting coordinates to prevent the backbone from making substantial movements. Constraints were placed on all Cα atoms with a standard deviation of 1.0 Å. Ten models were generated for each modeled sequence and a total of 5,020 models were used to generate each P3SM. To generate a P3SM ridge regression was used to fit a coefficient to each of the 20 amino acids at 19 positions of the CDRH3 loop, with an l2 penalty to enforce sparsity. The coefficients were fit to optimize prediction of ROSETTA score of the CDRH3 loop from the sequence. The Python package scikit-learn was used to perform ridge regression (Pedregosa et al., 2011).

CHAPTER VI.

Computationally designed cyclic peptides derived from an antibody loop increase breadth of binding for influenza variants

Sevy, A. M., Gilchuk I.M., Nargi R., Jensen M., Meiler J., Crowe J.E. Jr. Computationally designed cyclic peptides derived from an antibody loop increase breadth of binding for influenza variants. Manuscript in preparation.

Author contributions: I performed all computational experiments described in this chapter under the mentorship of Jens Meiler and James Crowe. I also performed all binding experiments described herein. I was responsible for experimental design, analyzed data with my co-mentors, and created all figures in this chapter. I collaborated with I.M.G. for hemagglutination inhibition assays. R.N. and M.J. assisted in protein expression and purification.

Abstract

The influenza hemagglutinin (HA) glycoprotein is the target of many known broadly neutralizing antibodies. However, influenza viruses can rapidly escape antibody recognition by mutation of hypervariable regions of HA that overlap with the antibody binding epitope. In this work, we hypothesized that by designing peptides to mimic antibody loops, we could enhance breadth of binding to HA antigenic variants by reducing contact with hypervariable residues on HA that mediate escape. We designed cyclic peptides that mimic the heavy chain complementarity-determining region 3 (CDRH3) of anti-influenza broadly neutralizing

monoclonal antibody C05 and show that these cyclic peptides bound to HA molecules with < 100 nM affinity, comparable to that of the full-length parental C05 IgG. In addition, these peptides exhibited increased breadth of recognition to influenza H4 and H7 subtypes by eliminating clashes between the hypervariable antigenic regions and the antibody CDRH1 loop.

Introduction

Since the mid-1990s, the role of monoclonal antibodies (mAbs) as therapeutic agents has been growing rapidly. MAbs have many favorable properties as therapeutics, as they can be generated quickly by a number of different discovery strategies, and it is possible to isolate mAbs that bind to virtually any target of interest. To date, over 60 antibody-based drugs have been approved for therapeutic use, and over 550 antibodies are currently being developed as therapeutics (Carter and Lazar, 2018). Part of the appeal of antibodies as drugs is the fact that they bind targets with both high affinity and remarkable specificity.

However, antibody therapeutics also face major limitations. As large (~ 150 kDa) biological molecules, they can be difficult to produce and store compared to small molecules. MAbs are susceptible to degradation and modification over time, and the cost of production is high (Elgundi et al., 2017). Antibodies typically are effective only against targets accessible on the surface of microorganisms or cells, as they cannot cross cell membranes efficiently (Carter and Lazar, 2018). In addition, in cases where antibodies are needed to recognize multiple related proteins with sequence polymorphisms, such as viral surface proteins, the specificity that is a cardinal feature of the antibody-antigen interaction can be viewed as a limitation.

To address these limitations, we developed a new method of generation of mini-antibodies using computational design of cyclic peptides based on the structure of an antibody hypervariable

loop. Antibodies bind their targets with a surface comprising six complementarity-determining regions (CDRs), the most diverse of which is the third loop on the heavy chain (CDRH3) (Murphy et al., 2012). Potent and cross-reactive antiviral antibodies commonly have long CDRH3 loops that mediate antigen recognition (Corti and Lanzavecchia, 2013), a trend that is also evident for other non-viral targets (Bonsignori et al., 2014; Shih et al., 2012; Thomas, 1993). Long CDRH3 loops typically contact a conserved portion of the antigen and can mediate broad recognition. However, the large overall footprint of such antibodies often includes contacts made by CDRH1 and CDRH2 and by light chain CDRs. If antigen residues in the periphery of the contact area are hypervariable (*i.e.*, vary among viral field strains), minor structural variations in those strains result in loss of binding and thus prevent broad reactivity for that mAb. We hypothesized that we could enhance the breadth of activity of a neutralizing antibody against a hypervariable viral protein by designing a peptide that adopts the conformation of the antibody CDRH3 loop but eliminates interactions with variable antigenic residues. As a proof-of-principle exercise, we used computational modeling to design a peptide derived from anti-influenza human mAb C05 whose CDRH3 binds to the highly conserved receptor-binding site on hemagglutinin (HA) (Ekiert et al., 2012). We show that this peptide binds with high affinity to influenza HA and exhibits increased breadth, as it binds to viral subtypes not recognized by the parental antibody, due to the highly conserved footprint of the peptide.

Results

Experimental workflow

To design antibody loop-based peptides, we performed computational modeling simulations using the ROSETTA software suite (Alford et al., 2017) (Figure VI.1). We focused our

efforts on anti-influenza mAb C05, as this mAb binds its antigen primarily with a long (26 amino acid) CDRH3 loop. We first removed the CDRH3 loop from the structure of C05 in the unbound state (PDB ID 4fnl) and added cysteines to the N- and C-termini *in silico* (Figure VI.1A). We reasoned that a disulfide-stabilized cyclic peptide would have reduced conformational entropy compared to a linear peptide and would more readily adopt the correct conformation (Bogdanowich-Knipp et al., 1999). We used ROSETTA to fold these peptides 1,000 times, modeling either the full peptide or a truncated version of the tip of the loop, and analyzed the folding energy landscape of the peptides, where a peptide likely to fold into its active conformation had low scoring decoys with low Cα RMSD (Figure VI.1B, lower left quadrant). After folding the wild-type peptide sequence, we used ROSETTADESIGN to optimize the peptide sequence to increase stability in the active conformation. We then refolded the sequence-optimized peptides to assess the likelihood that these spontaneously fold into the bioactive conformation and cannot adopt low energy alternative conformations. Based on the computational analysis we selected eight redesigned peptides for synthesis and experimental characterization (Figure VI.1C).

Peptide folding simulations

Starting with the full-length and truncated CDRH3 peptides, we folded these peptides *in silico* to predict their low energy conformations (Figure VI.1B). The peptides failed to converge on the active conformation, suggesting that alternative low energy conformations exist that would presumably not function. Notably, this finding agrees with previous experiments showing that the C05 CDRH3 peptide is not active in cyclized or linear formats (Ekiert et al., 2012), effectively serving as a negative control for our computational protocol. Next, we redesigned the group of peptide models within 2 Å of the native conformation, consisting of 36 full-length and 49 truncated models, and refolded the top 10 and 23 scoring designs, respectively. Out of these models, three

full-length and five truncated peptides displayed an improved folding profile (Supplementary Figure VI.1). These variants featured a decreased overall ROSETTA score compared to the wild-type peptides, suggesting a stabilization of the active conformation. The variants also exhibited a characteristic "funnel" shape in the ROSETTA score plots, where a decrease in energy is correlated with low RMSD models, suggesting that they primarily fold into the active conformations and not into alternative, competing conformations. In addition, we modeled these peptide variants in the context of the antibody-antigen interaction, eliminated any mutations that negatively affected the antigen binding interface in the simulations, and refolded all variants to ensure they retained desirable folding energies and landscapes (Supplementary Figure VI.2). This analysis prioritized eight redesigned peptides for further characterization.

Molecular dynamics simulations

To further assess the expected stability of these peptides in solution, we used molecular dynamics (MD) simulation as a complementary modeling approach. We modeled the behavior of two wild-type peptides and eight redesigned peptides in explicit solvent over a time scale of 50 ns and measured the fluctuation in Cα RMSD (Supplementary Figure VI.3). The MD simulations of full-length peptides differed from the ROSETTA folding data, suggesting that the wild-type peptide was relatively stable and that peptides d1 and d7 would not adopt the active conformation. In contrast, the MD simulations of truncated peptides matched well with the ROSETTA simulations. The wild-type truncated peptide appeared to adopt an alternate conformation after roughly 40 ns. Several of the redesigned truncated peptides (d2, d4, d5) remained stable in the active conformation after 50 ns, and others (d8, d10) adopted an alternate conformation similar to the wild-type peptide.

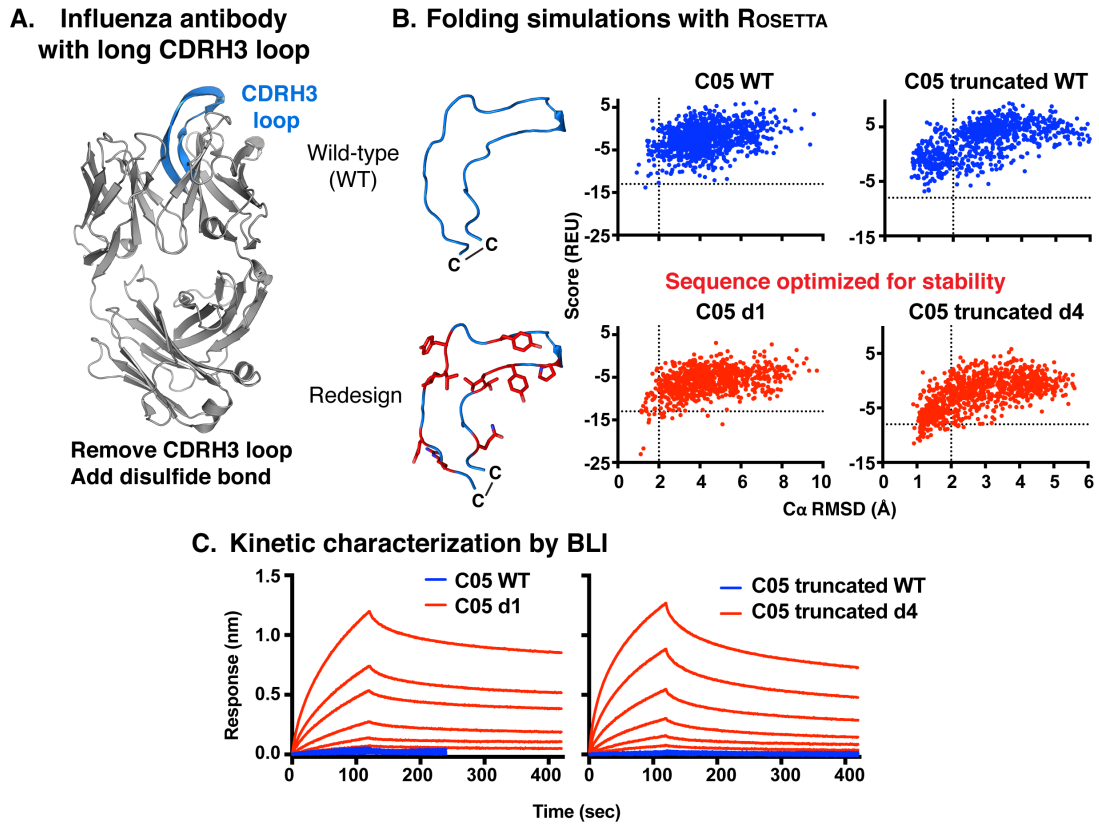


Figure VI.1. Experimental workflow of designing CDRH3-derived cyclic peptides. A. Influenza antibody C05 was chosen due to its long CDRH3 loop involved in antigen recognition. The CDRH3 loop was removed from the antibody and cyclized with a disulfide bond in silico. B. Folding simulations of the full-length and truncated C05 cyclic peptide were performed using Rosetta loop modeling. The wild-type (WT) CDRH3 sequence (blue) was redesigned to stabilize the peptide and improve the energy landscape (red). Favorable energy landscapes have native-like models (low Cα RMSD) with low Rosetta score (lower left quadrant). Score is expressed in Rosetta energy units (REU). C. WT (blue) and redesigned (red) cyclic peptides were synthesized and characterized for binding kinetics to recombinant influenza HA using biolayer interferometry (BLI).

Experimental characterization of redesigned peptides

Based on the *in silico* modeling of these peptides by two complementary approaches, we synthesized and characterized eight redesigned and two wild-type peptides, four of the full-length and six of the truncated form, cyclized by a disulfide linkage between the N- and C-termini. We then tested binding of these peptides to recombinant influenza HA by biolayer interferometry

(BLI). We found that while binding of the wild-type peptides to HA was not detected, two redesigned peptides (one full-length and one truncated) displayed high-affinity binding (Figure VI.2). These peptides bound to HA proteins of both the H1 and H3 antigenic subtype with an affinity better than 100 nM. To verify that this binding activity was not an artifact of the BLI system, we repeated the binding assay using ELISA on streptavidin-coated plates (Supplementary Figure VI.4). We observed a clear binding signal for C05 d1 on this system, albeit at lower apparent affinity. C05 truncated d4 showed low but observable binding in ELISA, however the EC_{50} could not be calculated due to lack of saturation. We attribute this finding to the fact that the truncated peptide lacks the torso of the CDRH3 loop, reducing the total linker distance between the biotin tag and functional portion of the peptide. Therefore, we conclude that peptides C05 d1 and C05 truncated d4 bind with high affinity to influenza HA.

To verify specificity of binding to the receptor-binding site on influenza HA, we performed competition binding on a biosensor using BLI. We first immobilized peptides to the biosensor, then bound recombinant HA followed by either a receptor-binding site antibody (mAb C05) or stem binding antibody (mAb CR6261, Figure VI.2D and H). We observed that, while CR6261 bound the HA tethered to peptide, binding of C05 was not detected, indicating that peptides bind specifically to the receptor-binding site as predicted.

To directly compare the affinity of the redesigned peptides to the affinity of C05 IgG, we measured binding to a monomeric head domain of HA from the H1 influenza virus strain A/Solomon Islands/03/2006. We observed an avidity effect from using a trimeric HA that depended on the density of the immobilized molecule (either peptide or IgG), suggesting monomer binding is the most unbiased method to directly compare affinities. The peptides bound monomeric HA head domain with an affinity comparable to that of C05 IgG (Supplementary Figure VI.5).

C05 IgG bound with an affinity of 88 nM, while the peptide affinities were 124 nM or 506 nM for full-length or truncated peptides, respectively. Notably the C05 IgG bound with high on and off rates, consistent with previous work on monovalent binding of C05 (Ekiert et al., 2012), whereas the peptides bound with slower on and off rates.

We repeated binding assays with a linear peptide of the same sequence to test our hypothesis that cyclization increases affinity by stabilization of proper conformation. We reduced the disulfide bond in the peptides before coupling to the biosensor and repeated binding to HA. Although we could still observe low levels of binding with the linear peptides, the affinity was reduced significantly (Supplementary Figure VI.6), suggesting that cyclization does improve peptide activity.

To investigate the biological activity of these peptides, we performed hemagglutination inhibition (HAI) assays with influenza H1 virus. Neither of the peptides showed HAI activity when tested in concentrations up to 50 μM (data not shown). In an effort to improve the avidity of the peptides we preloaded peptides onto streptavidin tetramers and repeated HAI, which also did not show activity when tested in concentrations up to 2.5 μM (data not shown).

Peptides have increased breadth compared to IgG

C05 IgG is known to have unusual breadth for an antibody to the influenza receptor-binding site, binding to both group 1 and 2 HA molecules (Ekiert et al., 2012). After confirming the binding activity of two redesigned peptides, we tested binding to a diverse panel of influenza HA, including strains known to bind or not bind C05 IgG. Remarkably, we found that the redesigned peptides not only maintained the breadth of C05 IgG, but they exhibited increased breadth (Table VI.1). The peptides recognized new strains within the H1 subtype, increasing breadth to A/Puerto Rico/8/1934 for peptides d1 and d4 and A/California/04/2009 for peptide d4.

The binding to A/Puerto Rico/8/1934 is consistent with previous work showing that a computationally redesigned C05 mutant antibody gains binding to this strain (manuscript submitted). In addition, HAs from two new influenza virus A subtypes, H4 and H7, were bound by peptides but not by IgG. As a control we performed binding to an irrelevant antigen from HIV (Supplementary Figure VI.7). We did observe a low level of binding to the irrelevant antigen that was above background signal, indicating that there is a weak nonspecific component to the peptide interaction. However, the signal from specific binding to HA was clearly distinct from the nonspecific signal (Supplementary Figure VI.7). Therefore, we only considered binding to variant HAs if the signal was at least twice as strong as that to the irrelevant antigens.

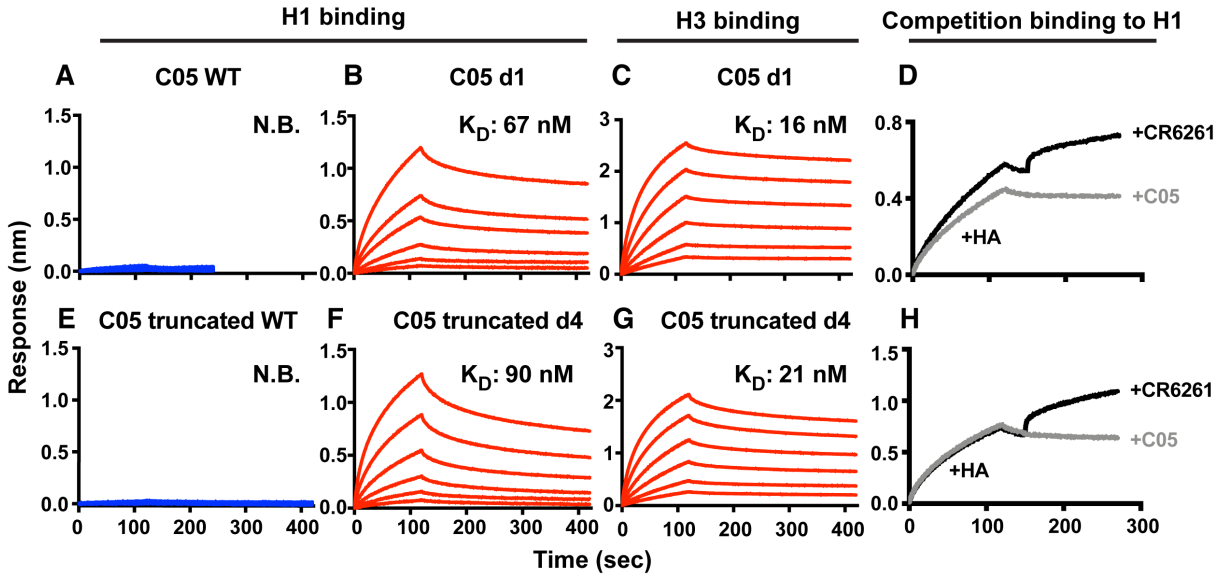


Figure VI.2. Redesigned cyclic peptides bind with high affinity to group 1 and 2 HAs. Redesigned cyclic peptides C05 d1 (upper panels) and C05 truncated d4 (lower panels) bind to H1 (B, F) or H3 (C, G) HAs with high (<100 nM) affinity. The wild-type CDRH3 sequence does not bind to H1 HA in either full-length (A) or truncated (E) formats. To identify the peptide epitope, peptides were loaded onto a biosensor, which then was treated with recombinant HA from H1 A/Solomon Islands/03/2006 virus followed by a receptor-binding site (C05) or stem (CR6261) antibody (D, H).

Table VI.1. C05-based cyclic peptides have increased breadth of recognition of diverse influenza HA molecules compared to the parental IgG molecule.

Group	Subtype	Strain	C05 d1	C05 d4	C05 IgG
1	H1N1	A/Solomon Islands/03/2006	+++	+++	++++
		A/Solomon Islands/03/2006 head domain	++	++	+++
		A/Brevig Mission/1/1918	-	-	-
		A/Tottori/YK012/2011	-	-	-
		A/mallard/Alberta/35/1976	-	-	++
		A/Puerto Rico/8/1934	+++	+++	-
		A/Texas/36/1991	-	-	-
		A/New Caledonia/20/1999	+++	+++	++
		A/California/04/2009	-	++++	-
	H2N2	A/Japan/305/1957	+++	++	++++
		A/Singapore/1/1957	+++	+++	++++
	H5N1	A/Vietnam/1203/2005	-	-	-
		A/Indonesia/5/2005	-	-	-
	H9N2	A/turkey/Wisconsin/1/1966	++++	+++	++
	H16N3	A/black-headed gull/Sweden/4/1999	-	-	-
2	H3N2	A/Hong Kong/1/68	+++	+++	++++
		A/Brisbane/10/2007	+++	+++	++++
		A/Perth/16/2009	+++	-	++++
		A/Panama/2007/1999	-	-	++++
		A/Bangkok/1/1979	-	-	-
	H4N6	A/duck/Czechoslovakia/1956	+++	+++	-
	H7N9	A/Shanghai/02/2013	+++	+++	-
		A/Netherlands/219/2003	-	-	-
	H15N8	A/shearwater/Western Australia/2576/1979	-	-	-

Legend

++++	<10 nM
+++	10-100 nM
++	100-1,000 nM
-	Specific binding not detected

No change in breadth compared to IgG
Gain of breadth compared to IgG
Loss of breadth compared to IgG

Structural analysis of peptides

Peptides that showed binding activity, d1 and d4, were highly mutated compared to the wild-type CDRH3 sequence (Figure VI.3A), with 11/28 or 10/18 amino acids mutated, respectively. Interestingly, the designed mutations in these peptides converged on the same amino acid at three positions, converging on hydrophobic residues (Figure VI.3A). A structural analysis of the peptide models suggests that the ROSETTA-designed mutations function by creating hydrophobic patches on the peptide to induce proper folding. Mutations in d1 are predicted to create a patch between Y9, P11, and Y16, and another patch involving L8, Y19, V20, and I21, which both cross opposing strands of the loop and encourage proper loop closure (Figure VI.3B). In addition, mutation Q3 is predicted to create a hydrogen bond with the neighboring main chain, and mutations to E24 and R26 are predicted to create an electrostatic interaction in the torso region of the loop (Figure VI.3B). In peptide d4, two hydrophobic patches are predicted to form, consisting of residues Y5 and L7, and residues L4, Y12, L15, P16, and L17 (Figure VI.3C). Both of these peptides are predicted to fold into conformations similar to that of the C05 CDRH3 loop, with Cα RMSDs of 1.6 and 1.8 Å for d1 and d4, respectively.

Evasion of HA hypervariable elements by peptides

To compare the predicted binding poses of peptides d1 and d4 to that of C05 IgG, we used ROSETTADOCK to model the bound conformation of the peptides in the receptor-binding site of the HA from nine different subtypes (Supplementary Figure VI.8). The peptides docked to H3 A/Hong Kong/1/1968 are predicted to adopt very similar conformations to C05 IgG (Figure VI.4A). Peptides d1 and d4 have a Cα RMSD of 1.7 and 2.4 Å, respectively, when aligning the HA component and calculating RMSD of the peptide to the C05 CDRH3 loop. These docked poses

therefore agree well with the experimental binding data and suggest that the redesigned peptides mimic C05 in their recognition of HA.

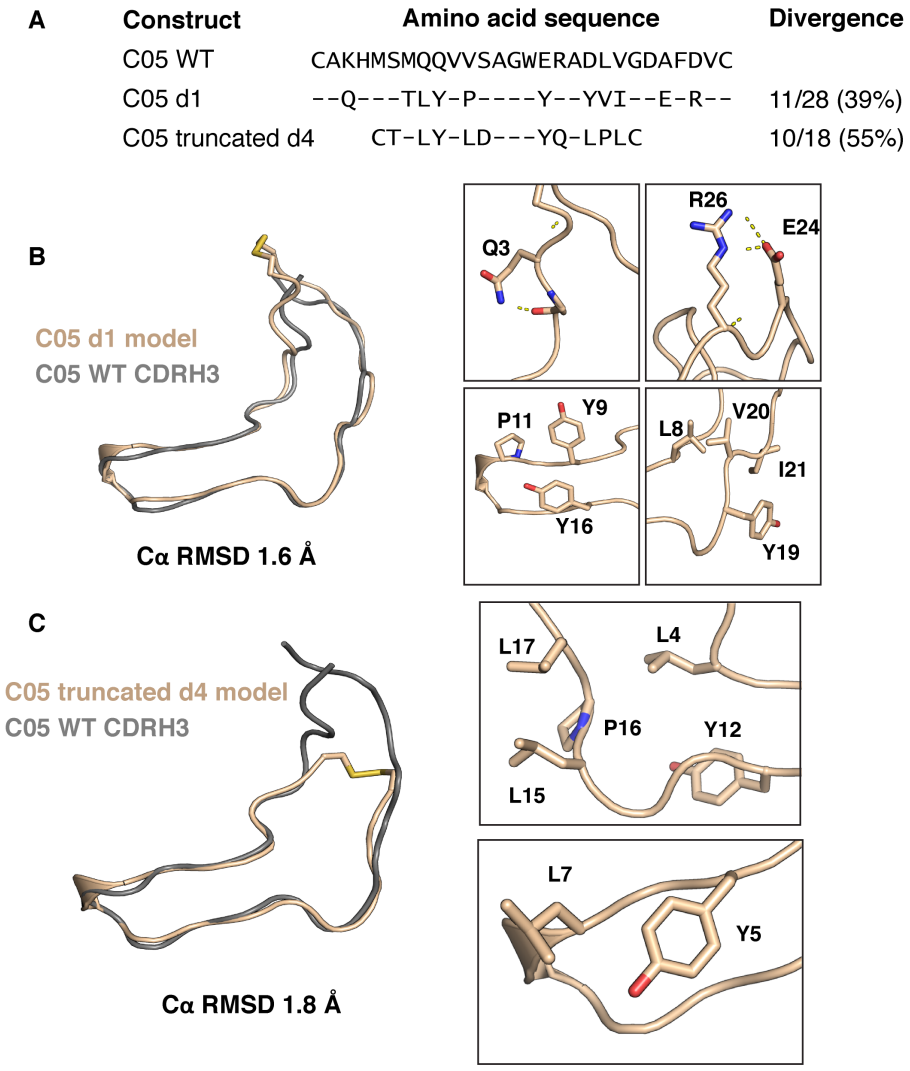


Figure VI.3. Structural analysis of redesigned cyclic peptides. A. Sequence alignment of redesigned peptides compared to the wild-type C05 CDRH3 sequence. Residues with the same identity as wild-type are shown as dashes. The total number of mutations in each peptide is also shown. B-C. Models of redesigned cyclic peptides (tan) are compared to the C05 CDRH3 structure (gray), and Cα RMSD over all residues in the peptide is shown below. The mutations introduced into the peptide sequence in their structural contexts are highlighted.

To explain the enhanced breadth of the peptides, we compared the conformation of hypervariable antigenic elements in strains H4 A/duck/Czechoslovakia/1956 and H7 A/Shanghai/02/2013 to the binding pose of C05. We identified two antigenic elements, the loop at position 150 of the HA (150-loop) and the helix at position 190 (190-helix), which are known to influence binding of receptor-binding site antibodies (Lee et al., 2014; Wu et al., 2017; 2018). H7 A/Shanghai/02/2013 has an insertion in the 150-loop compared to H3 A/Hong Kong/1/1968 that directly clashes with the CDRH1 of C05 (Supplementary Figure VI.9). In H4 A/duck/Czechoslovakia/1956, the 150-loop has amino acid substitutions that clash with the CDRH3 of C05, and the 190-helix has substitutions clashing with the CDRH1. By removing the CDRH1 and introducing mutations into the CDRH3 of the redesigned peptides, we reduced the binding footprint to avoid these antigenic elements.

Based on our hypothesis that the peptides achieve increased breadth by reducing the binding footprint on the HA surface, we compared the footprint of C05 IgG and peptides d1 and d4 docked to nine HAs of the H1, H2, H3, H4, and H7 subtypes (Table VI.2). The IgG tended to contact a larger surface area than the peptides as predicted, with average buried surface areas of 730, 627, or 618 $Å^2$ for IgG, d1, or d4, respectively. We then compared the buried surface area of peptides docked into the subtypes with increased binding breadth, H4 and H7 (Figure VI.4B, C). In the docked models the peptides have a greatly reduced footprint compared to IgG, primarily due to reduced contacts on the 150-loop and 190-helix. The binding footprint of these peptides in the docked conformations therefore represents the minimal binding epitope of HA that is conserved across H1, H2, H3, H4, and H7 subtypes.

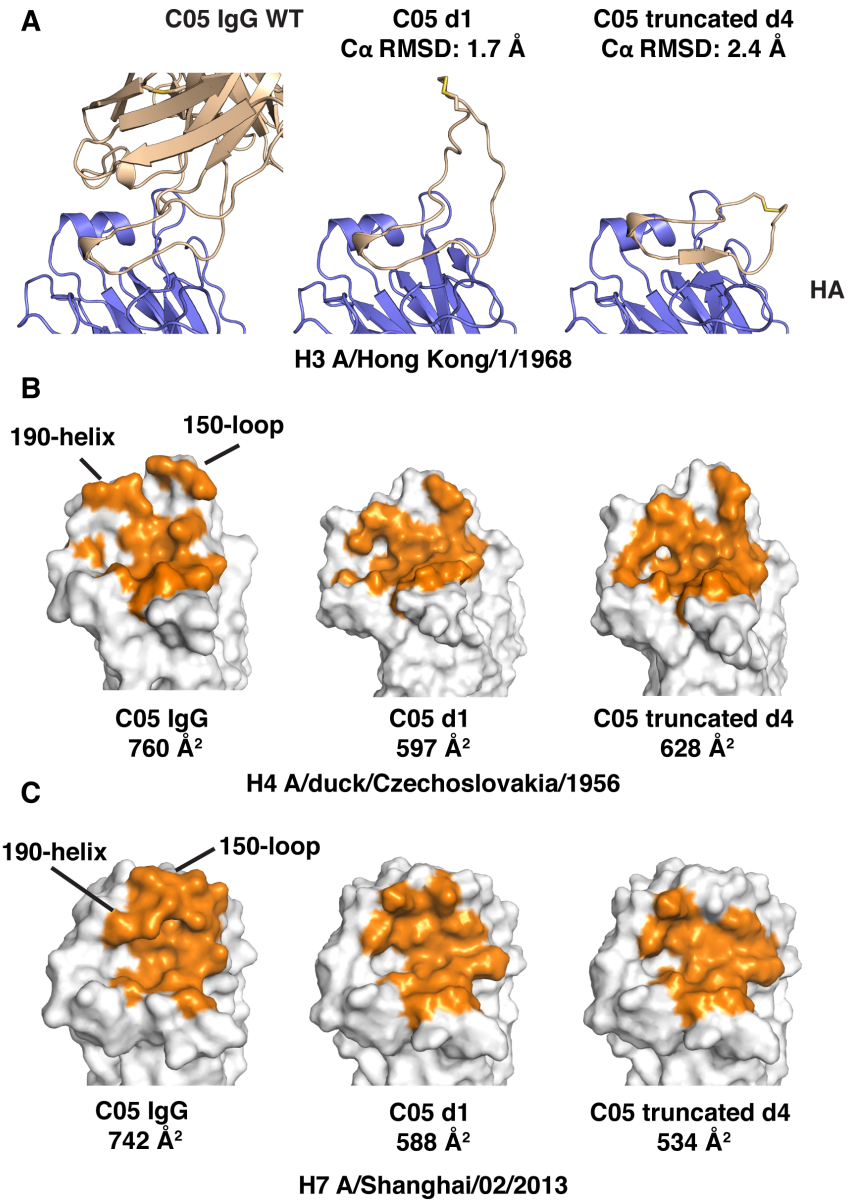


Figure VI.4. Cyclic peptides contact a minimal epitope on the surface of influenza HA. A. Models of peptides d1 and d4 (tan) were docked into the receptor-binding site of influenza HA (blue) from H3 A/Hong Kong/1/1968 (PDB ID 4fnk) and compared to the co-crystal structure of C05 IgG (PDB ID 4fp8). Cα RMSD was calculated by superimposing the HA and measuring RMSD over all residues on the peptide compared to the IgG co-crystal structure. B, C. The binding footprint (orange) of either C05 IgG or peptides d1 and d4 was calculated for two antigens for which the peptides have increased breadth, H7 A/Shanghai/02/2013 (B) and H4 A/duck/Czechoslovakia/1956 (C). Buried surface area on the HA surface was calculated and is shown below each structure.

Table VI.2. Buried surface area on the HA of various subtypes.

	Buried surface area on the HA (Å^2)		
Strain	**C05 IgG**	**C05 d1**	**C05 truncated d4**
H1 A/Solomon Islands/03/2006	639	638	586
H1 A/California/04/2009	777	580	569
H1 A/Puerto Rico/8/1934	734	720	593
H2 A/Japan/305/1957	753	596	800
H2 A/Singapore/1/1957	731	666	634
H3 A/Perth/16/2009	780	686	614
H3 A/Hong Kong/1/1968	651	568	609
H4 A/duck/Czechoslovakia/1956	760	597	628
H7 A/Shanghai/02/2013	742	588	534
Average	**730**	**627**	**618**

Mock co-complexes of C05 IgG with each subtype were created by aligning the HA structure to antigen in the co-crystal structure of C05 (PDB ID 4fp8). Peptides C05 d1 and C05 truncated d4 were docked into the receptor-binding site of each HA and the buried surface area of the lowest energy model was calculated.

Discussion

Summary of results

In this chapter, we show that computational design can be used to engineer antibody CDRH3-based peptides with high affinity and enhanced breadth of recognition for antigenic variants. We designed variants of the full-length and truncated CDRH3 loop of anti-influenza antibody C05 and identified two variants with potent activity. These peptides bound specifically to the influenza receptor-binding site and gained recognition for two HA subtypes, H4 and H7, not bound by the IgG molecule. Models of these two peptides suggest that they bind in a similar orientation to the C05 CDRH3 loop and achieve increased breadth of recognition by avoiding contact with the HA 150-loop and 190-helix hypervariable antigenic elements.

Although the peptides exhibited high-affinity binding on BLI they were not able to inhibit viral hemagglutination. We attribute this lack of activity to several factors. First, BLI may overestimate affinity compared to other kinetic platforms (Yang et al., 2016), therefore the peptides may have lower affinity than reported here. Second, antibodies targeting the influenza receptor-binding site are well known to rely on avidity for their neutralization activity, as many potently neutralizing antibodies show weak binding as monovalent antibody fragments (Fab) (Ekiert et al., 2012; Lee et al., 2014; Schmidt et al., 2013; Whittle et al., 2011). The peptides in this study were monomeric and therefore suffer from the same lack of avidity as a Fab. We repeated HAI assays with peptide loaded onto streptavidin tetramers in an attempt to increase avidity, which also failed to show activity, presumably due to incorrect geometry compared to the HA trimer. In future work, we plan to optimize the trimeric geometry of our peptides using scaffold proteins, which has been shown to impart neutralizing activity to computationally designed proteins (Strauch et al., 2017).

Cyclic peptide implications

Cyclic peptides have long been pursued as inhibitors of protein-protein interactions (Crook et al., 2017; Owens et al., 2017), modulators of antibody activity (van Rosmalen et al., 2017), mimics of viral antigenic loops (Bird et al., 2014), and antibody mimics (Casset et al., 2003; Kadam et al., 2017; Levi et al., 1993). However, cyclic peptide design has been met with many challenges. Most protein loops do not readily assume their active conformations as peptides, severely limiting the scope of which antibodies can be mimicked by peptides. The work in this study surpasses this limitation using structure-based computational design, by showing that the C05 CDRH3 peptide has no activity when using the wild-type sequence and only functions after introducing designed variations. Instead of limiting the mimicry of antibodies by use of the

naturally occurring CDRH3 peptide, we introduce a more systematic approach to create stable peptides from loops based on structure-based design principles. This technology has the potential to be applied to a wide variety of projects involving antibody therapeutics.

Minimal epitope for influenza HA receptor-binding site recognition

We identified a minimal epitope that is conserved across many influenza A subtypes, including H1, H2, H3, H4, H7, and H9. Influenza antibodies, especially those that target the highly conserved receptor-binding site, typically have restricted breadth to a single subtype (Lee et al., 2014; Schmidt et al., 2015; Whittle et al., 2011; Winarski et al., 2015) or to a subset of strains within a subtype (Ekiert et al., 2012; Krause et al., 2012; Lee et al., 2012). This restriction of breadth is due to contacts outside of the receptor-binding site that clash with hypervariable antigenic elements. The peptides designed in this study are of significant interest since they avoid contact with these variable antigenic elements and target a minimal epitope capable of achieving high-affinity binding against a variety of diverse HA antigens. Identification of this highly conserved epitope can be applied to the design of highly potent small molecule and protein inhibitors (Kadam and Wilson, 2018). The computational methods used to engineer these peptides also can be applied to other systems to identify minimal epitopes required for broad and potent activity.

Methods

Structure preparation

To generate templates for peptide modeling, we first extracted the CDRH3 loop from the crystal structure of antibody C05 from the Protein Data Bank (PDB, ID 4fnl) using PyMol (Schrodinger, LLC, 2015). The loop was renumbered starting from residue 1. For the truncated

form, the CDRH3 was truncated based on visual inspection. Cysteines were added to the N and C termini using the PeptideStubMover functionality in ROSETTA (see Appendix E for details).

Cyclic peptide closure and redesign

Once peptides were processed, they were subjected to loop closure simulations using the ROSETTA Generalized Kinematic Closure (GeneralizedKIC) protocol (Bhardwaj et al., 2016; Stein and Kortemme, 2013). Peptides were closed using a residue at the tip of the CDRH3 as the anchor point and perturbed using random perturbations of φ and ψ angles along the loop. Full details of the loop modeling protocol can be found in the Protocol Capture in Appendix E. 1,000 decoys were generated for each peptide, and the score and Cα RMSD compared to the wild-type loop were calculated.

After modeling wild-type peptides, we redesigned the sequence to increase convergence on the active conformation. We identified all folded loops from the previous simulation that were within 2 Å of the native loop and subjected the top ten by score to ROSETTA fixed backbone design (Leaver-Fay et al., 2013). We then simulated folding of the redesigned peptides and calculated score and Cα RMSD compared to the wild-type loop. To identify sequences that converged more readily on the active conformation, we used the funnel discrimination metric described in Conway *et al.* (Conway et al., 2014).

Binding energy calculations

To ensure the redesign of peptides did not introduce residues that would clash with the antigen, we modeled the designed sequence in the context of the antibody-antigen interface. We modeled the isolated CDRH3 loop (PDB ID 4fnl) in complex with H1 A/Solomon Islands/03/2006 (PDB ID 4hkx). We created the mock complex of these proteins by aligning to the co-crystal structure of C05 in complex with an H3 antigen (PDB ID 4fp8). We then threaded the mutated

residues and refined the structure using ROSETTA relax with a 1.0 Å backbone constraint to the starting coordinates. Binding energy (ΔΔG) was calculated as below:

$$\Delta\Delta G = E_{complex} - (E_{Ab} + E_{Ag})$$

where E_{Ab} and E_{Ag} are the energies of the antibody and antigen alone, respectively. When mutated residues affected the binding energy of the complex, we reverted these mutants individually and determined which contributed most to the increase in binding energy. We then modeled revertants with the same loop modeling protocol, and mutants that were favorable in both convergence on the active conformation and binding energy were selected for experimental characterization.

Molecular dynamics simulations

To test whether the cyclic peptides would remain stable in solution, we subjected them to molecular dynamics (MD) simulation. Input files and structure were prepared using the VMD software (Humphrey et al., 1996) with the QwikMD plugin. As input structures, we used the lowest energy models from the ROSETTA loop closure simulations. The system was solvated a using cubic water box with a 7.5 Å buffer with a salt concentration of 0.15 mol/L. MD simulations were performed using the NAMD package (Phillips et al., 2005) with the CHARMM36 force field (Feller et al., 1995; Jorgensen et al., 1983). The MD simulation without constraints was performed with explicit solvent using the TIP3 water model (Jorgensen et al., 1983) in the NpT ensemble. The temperature was maintained at 300.00 K using Langevin dynamics. The pressure was maintained at 1 atm using Nosé-Hoover Langevin piston (Feller et al., 1995; Martyna et al., 1994). A distance cut-off of 12.0 Å was applied to short-range, non-bonded interactions, and 10.0 Å for the smothering functions. Long-range electrostatic interactions were treated using the particle-

mesh Ewald (PME) (Darden et al., 1993) method. The equations of motion were integrated using the r-RESPA multiple time step scheme (Phillips et al., 2005) to update the short-range interactions every 1 steps and long-range electrostatics interactions every 2 steps. The time step of integration was chosen to be 2 fs for all simulations. In this step consisting of 10.0 ns of simulation, no atoms were constrained. All peptides were simulated for a total of 50 ns.

Docking

To generate models of the bound peptides the lowest energy peptide models were aligned to the CDRH3 loop position of C05 in PDB ID 4fp8 and saved in complex with the HA from nine different subtypes (PDB IDs 4hkx, 3ubq, 1rvx, 3ku3, 2wr7, 4fnk, 4kvn, 5xl3, and 4ln3). 1,000 decoys were generated of the two peptides docked into the five antigens using ROSETTADOCK (Gray et al., 2003). The protein-peptide interface then was refined using the ROSETTA relax protocol (Combs et al., 2013). The ROSETTA score and Cα RMSD was calculated for each decoy compared to the wild-type C05 CDRH3 loop. Buried surface area calculations were performed using PyMol (Schrodinger, LLC, 2015).

Experimental characterization

Eight redesigned peptide candidates were selected for characterization based on the previously described criteria, along with two control peptides with the wild-type C05 sequence. Peptides were synthesized by Genscript with a disulfide linkage between residues at the N and C termini, and a C-terminal polyethylene glycol (PEG) 6 linker connected to a lysine-linked biotin group.

Recombinant HA expression

Sequences encoding the HA genes of interest were optimized for expression in human cells and synthesized (Genscript). Genes were constructed as soluble trimer constructs by replacing the

transmembrane and cytoplasmic domain sequences with a GCN4 trimerization domain and a 6x-His tag at the C-terminus. Synthesized genes were cloned into the pcDNA3.1(+) mammalian expression vector (Invitrogen). HA protein was expressed by transient transfection of Expi293F cells (ThermoFisher Scientific). Supernatants were harvested after 7 days, filter-sterilized with a 0.2-μm filter, and purified using affinity chromatography with a 5 mL HisTrap excel column (GE Healthcare). HA head domain was synthesized as a maltose-binding protein (MBP) fusion in pMAL-c5x vector (New England BioLabs). Head domain was expressed in SHuffle T7 Express competent *E. coli* (New England BioLabs) to enable disulfide formation in the cytoplasm, induced by the addition of 1 mM IPTG overnight at 18 °C, and purified using amylose resin (New England BioLabs).

Biolayer interferometry assay

Binding kinetics were determined using biolayer interferometry (BLI) with an Octet Red instrument (FortéBio, Menlo Park, CA). Peptides were loaded onto streptavidin biosensors at 5 μM in kinetics buffer (PBS + 1% BSA, 0.05% Tween 20). The binding experiments were performed with the following steps: 1) baseline in kinetics buffer for 60 s, 2) loading of peptide for 30 s, 3) baseline for 60 s, 4) association of HA for 120 s, and 5) dissociation of HA into kinetics buffer for 300 s. A reference well was run in all experiments, where peptide was loaded onto the biosensor, but antigen was not present, and was subtracted from all sample wells to correct for drift and buffer evaporation. Trimeric HAs were diluted two-fold starting from a concentration of 1.25 μM, and monomeric HA was diluted two-fold from a starting concentration of 20 μM. At least four dilutions of HA were used to fit kinetic curves. To eliminate nonspecific effects, the binding curves were compared to binding to an irrelevant antigen (HIV gp120), and binding was only considered significant if the signal was >2x as strong as the irrelevant signal. To test the effect

of cyclization on peptide affinity, the peptides were reduced using 2.5 mM TCEP before loading onto the biosensor. Curves were fit to a 1:1 or 2:1 binding model using the FortéBio software. Curve fits were accepted only if they fulfilled an R^2 of > 0.9. To perform competition binding, the peptide was loaded to streptavidin tips as previously described, and HA was bound at a concentration of 50 μg/mL followed by baseline and binding of either C05 or CR6261 IgG at a concentration of 50 μg/mL.

ELISA binding assay

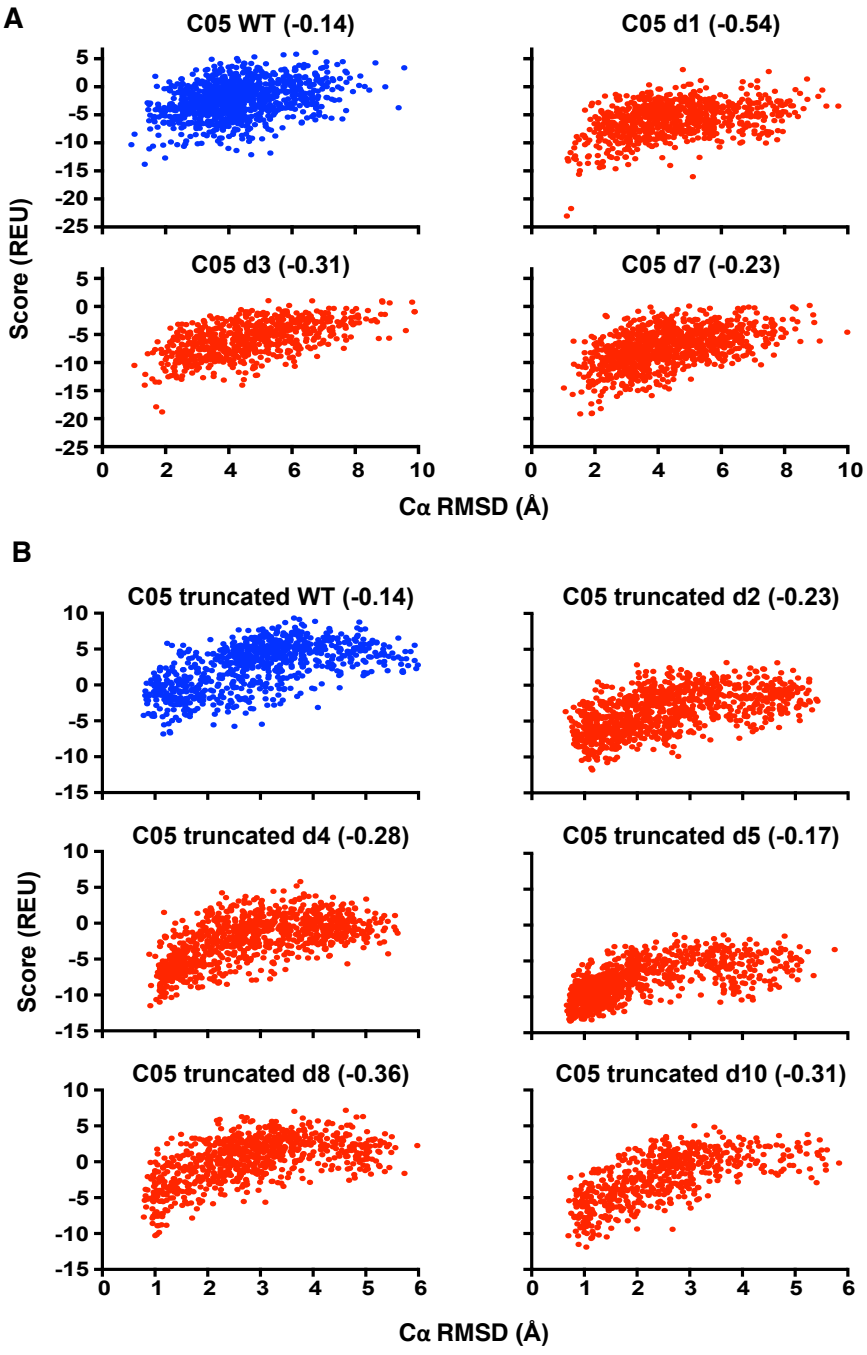
Biotin-labeled cyclic peptides were bound to a pre-coated streptavidin ELISA plate (Streptavidin Coated High Capacity Plates, ThermoFisher Scientific) at 1 μM and incubated for 1 hour at 37 °C. The plates then were blocked with 10% goat serum (Gibco) in PBS for 1 hour at 37 °C. HA monomeric head domain protein was diluted serially 2-fold in blocking buffer at a starting concentration of 14 μM. To detect binding, plates were incubated with a mouse anti-His mAb coupled to HRP (ThermoFisher Scientific). Binding was detected by addition of 100 μL of TMB substrate (ThermoFisher Scientific) and incubated for 5-10 min before quenching the reaction with 100 μL of 1 N HCl. Plates were read at 450 nm using a BioTek plate reader. After plate coating and primary and secondary antibody incubation, plates were washed 3x with wash buffer (PBS +0.05% Tween 20, Cell Signaling Technologies). EC_{50} values were calculated in GraphPad Prism using robust nonlinear regression. All ELISAs were performed in triplicate.

Viruses and hemagglutination assay

Influenza virus strain A/Solomon Island/3/2006 H1N1 strain was provided by Influenza Reagent Resource of US CDC. The working stocks used for hemagglutination inhibition assay (HAI) were made in MDCK cell culture. For HAI, 25 μL of four hemagglutination units of virus were incubated for 1 hour at room temperature with 25 μL two-fold serial dilutions of peptides

starting at 50 μM in PBS. The 50 μL of antibody-virus mixture was incubated for 45 minutes at 4 °C with 50 μL of turkey red blood cells (Rockland) diluted in PBS. The HAI titer was defined as the highest dilution of antibody that inhibited hemagglutination of red blood cells.

Supplemental Information



Supplementary Figure VI.1. Folding energy landscapes for C05-derived cyclic peptides. The wild-type (WT) C05 CDRH3 loop was folded using ROSETTA loop modeling and redesigned to improve stability and folding funnel, in both full-length (A) and truncated (B) formats. In parenthesis is the decoy discrimination score from Conway *et al.* 2014, with more negative values representing more desirable energy landscapes.

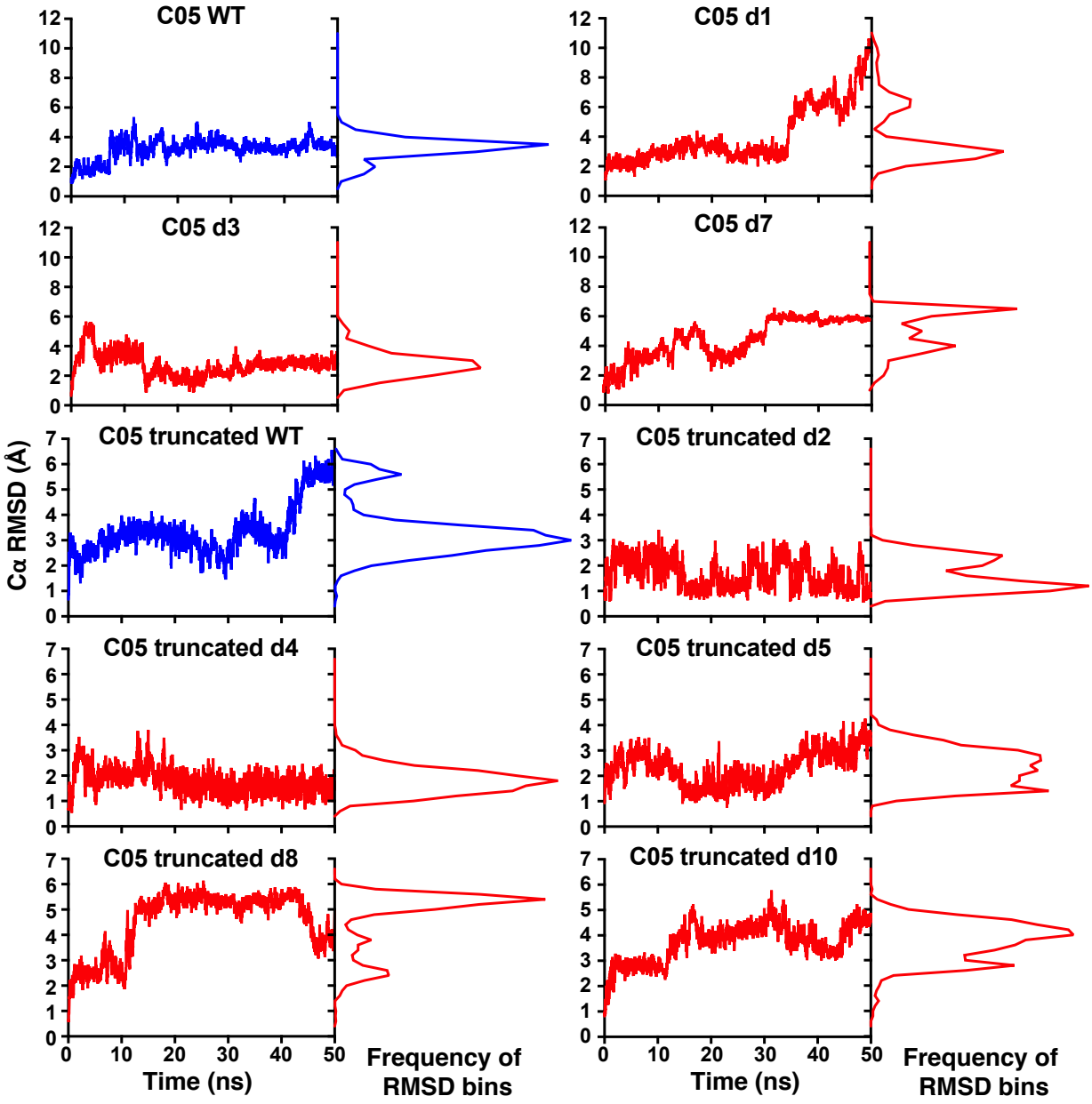
A

	Score	ΔΔG	Funnel statistic	Amino acid sequence
WT	-332.9	-10.2	-0.14	CAKHMSMQQVVSAGWERADLVGDAFDVC
d1	-312.6	5.0	-0.37	--Q-L-TLYIPVL--Y-PYVI--E-RK-
d1v1	-339.4	-11.1	-0.54	--Q---TLY-P----Y--YVI--E-R--
d1v2	-340.1	-11.4	-0.27	--Q---TLY-P----Y--YV---E-R--
d1v3	-339.8	-11.4	-0.22	--Q----LY-P----Y--YV---E-R--
d3	-307.4	5.9	-0.18	-ST-QDSLKDEDK-LETDLEK-GP--N-
d3v1	-333.7	-10.2	-0.11	-ST-QD-LK-ED---ET-LEK-GP--N-
d3v2	-332.8	-10.0	-0.31	-ST-QD--K-ED---ET-LEK-GP--N-
d3v3	-335.4	-10.8	-0.06	-ST-Q---K-ED---ET-LEK-GP--N-
d7	-316.2	5.1	-0.16	--V-EKT-H-PDE-NY-T-EE--Q-KR-
d7v1	-339.3	-10.6	-0.23	--V-E-T-H-PD---Y---EE--Q-KR-
d7v2	-337.8	-10.9	-0.08	--V-E---H-P----Y---EE--Q-KR-

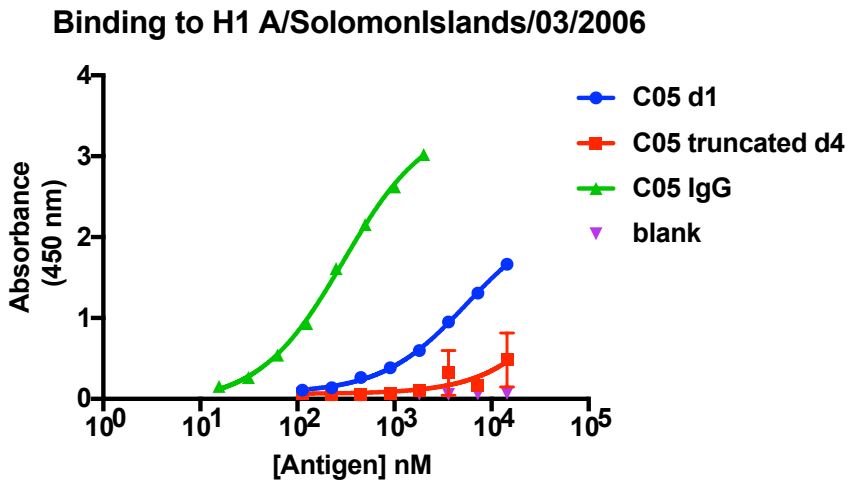
B

	Score	ΔΔG	Funnel statistic	Amino acid sequence
WT	-331.2	-9.9	-0.14	CSMQQVVSAGWERADLVC
d2	-309.6	7.9	-0.22	-KLTHIPNK---D-VP-
d2v1	-335.4	-10.1	-0.13	--LTH-PN------VP-
d2v2	-335.7	-10.1	-0.16	--LTH-P-------VP-
d2v3	-334.9	-10.1	-0.23	---TH-P--------P-
d2v4	-335.4	-10.1	-0.12	------P--------P-
d4	-319.9	2.1	-0.20	-TDLYRLDE--YQDLPL-
d4v1	-332.2	-9.5	-0.27	-TDLY-LD---YQ-LPL-
d4v2	-332.8	-9.9	-0.28	-T-LY-LD---YQ-LPL-
d4v3	-333.8	-10.3	-0.21	-TDLY-LD---Y--LPL-
d5	-311.3	3.5	-0.23	-GLTSIPDK--YSQTKP-
d5v1	-337.7	-11.9	-0.19	-GLTS-PD---YS-TKP-
d5v2	-337.0	-11.7	-0.17	-G-TS-PD---YS-TKP-
d8	-314.3	4.8	-0.55	-TK--DP-E--D-E-KR-
d8v1	-329.9	-10.1	-0.36	-T----P----D---KR-
d10	-313.1	6.3	-0.31	-TDKKDP-Q--S-EKDE-
d10v1	-335.0	-8.8	-0.31	-TDKK-P----S--KDE-
d10v2	-336.0	-9.2	-0.31	-T-KK-P----S--KDE-
d10v3	-336.3	-10.9	-0.24	-TDKK-P-------KDE-

Supplementary Figure VI.2. Redesigned peptides modeled in the context of the antibody-antigen complex. Either full-length (A) or truncated (B) peptides were modeled in the context of the antibody-antigen complex to retain antigen binding. Mutations were reverted back to wild-type (WT) sequentially (v1, v2, etc.) and the total complex score, binding energy (ΔΔG) and funnel discrimination statistic were calculated. Funnel discrimination score is calculated as in Conway *et al.* 2014, with more negative values representing more desirable energy landscapes. Highlighted are the sequences with the best folding profile and binding energy that were selected for experimental characterization.

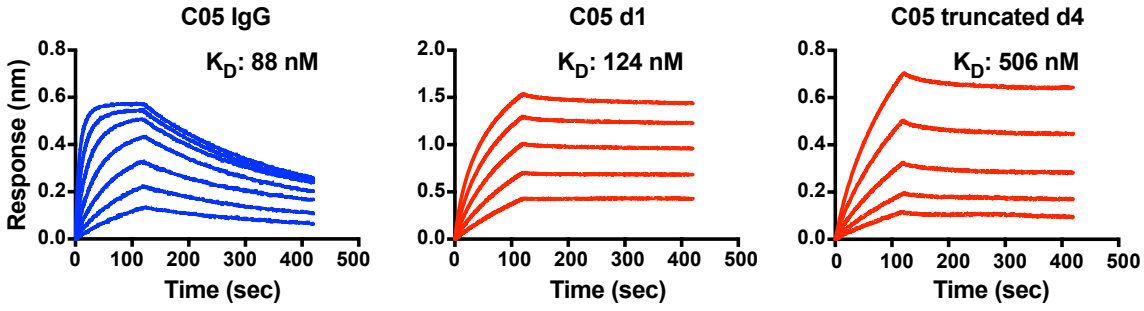


Supplementary Figure VI.3. Molecular dynamics simulations of cyclic peptides. Two wild-type (WT; blue) and eight redesigned peptides (red) were simulated in explicit solvent for 50 ns. Plotted is the fluctuation in Cα RMSD for each peptide over the simulation. Histograms of RMSD over the course of the simulation are shown on the right of the plots for each peptide.

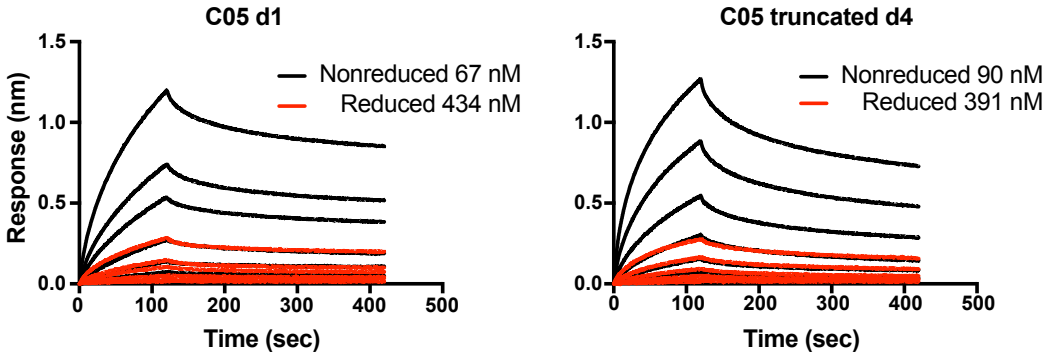


Construct	EC_{50} (nM)	95% CI
C05 d1	5,483	4,540 - 6,677
C05 truncated d4	Ambiguous	
C05 IgG	299	268-333

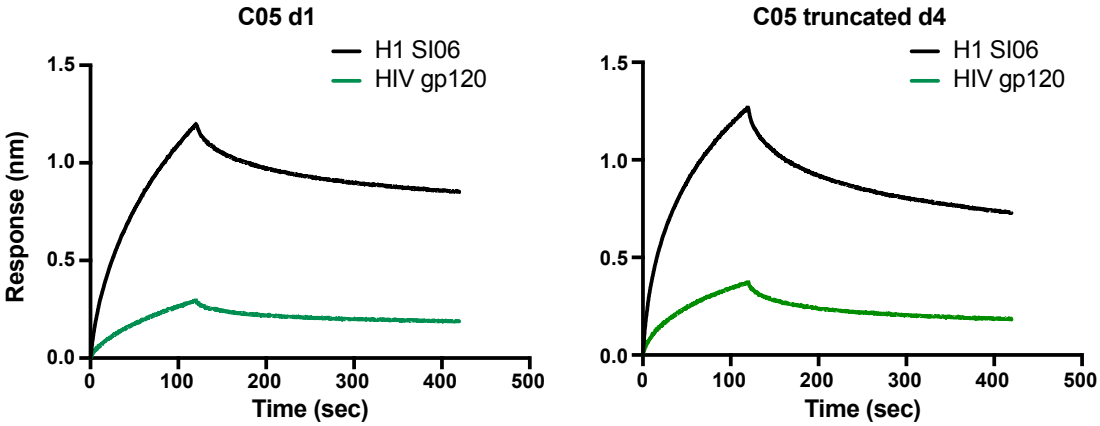
Supplementary Figure VI.4. Binding of peptides or IgG was repeated in ELISA format. Either peptides or biotinylated IgG were loaded onto a streptavidin-coated ELISA plate, which then bound to recombinant H1 HA. Binding was detected by an anti-His HRP-conjugated secondary. EC50 values and 95% confidence intervals are shown below.



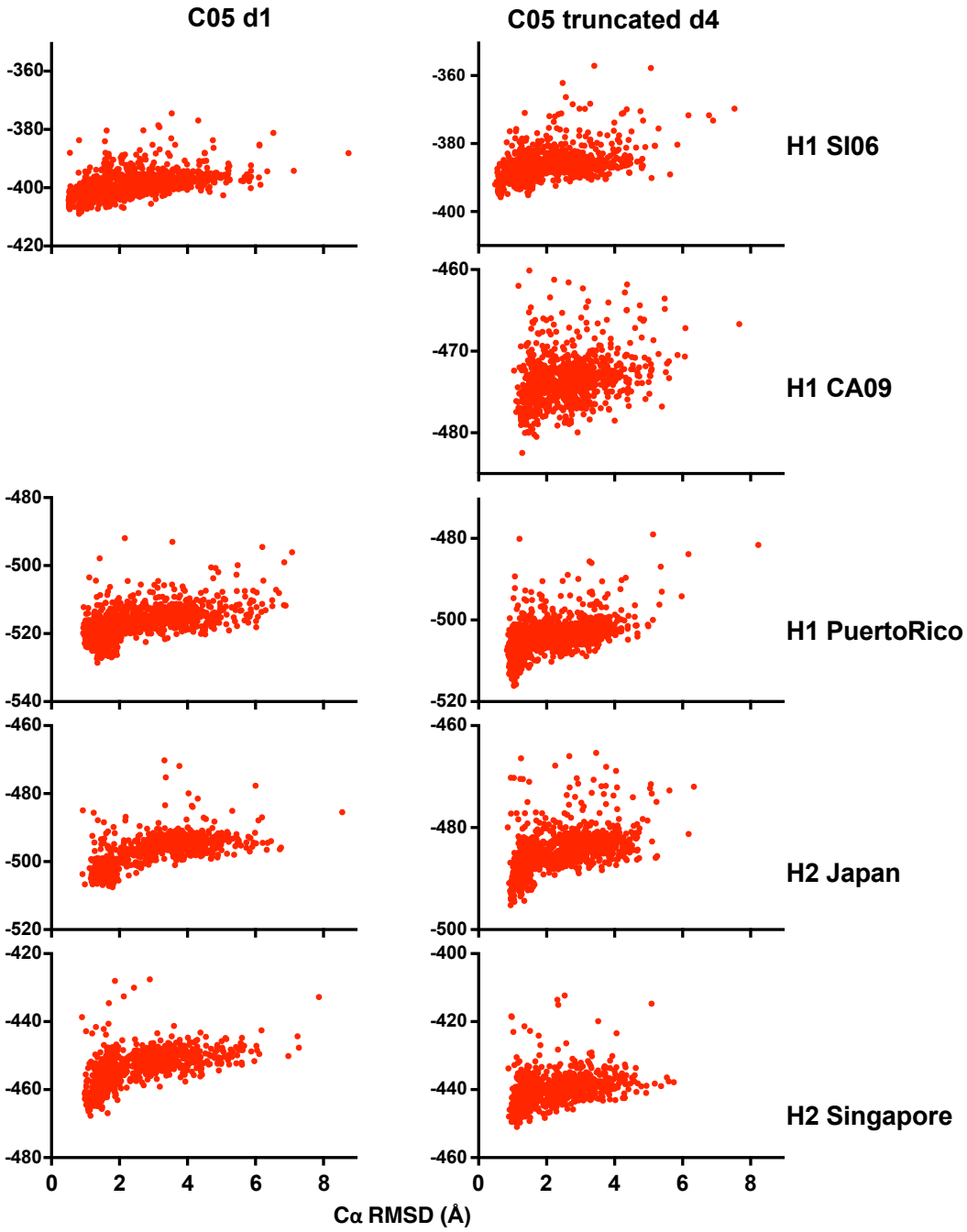
Supplementary Figure VI.5. Binding assays of C05 IgG and redesigned peptides to monomeric HA from H1 A/Solomon Islands/03/2006. Monomeric HA was used to eliminate avidity effects on the instrument and provide a direct comparison for affinity between peptides and IgG.

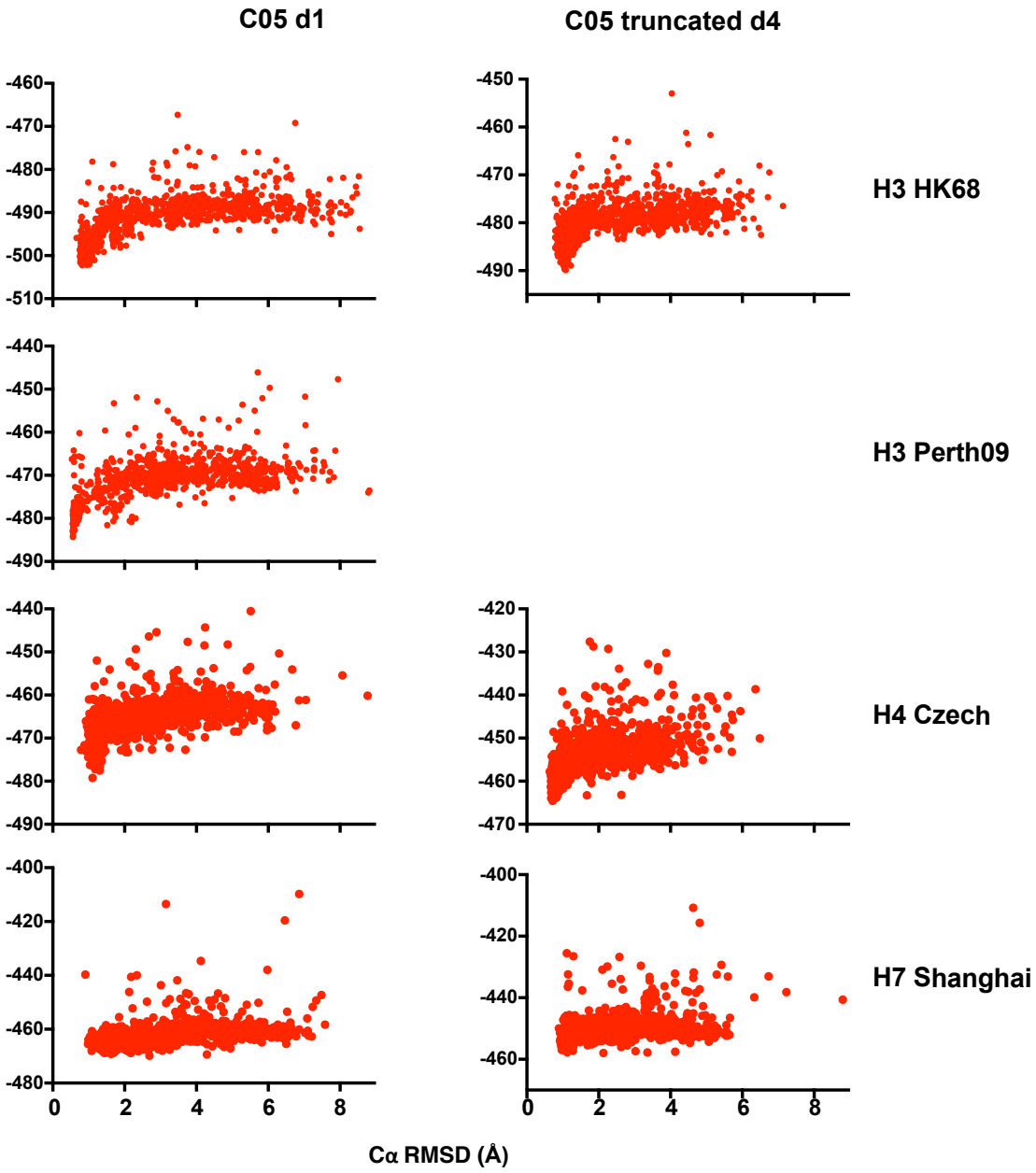


Supplementary Figure VI.6. Binding of redesigned peptides was repeated in the presence of a reducing agent to test affinity of linear peptides. Reduction was performed by addition of 2.5 mM TCEP. Nonreduced or reduced peptides are shown in black or red, respectively, with K_D values shown in the legend. Binding was performed with trimeric HA from H1 A/Solomon Islands/03/2006.

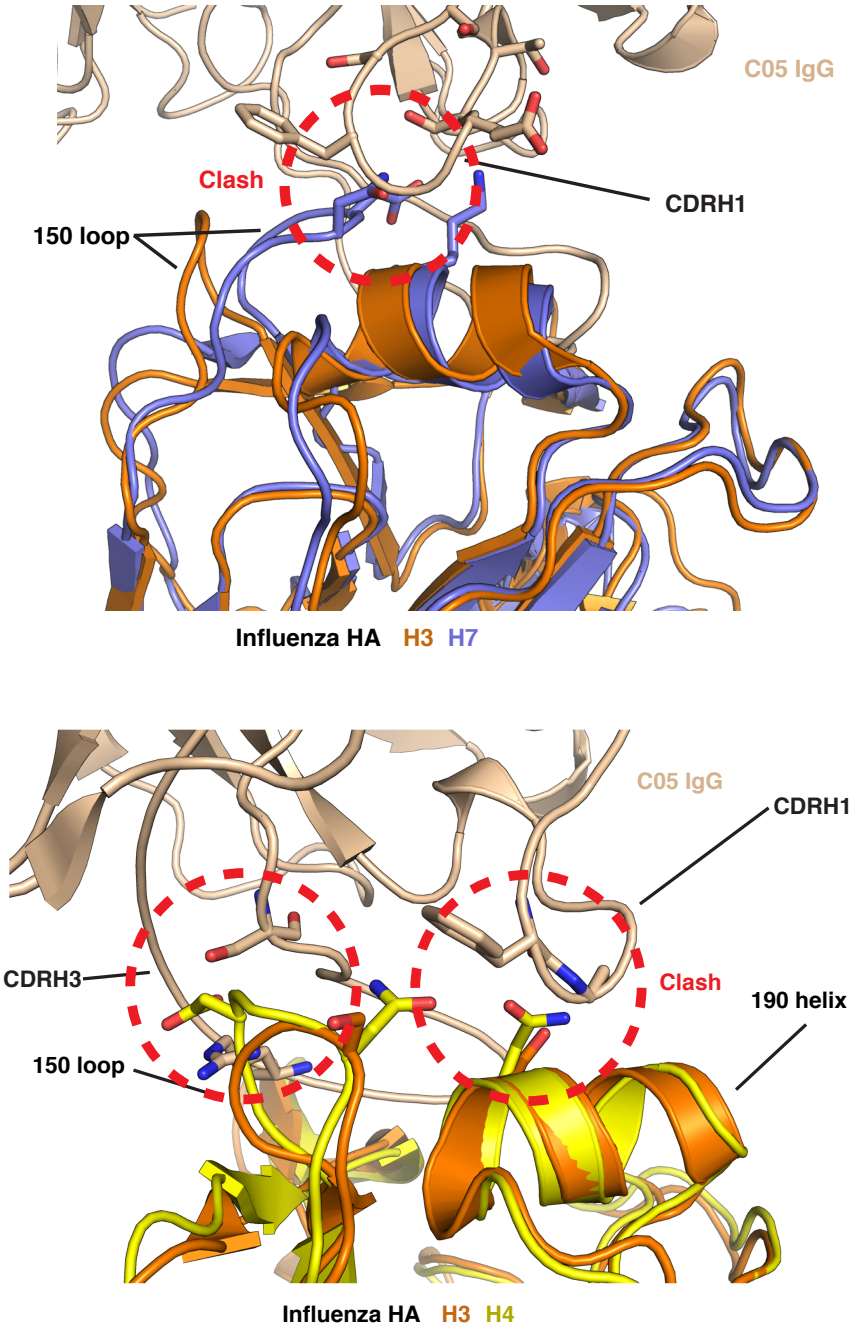


Supplementary Figure VI.7. Nonspecific binding was tested by binding to an irrelevant antigen. Binding of C05 designed peptides was tested with H1 A/Solomon Islands/03/2006 (H1 SI06) or HIV gp120, both at 1.25 μM. The level of binding to variant HAs was compared to gp120 binding to correct for the nonspecific component of binding.


C05 d1
C05 truncated d4
H1 SI06
H1 CA09
H1 PuertoRico
H2 Japan
H2 Singapore
Cα RMSD (Å)



Supplementary Figure VI.8. Docking funnels from peptide models docked into the receptor binding site of HA antigens from different subtypes. Docking funnels for C05 d1 (left column) and C05 truncated d4 (right column) are shown for docking to antigens in the panel. Only peptide-antigen pairs that bind are shown. Pictured: H1 A/Solomon Islands/03/2006 (PDB ID 4hkx), H1 A/California/04/2009 (3ubq), H1 A/Puerto Rico/8/1934 (1rvx), H2 A/Japan/305/1957 (3ku3), H2 A/Singapore/1/1957 (2wr7), H3 A/Hong Kong/1/68 (PDB ID 4fnk), H3 A/Perth/16/2009 (4kvn), H4 A/duck/Czechoslovakia/1956 (5xl3), and H7 A/Shanghai/02/2013 (4ln3). Cα RMSD compared to the CDRH3 from C05 is shown on the X axis, and ROSETTA score is shown on the Y axis in ROSETTA energy units (REU).



Supplementary Figure VI.9. Comparison of the structures of variant HAs. Top: Comparison of the structures of HA from H3 A/Hong Kong/1/1968 (orange) and H7 A/Shanghai/02/2013 (blue). The binding pose of C05 IgG is shown in tan, and clashes between C05 CDRH1 loop and H7 are highlighted in red. PDB IDs used for H3 and H7 structures are 4fp8 and 4ln3, respectively. Bottom: Comparison of the structures of HA from H3 A/Hong Kong/1/1968 (orange) and H4 A/duck/Czechoslovakia/1956 (yellow)**.** Binding pose of C05 IgG is shown in tan, and clashes between C05 and H4 are highlighted in red. Antigenic elements 150 loop and 190 helix on the HA are labeled. PDB IDs used for H3 and H4 structures are 4fp8 and 5xl3, respectively.

CHAPTER VII.

Conclusions and Future Directions

Summary of results

In this thesis I describe my work in engineering cross-reactivity into antibodies with a focus on computational techniques. Computational antibody design has long been possible, and in fact has achieved success in several cases in the past (Clark et al., 2006; Lippow et al., 2007; Marvin and Lowman, 2003; Midelfort et al., 2004; Willis et al., 2015). However, there was a great need for a method to design antibodies to not only achieve binding to a single target, but to incorporate multiple targets to impart cross-reactivity to the starting antibody, a paradigm known as multistate design. In Chapter II, I developed a computational method for multistate design within the ROSETTA software suite, called the RECON method, that reduced the amount of sampling needed to achieve a low energy multistate solution. Through two benchmark cases I show that the RECON method increased the computational efficiency of multistate design while also improving the biological relevance of designed sequences, compared to an existing multistate design method in ROSETTA. I then applied this method to a test case of improving breadth of anti-influenza antibodies to recognize a large panel of seasonal variants, as described in Chapter III. I showed that redesigned variants of anti-influenza antibody C05 exhibit increased breadth for an additional viral strain, as well as improved affinity for another strain, while maintaining high-affinity binding to several strains of the H1, H2, and H3 subtypes. Chapter IV describes a complementary approach, known as BROAD, for performing multistate design against large viral panels, by creating a large number of structural models and using machine learning and integer linear programming to

optimize breadth and affinity. I benchmark this approach by redesigning an anti-HIV antibody, VRC23, against a panel of viral variants of HIV. The BROAD method was able to search more exhaustively through sequence space and generate low energy solutions that were never encountered by structure-based multistate design using RECON. In Chapter V I address an immunological hypothesis based on antibody structural data, that anti-HIV antibodies may have precursors that are able to recognize influenza HA, and that HA may have been the original stimulating antigen for such anti-HIV antibodies. I performed next-generation sequencing to identify antibody sequences potentially capable of recognizing both influenza and HIV antigens and use a computational protocol known as a position-specific structural scoring matrix to identify many sequences with influenza-HIV cross-reactive potential. Lastly, in Chapter VI I describe a different way of achieving cross-reactivity with an antibody, by isolating the CDRH3 loop and redesigning the sequence for stability in the format of a cyclized peptide. Two antibody-based cyclic peptides showed increased binding breadth to novel influenza HA subtypes by reducing contact with hypervariable loops on the HA surface.

Energy functions in ROSETTA design

The work in this thesis highlighted several strengths and limitations of the energy functions used in ROSETTA protein modeling and design. Ultimately the predictive value of the energy function was suboptimal, as many antibody variants modeled *in silico* failed to exhibit their predicted behavior *in vitro*. Of the 27 variants of antibody C05 generated in Chapter II, only 2 showed a significant increase in affinity for any of the antigens tested, and 7 variants had an increase in either affinity or thermostability, a hit rate of 7-25% depending on the metric. Although this hit rate is low, it is in line with previous studies using ROSETTA and other software packages reporting a hit rate of ~ 10% (Adolf-Bryfogle et al., 2018; Baran et al., 2017; Chevalier et al., 2017;

Correia et al., 2010; Entzminger et al., 2017; Rocklin et al., 2017; Strauch et al., 2017). Therefore, it is clear that the ROSETTA energy function can be substantially improved. Although the work I described in Chapters II and IV improve the sampling algorithms during multistate design, they are ultimately limited by the accuracy of the energy function. I did observe that some energetic terms tended to have more predictive power than others. For example, successful mutations tended to improve van der Waals interactions and $\varphi - \psi$ angle favorability in loops. Many mutations were predicted to improve hydrogen bonding or electrostatic interactions at the antibody-antigen interface, and all but one of these mutants failed experimentally. Many in the ROSETTA community have reported similar results about the reliability of hydrogen bonding terms, which are more difficult to model than short-range van der Waals interactions (Alford et al., 2017; Rocklin et al., 2017; Stranges and Kuhlman, 2013). However, there have been improvements in these energetic terms, including work on improving hydrogen bonding terms (O'Meara et al., 2015), Boyken:2016ib, Maguire:2018gr} that have led to improvements. In addition, a longtime challenge in the ROSETTA score function has been accurate representation of water molecules at a protein-protein interface. Existing iterations of the ROSETTA score function rely on an implicit model of protein solvation rather than explicitly placing water molecules (Lazaridis and Karplus, 1999). There has recently been work to both improve the implicit solvation model (Bazzoli and Karanicolas, 2017) and add explicit water to simulations (Marze et al., 2016). Given that there have been substantial improvements in energetic terms since this work began, it would be worthwhile to predict the activity of the mutants reported in this thesis with the improved energetic terms to see if they improve accuracy.

High-throughput assays for experimental validation

In future work, the computational approaches described in this thesis could be greatly strengthened by incorporating high-throughput experimental validation. As previously discussed, the success rate of ROSETTA-designed mutations is relatively low, historically around 10%. Therefore, to improve the odds of obtaining a successful mutant it is necessary to express and test mutants at a larger scale. In this work, mutants were made by synthesizing and cloning individual genes in Chapter II and by solid-phase peptide synthesis in Chapter VI, and testing was done by ELISA or biolayer interferometry binding assays. This led to a low overall throughput of testing, with 27 mutants tested in Chapter II and 8 mutants in Chapter VI. Recently there have been advances in gene synthesis that can synthesize 10,000 individually specified genes (Kosuri and Church, 2014), allowing expression and testing of designed proteins on a much larger scale (Rocklin et al., 2017; Sun et al., 2016). These variants can then be cloned into a yeast or phage display vector and the library can be screened for improved thermostability or binding to a given target. In future multistate design experiments, it would be very useful to take advantage of high-throughput synthesis to create libraries of ROSETTA-designed mutations and screen against large panels of antigen variants. One of the principal benefits of multistate design over experimental screening is the ability to incorporate many more targets during *in silico* simulations. The ability to simulate large viral panels was one of the main motivations for the algorithms developed in Chapters II and IV. However, multistate design and experimental screening are complementary approaches that should be combined in a workflow to maximize success of designed proteins.

The inability to test antibody mutants in a high-throughput assay was one of the main limitations of the work done in Chapter V. Many of the antibody sequences obtained from human donors in Chapter V were predicted to be cross-reactive to HIV and influenza based on ROSETTA

modeling. However, I did not express these antibodies to verify their binding activity. This was based on several factors. The computational binding predictions were based on modeling the CDRH3 of sequenced antibodies chimerized onto the backbone of either anti-influenza mAb 641 I-9 or anti-HIV mAb Z13e1. To identify a single cross-reactive antibody, it would be necessary to model a given CDRH3 loop in the context of both mAb backbones and use RECON multistate design to optimize the compatibility with the two backbones. Another factor, as previously discussed, was low throughput of experimental validation. To guarantee success of identifying several cross-reactive clones, the scale of antibody expression and testing would have to be significantly increased to the range of 100-500 variants. I expect testing these predictions on a large scale would identify antibodies cross-reactive to HIV and influenza.

Affinity vs. breadth in antibody recognition

The work in this thesis supports the idea that there is a tradeoff between affinity and breadth in antibody recognition of viral variants. Previous computational work has suggested that multi-specific proteins need to achieve a compromise between affinity for one target and recognition of multiple targets (Fromer and Shifman, 2009; Shifman and Mayo, 2002; Willis et al., 2013). The computational multistate design methods developed in Chapter II and Chapter IV made it possible to directly test this hypothesis in the context of antibody-antigen complexes. In Chapter III the experimentally validated C05 mutants were able to increase both breadth and affinity across a seasonal influenza panel. However, the magnitude of the affinity increase was relatively modest, which we attribute to the need to restrict mutations to those which are compatible with all viral strains. When we repeated the design simulations with a single target rather than a panel, we observed that the single-state designs achieved a lower energy than multistate designs. This indicates that there was a tradeoff at play, and that mutants with a greater improvement in affinity

could be achieved, but at the expense of the other viral targets. In addition, using multistate design we were able to identify specific antigenic residues that were accessible to the antibody CDRH3 loop, and were either conserved over a large panel, or when not conserved would not clash with the antibody. An example of such a residue is K125a on the HA surface. Wu *et al.* presented an elegant study using complementary experimental approaches to affinity mature the same mAb, C05 (Wu et al., 2017). They observed that the affinity could be increased for H1 strains only at the cost of affinity for H3 strains, and they identified the specific amino acids responsible for this tradeoff. Taken together, this work indicates that when performing antibody design, multiple viral proteins should be included as target states to ensure that mutations do not negatively impact any of the target panel.

The work in Chapter V supports the same supposition regarding affinity and breadth. When measuring the predicted proclivity of antibody sequences for binding either HIV or influenza, it was clear that increasing breadth for one antigen came at the expense of affinity for the other. We could clearly visualize the tradeoff frontier between affinity for either antigen. In future work it would be interesting to measure the experimental binding affinity of the sequences at this tradeoff frontier to assess whether the binding tradeoff adheres to the anticipated pattern. Although the sequences from human donors tended to adhere to the binding tradeoff frontier, when these sequences were subjected to multistate design they could be improved substantially. This result suggests that naturally evolved antibodies may encounter an affinity-breadth tradeoff prematurely, that can then be further improved by computational design. However, it is also possible that antibodies which are further matured for binding to both targets gain too much breadth, to the point where they are autoreactive against self-antigens and are eliminated by negative selection. In future work the affinity and autoreactivity of computationally designed sequences should be tested.

In Chapter VI, although the computational methods were different I observed a similar affinity-breadth tradeoff. The CDRH3-based cyclic peptides achieved increased breadth for several HAs, notably of the H3 and H7 subtype. I show through computational docking experiments that this increased breadth was achieved through contact of a minimal binding surface on the HA. However, these peptides also exhibited reduced affinity for several HA targets, and in fact lost breadth to three HA strains. This loss of affinity and breadth occurs when chemical interactions at the periphery of the antibody-antigen interface are eliminated to achieve a minimal binding surface. In future work, these cyclic peptides could be redesigned using multistate design to maintain a minimal binding surface while potentially introducing new side chain interactions to increase affinity. As previously mentioned, when a molecule is engineered for increased breadth it is possible that at some point it becomes autoreactive to self-antigens. The peptides described in Chapter VI did exhibit a degree of nonspecific binding to irrelevant viral targets, suggesting they may have off-target effects. However, a direct measurement of binding to human self-antigens was never performed. In future work the binding of cyclic peptides to common self-antigens should be assessed.

Implications for reverse vaccinology

A major motivation for computational design of antibodies is to better understand the principles governing antibody-antigen interactions to enable design of more effective vaccines. The work performed in this thesis aids in this understanding in several respects. In Chapter II I engineered a mutant antibody with increased affinity and breadth. The co-crystal structure of this mutant reveals the amino acids on the antigen that are necessary for binding to this engineered variant. An influenza HA immunogen could be designed that specifically incorporates the residues necessary for C05 variant recognition, with the goal of inducing antibodies similar to the C05

variant. These antibodies would in theory recognize many circulating strains, and a vaccine capable of inducing them at high titer would be very effective. In Chapter VI, the minimal epitope on the surface of influenza HA was identified. The footprint of the cyclic peptides is presumably close to the smallest epitope that can be contacted by an antibody while still maintaining high affinity. An HA immunogen consisting of only this minimal epitope would be a logical candidate to induce cross-reactive antibodies mimicking the activity of the engineered cyclic peptides.

References

Acheson, N.H. (2011). Fundamentals of Molecular Virology (2nd ed.). 1–528.

Adolf-Bryfogle, J., Kalyuzhniy, O., Kubitz, M., Weitzner, B.D., Hu, X., Adachi, Y., Schief, W.R., and Dunbrack, R.L. (2018). RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. PLoS Comput. Biol. *14*, e1006112.

Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput *13*, 3031–3048.

Allcorn, L.C., and Martin, A.C.R. (2002). SACS--self-maintaining database of antibody crystal structure information. Bioinformatics *18*, 175–181.

Allen, B.D., and Mayo, S.L. (2010). An efficient algorithm for multistate protein design based on FASTER. J. Comput. Chem. *31*, 904–916.

Allen, B.D., Nisthal, A., and Mayo, S.L. (2010). Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proc. Natl. Acad. Sci. U.S.a. *107*, 19838–19843.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Ambroggio, X.I., and Kuhlman, B. (2006). Computational Design of a Single Amino Acid Sequence that Can Switch between Two Distinct Protein Folds. J. Am. Chem. Soc. *128*, 1154–1161.

American Academy of Pediatrics Committee on Infectious Diseases (2014). Updated guidance for palivizumab prophylaxis among infants and young children at increased risk of hospitalization for respiratory syncytial virus infection. Pediatrics *134*, 415–420.

Andrews, S. FastQC: A quality control tool for high throughput sequence data. Available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ashworth, J., Taylor, G.K., Havranek, J.J., Quadri, S.A., Stoddard, B.L., and Baker, D. (2010). Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. Nucleic Acids Res. *38*, 5601–5608.

Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J., Stoddard, B.L., and Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. Nature *441*, 656–659.

Azoitei, M.L., Correia, B.E., Ban, Y.-E.A., Carrico, C., Kalyuzhniy, O., Chen, L., Schroeter, A., Huang, P.-S., McLellan, J.S., Kwong, P.D., et al. (2011). Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. Science *334*, 373–376.

Babor, M., and Kortemme, T. (2009). Multi-constraint computational design suggests that native sequences of germline antibody H3 loops are nearly optimal for conformational flexibility. Proteins *75*, 846–858.

Baran, D., Pszolla, M.G., Lapidoth, G.D., Norn, C., Dym, O., Unger, T., Albeck, S., Tyka, M.D., and Fleishman, S.J. (2017). Principles for computational design of binding antibodies. Proc. Natl. Acad. Sci. U.S.a.

Barderas, R., Desmet, J., Timmerman, P., Meloen, R., and Casal, J.I. (2008). Affinity maturation of antibodies assisted by in silico modeling. Proc. Natl. Acad. Sci. U.S.a. *105*, 9029–9034.

Barouch, D.H., Whitney, J.B., Moldt, B., Klein, F., Oliveira, T.Y., Liu, J., Stephenson, K.E., Chang, H.-W., Shekhar, K., Gupta, S., et al. (2013). Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. Nature *503*, 224–228.

Bazzoli, A., and Karanicolas, J. (2017). "Solvent hydrogen-bond occlusion": A new model of polar desolvation for biomolecular energetics. J. Comput. Chem. *38*, 1321–1331.

Belongia, E.A., Simpson, M.D., King, J.P., Sundaram, M.E., Kelley, N.S., Osterholm, M.T., and McLean, H.Q. (2016). Variable influenza vaccine effectiveness by subtype: a systematic review and meta-analysis of test-negative design studies. Lancet Infect Dis *16*, 942–951.

Bender, B.J., Cisneros, A., Duran, A.M., Finn, J.A., Fu, D., Lokits, A.D., Mueller, B.K., Sangha, A.K., Sauer, M.F., Sevy, A.M., et al. (2016). Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. Biochemistry *55*, 4748–4763.

Benichou, J., Glanville, J., Prak, E.T.L., Azran, R., Kuo, T.C., Pons, J., Desmarais, C., Tsaban, L., and Louzoun, Y. (2013). The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. J. Immunol. *190*, 5567–5577.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bhardwaj, G., Mulligan, V.K., Bahl, C.D., and Gilmore, J.M. (2016). Accurate de novo design of hyperstable constrained peptides. Nature.

Binley, J.M., Wrin, T., Korber, B., Zwick, M.B., Wang, M., Chappey, C., Stiegler, G., Kunert, R., Zolla-Pazner, S., Katinger, H., et al. (2004). Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. J. Virol. *78*, 13232–13252.

Bird, G.H., Irimia, A., Ofek, G., Kwong, P.D., Wilson, I.A., and Walensky, L.D. (2014). Stapled HIV-1 peptides recapitulate antigenic structures and engage broadly neutralizing antibodies. Nat. Struct. Mol. Biol. *21*, 1058–1067.

Boder, E.T., Midelfort, K.S., and Wittrup, K.D. (2000). Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. Proceedings of the National Academy of Sciences *97*, 10701–10705.

Bogdanowich-Knipp, S.J., Chakrabarti, S., Williams, T.D., Dillman, R.K., and Siahaan, T.J. (1999). Solution stability of linear vs. cyclic RGD peptides. J. Pept. Res. *53*, 530–541.

Bolon, D.N., Grant, R.A., Baker, T.A., and Sauer, R.T. (2005). Specificity versus stability in computational protein design. Proc. Natl. Acad. Sci. U.S.a. *102*, 12724–12729.

Bonsignori, M., Hwang, K.-K., Chen, X., Tsao, C.-Y., Morris, L., Gray, E., Marshall, D.J., Crump, J.A., Kapiga, S.H., Sam, N.E., et al. (2011). Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. J. Virol. *85*, 9998–10009.

Bonsignori, M., Wiehe, K., Grimm, S.K., Lynch, R., Yang, G., Kozink, D.M., Perrin, F., Cooper, A.J., Hwang, K.-K., Chen, X., et al. (2014). An autoreactive antibody from an SLE/HIV-1 individual broadly neutralizes HIV-1. J. Clin. Invest. *124*, 1835–1843.

Bostrom, J., Yu, S.-F., Kan, D., Appleton, B.A., Lee, C.V., Billeci, K., Man, W., Peale, F., Ross, S., Wiesmann, C., et al. (2009). Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. Science *323*, 1610–1614.

Briney, B., Sok, D., Jardine, J.G., Kulp, D.W., Skog, P., Menis, S., Jacak, R., Kalyuzhniy, O., de Val, N., Sesterhenn, F., et al. (2016). Tailored Immunogens Direct Affinity Maturation toward HIV Neutralizing Antibodies. Cell *166*, 1459–1470.e11.

Bryson, S., Julien, J.-P., Hynes, R.C., and Pai, E.F. (2009). Crystallographic definition of the epitope promiscuity of the broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2F5: vaccine design implications. J. Virol. *83*, 11862–11875.

Calarese, D.A., Scanlan, C.N., Zwick, M.B., Deechongkit, S., Mimura, Y., Kunert, R., Zhu, P., Wormald, M.R., Stanfield, R.L., Roux, K.H., et al. (2003). Antibody domain exchange is an immunological solution to carbohydrate cluster recognition. Science *300*, 2065–2071.

Carter, D.M., Darby, C.A., Lefoley, B.C., Crevar, C.J., Alefantis, T., Oomen, R., Anderson, S.F., Strugnell, T., Cortés-Garcia, G., Vogel, T.U., et al. (2016). Design and Characterization of a Computationally Optimized Broadly Reactive Hemagglutinin Vaccine for H1N1 Influenza Viruses. J. Virol. *90*, 4720–4734.

Carter, P.J., and Lazar, G.A. (2018). Next generation antibody drugs: pursuit of the 'high-hanging fruit'. Nature Publishing Group *17*, 197–223.

Caskey, M., Klein, F., Lorenzi, J.C.C., Seaman, M.S., West, A.P., Buckley, N., Kremer, G., Nogueira, L., Braunschweig, M., Scheid, J.F., et al. (2015). Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. Nature *522*, 487–491.

Casset, F., Roux, F., Mouchet, P., Bes, C., Chardes, T., Granier, C., Mani, J.-C., Pugnière, M., Laune, D., Pau, B., et al. (2003). A peptide mimetic of an anti-CD4 monoclonal antibody by rational design. Biochem. Biophys. Res. Commun. *307*, 198–205.

Caton, A.J., Brownlee, G.G., Yewdell, J.W., and Gerhard, W. (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell *31*, 417–427.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66*, 12–21.

Chevalier, A., Silva, D.-A., Rocklin, G.J., Hicks, D.R., Vergara, R., Murapa, P., Bernard, S.M., Zhang, L., Lam, K.-H., Yao, G., et al. (2017). Massively parallel de novo protein design for targeted therapeutics. Nature *550*, 74–79.

Cho, K.J., Lee, J.-H., Hong, K.W., Kim, S.-H., Park, Y., Lee, J.Y., Kang, S., Kim, S., Yang, J.H., Kim, E.-K., et al. (2013). Insight into structural diversity of influenza virus haemagglutinin. J. Gen. Virol. *94*, 1712–1722.

Chuang, G.-Y., Acharya, P., Schmidt, S.D., Yang, Y., Louder, M.K., Zhou, T., Kwon, Y.D., Pancera, M., Bailer, R.T., Doria-Rose, N.A., et al. (2013). Residue-level prediction of HIV-1 antibody epitopes based on neutralization of diverse viral strains. J. Virol. *87*, 10047–10058.

Clark, L.A., Boriack-Sjodin, P.A., Eldredge, J., Fitch, C., Friedman, B., Hanf, K.J.M., Jarpe, M., Liparoto, S.F., Li, Y., Lugovskoy, A., et al. (2006). Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. Protein Sci. *15*, 949–960.

Combs, S.A., DeLuca, S.L., Deluca, S.H., Lemmon, G.H., Nannemann, D.P., Nguyen, E.D., Willis, J.R., Sheehan, J.H., and Meiler, J. (2013). Small-molecule ligand docking into comparative models with Rosetta. Nat Protoc *8*, 1277–1298.

Conley, A.J., Kessler, J.A., Boots, L.J., Tung, J.S., Arnold, B.A., Keller, P.M., Shaw, A.R., and Emini, E.A. (1994). Neutralization of divergent human immunodeficiency virus type 1 variants and primary isolates by IAM-41-2F5, an anti-gp41 human monoclonal antibody. Proceedings of the National Academy of Sciences *91*, 3348–3352.

Conway, P., Tyka, M.D., DiMaio, F., Konerding, D.E., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Science *23*, 47–55.

Correia, B.E., Ban, Y.-E.A., Holmes, M.A., Xu, H., Ellingson, K., Kraft, Z., Carrico, C., Boni, E., Sather, D.N., Zenobia, C., et al. (2010). Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. Structure *18*, 1116–1126.

Correia, B.E., Bates, J.T., Loomis, R.J., Baneyx, G., Carrico, C., Jardine, J.G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., et al. (2015). Proof of principle for epitope-focused vaccine design. Nature *507*, 201–206.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning.

Corti, D., and Lanzavecchia, A. (2013). Broadly Neutralizing Antiviral Antibodies. Annu. Rev. Immunol. *31*, 705–742.

Corti, D., Bianchi, S., Vanzetta, F., Minola, A., Perez, L., Agatic, G., Guarino, B., Silacci, C., Marcandalli, J., Marsland, B.J., et al. (2013). Cross-neutralization of four paramyxoviruses by a human monoclonal antibody. Nature *501*, 439–443.

Corti, D., Voss, J., Gamblin, S.J., Codoni, G., Macagno, A., Jarrossay, D., Vachieri, S.G., Pinna, D., Minola, A., Vanzetta, F., et al. (2011). A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. Science *333*, 850–856.

Crook, Z.R., Sevilla, G.P., Friend, D., Brusniak, M.-Y., Bandaranayake, A.D., Clarke, M., Gewe, M., Mhyre, A.J., Baker, D., Strong, R.K., et al. (2017). Mammalian display screening of diverse cystine-dense peptides for difficult to drug targets. Nat Commun *8*, 2244.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Dahiyat, B.I., and Mayo, S.L. (1997). De novo protein design: fully automated sequence selection. Science *278*, 82–87.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. The Journal of Chemical Physics *98*, 10089–10092.

Daugherty, P.S., Chen, G., Iverson, B.L., and Georgiou, G. (2000). Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. Proceedings of the National Academy of Sciences *97*, 2029–2034.

Davey, J.A., and Chica, R.A. (2014). Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. Proteins *82*, 771–784.

Davis, I.W., Arendall, W.B., Richardson, D.C., and Richardson, J.S. (2006). The backrub motion: how protein backbone shrugs when a sidechain dances. Structure *14*, 265–274.

Desiderio, S.V., Yancopoulos, G.D., Paskind, M., Thomas, E., Boss, M.A., Landau, N., Alt, F.W., and Baltimore, D. (1984). Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. Nature *311*, 752–755.

Desjarlais, J.R., and Handel, T.M. (1995). De novo design of the hydrophobic cores of proteins. Protein Sci. *4*, 2006–2018.

Dill, K.A., and MacCallum, J.L. (2012). The protein-folding problem, 50 years on. Science *338*, 1042–1046.

Diskin, R., Scheid, J.F., Marcovecchio, P.M., West, A.P., Klein, F., Gao, H., Gnanapragasam, P.N.P., Abadir, A., Seaman, M.S., Nussenzweig, M.C., et al. (2011). Increasing the potency and breadth of an HIV antibody by using structure-based rational design. Science *334*, 1289–1293.

Dreyfus, C., Ekiert, D.C., and Wilson, I.A. (2013). Structure of a classical broadly neutralizing stem antibody in complex with a pandemic H2 influenza virus hemagglutinin. J. Virol. *87*, 7149–7154.

Dreyfus, C., Laursen, N.S., Kwaks, T., Zuijdgeest, D., Khayat, R., Ekiert, D.C., Lee, J.H., Metlagel, Z., Bujny, M.V., Jongeneelen, M., et al. (2012). Highly conserved protective epitopes on influenza B viruses. Science *337*, 1343–1348.

DuBois, R.M., Aguilar-Yañez, J.M., Mendoza-Ochoa, G.I., Oropeza-Almazán, Y., Schultz-Cherry, S., Alvarez, M.M., White, S.W., and Russell, C.J. (2011). The receptor-binding

domain of influenza virus hemagglutinin produced in Escherichia coli folds into its native, immunogenic structure. J. Virol. *85*, 865–872.

Ducatez, M.F., Pelletier, C., and Meyer, G. (2015). Influenza D virus in cattle, France, 2011-2014. Emerging Infect. Dis. *21*, 368–371.

Dunbrack, R.L. (2002). Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. *12*, 431–440.

Edgar, R.C., and Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics *31*, 3476–3482.

Eggink, D., Goff, P.H., and Palese, P. (2014). Guiding the immune response against influenza virus hemagglutinin toward the conserved stalk domain by hyperglycosylation of the globular head domain. J. Virol. *88*, 699–704.

Ekiert, D.C., Bhabha, G., Elsliger, M.-A., Friesen, R.H.E., Jongeneelen, M., Throsby, M., Goudsmit, J., and Wilson, I.A. (2009). Antibody recognition of a highly conserved influenza virus epitope. Science *324*, 246–251.

Ekiert, D.C., Friesen, R.H.E., Bhabha, G., Kwaks, T., Jongeneelen, M., Yu, W., Ophorst, C., Cox, F., Korse, H.J.W.M., Brandenburg, B., et al. (2011). A highly conserved neutralizing epitope on group 2 influenza A viruses. Science *333*, 843–850.

Ekiert, D.C., Kashyap, A.K., Steel, J., Rubrum, A., Bhabha, G., Khayat, R., Lee, J.H., Dillon, M.A., O'Neil, R.E., Faynboym, A.M., et al. (2012). Cross-neutralization of influenza A viruses mediated by a single antibody loop. Nature *489*, 526–532.

Elgundi, Z., Reslan, M., Cruz, E., Sifniotis, V., and Kayser, V. (2017). The state-of-play and future of antibody therapeutics. Advanced Drug Delivery Reviews *122*, 2–19.

Ellebedy, A.H., Krammer, F., Li, G.-M., Miller, M.S., Chiu, C., Wrammert, J., Chang, C.Y., Davis, C.W., McCausland, M., Elbein, R., et al. (2014). Induction of broadly cross-reactive antibody responses to the influenza HA stem region following H5N1 vaccination in humans. Proc. Natl. Acad. Sci. U.S.a. *111*, 13133–13138.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

Entzminger, K.C., Hyun, J.-M., Pantazes, R.J., Patterson-Orazem, A.C., Qerqez, A.N., Frye, Z.P., Hughes, R.A., Ellington, A.D., Lieberman, R.L., Maranas, C.D., et al. (2017). De novo

design of antibody complementarity determining regions binding a FLAG tetra-peptide. Sci Rep *7*, 10295.

Fagète, S., Rousseau, F., Magistrelli, G., Gueneau, F., Ravn, U., Kosco-Vilbois, M.H., and Fischer, N. (2012). Dual specificity of anti-CXCL10-CXCL9 antibodies is governed by structural mimicry. J. Biol. Chem. *287*, 1458–1467.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research *9*, 1871–1874.

Farady, C.J., Sellers, B.D., Jacobson, M.P., and Craik, C.S. (2009). Improving the species cross-reactivity of an antibody using computational design. Bioorg. Med. Chem. Lett. *19*, 3744–3747.

Feller, S.E., Zhang, Y., Pastor, R.W., and Brooks, B.R. (1995). Constant pressure molecular dynamics simulation: The Langevin piston method. The Journal of Chemical Physics *103*, 4613–4621.

Fleishman, S.J., Leaver-Fay, A., Corn, J.E., Strauch, E.-M., Khare, S.D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., et al. (2011a). RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *6*, e20161.

Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.-M., Wilson, I.A., and Baker, D. (2011b). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science *332*, 816–821.

Flynn, N.M., Forthal, D.N., Harro, C.D., Judson, F.N., Mayer, K.H., Para, M.F., rgp120 HIV Vaccine Study Group (2005). Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. J. Infect. Dis. *191*, 654–665.

Foote, J., and Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. Proceedings of the National Academy of Sciences *91*, 10370–10374.

Fromer, M., and Shifman, J.M. (2009). Tradeoff between stability and multispecificity in the design of promiscuous proteins. PLoS Comput. Biol. *5*, e1000627.

Fromer, M., Yanover, C., Harel, A., Shachar, O., Weiss, Y., and Linial, M. (2010). SPRINT: side-chain prediction inference toolbox for multistate protein design. Bioinformatics *26*, 2466–2467.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Fu, Y., Zhang, Z., Sheehan, J., Avnir, Y., Ridenour, C., Sachnik, T., Sun, J., Hossain, M.J., Chen, L.-M., Zhu, Q., et al. (2016). A broadly neutralizing anti-influenza antibody reveals ongoing capacity of haemagglutinin-specific memory B cells to evolve. Nat Commun *7*, 12780–13.

Fung, H.K., Floudas, C.A., Taylor, M.S., Zhang, L., and Morikis, D. (2008). Toward full-sequence de novo protein design with flexible templates for human beta-defensin-2. Biophys. J. *94*, 584–599.

Gamblin, S.J., Haire, L.F., Russell, R.J., Stevens, D.J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D.A., Daniels, R.S., Elliot, A., et al. (2004). The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science *303*, 1838–1842.

Garcia-Rodriguez, C., Levy, R., Arndt, J.W., Forsyth, C.M., Razai, A., Lou, J., Geren, I., Stevens, R.C., and Marks, J.D. (2006). Molecular evolution of antibody cross-reactivity for two subtypes of type A botulinum neurotoxin. Nat. Biotechnol. *25*, 107–116.

Georgiev, I.S., Doria-Rose, N.A., Zhou, T., Kwon, Y.D., Staupe, R.P., Moquin, S., Chuang, G.-Y., Louder, M.K., Schmidt, S.D., Altae-Tran, H.R., et al. (2013). Delineating antibody recognition in polyclonal sera from patterns of HIV-1 isolate neutralization. Science *340*, 751–756.

Gilbert, P.B., McKeague, I.W., Eisen, G., Mullins, C., Guéye-NDiaye, A., Mboup, S., and Kanki, P.J. (2003). Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. Stat Med *22*, 573–593.

Giles, B.M., and Ross, T.M. (2011). A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. Vaccine *29*, 3043–3054.

Giles, B.M., Bissel, S.J., Dealmeida, D.R., Wiley, C.A., and Ross, T.M. (2012a). Antibody breadth and protective efficacy are increased by vaccination with computationally optimized hemagglutinin but not with polyvalent hemagglutinin-based H5N1 virus-like particle vaccines. Clin. Vaccine Immunol. *19*, 128–139.

Giles, B.M., Crevar, C.J., Carter, D.M., Bissel, S.J., Schultz-Cherry, S., Wiley, C.A., and Ross, T.M. (2012b). A computationally optimized hemagglutinin virus-like particle vaccine elicits broadly reactive antibodies that protect nonhuman primates from H5N1 infection. J. Infect. Dis. *205*, 1562–1570.

Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. *331*, 281–299.

Grigoryan, G., Reinke, A.W., and Keating, A.E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature *458*, 859–864.

Guntas, G., Hallett, R.A., Zimmerman, S.P., Williams, T., Yumerefendi, H., Bear, J.E., and Kuhlman, B. (2015). Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins. Proc. Natl. Acad. Sci. U.S.a. *112*, 112–117.

Hahn, B.H., Shaw, G.M., De Cock, K.M., and Sharp, P.M. (2000). AIDS as a zoonosis: scientific and public health implications. Science *287*, 607–614.

Hampson, A.W., and Mackenzie, J.S. (2006). The influenza viruses. Med. J. Aust. *185*, S39–S43.

Hannoun, C. (2013). The evolving history of influenza viruses and influenza vaccines. Expert Rev Vaccines *12*, 1085–1094.

Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., and Kim, P.S. (1998). High-resolution protein design with backbone freedom. Science *282*, 1462–1467.

Harris, L.J., Larson, S.B., Hasel, K.W., and McPherson, A. (1997). Refined structure of an intact IgG2a monoclonal antibody. Biochemistry *36*, 1581–1597.

Havranek, J.J., and Harbury, P.B. (2003). Automated design of specificity in molecular recognition. Nat. Struct. Biol. *10*, 45–52.

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. Trends Mol Med *18*, 182–192.

Hong, M., Lee, P.S., Hoffman, R.M.B., Zhu, X., Krause, J.C., Laursen, N.S., Yoon, S.-I., Song, L., Tussey, L., Crowe, J.E., et al. (2013). Antibody recognition of the pandemic H1N1 Influenza virus hemagglutinin receptor binding site. J. Virol. *87*, 12471–12480.

Hoot, S., McGuire, A.T., Cohen, K.W., Strong, R.K., Hangartner, L., Klein, F., Diskin, R., Scheid, J.F., Sather, D.N., Burton, D.R., et al. (2013). Recombinant HIV envelope proteins fail to engage germline versions of anti-CD4bs bNAbs. PLoS Pathog *9*, e1003106.

Horwitz, J.A., Bar-On, Y., Lu, C.-L., Fera, D., Lockhart, A.A.K., Lorenzi, J.C.C., Nogueira, L., Golijanin, J., Scheid, J.F., Seaman, M.S., et al. (2017). Non-neutralizing Antibodies Alter the Course of HIV-1 Infection In Vivo. Cell *170*, 637–648.e10.

Howell, S.C., Inampudi, K.K., Bean, D.P., and Wilson, C.J. (2014). Understanding thermal adaptation of enzymes through the multistate rational design and stability prediction of 100 adenylate kinases. Structure *22*, 218–229.

Hu, X., Wang, H., Ke, H., and Kuhlman, B. (2007). High-resolution design of a protein loop. Proc. Natl. Acad. Sci. U.S.a. *104*, 17668–17673.

Huang, J., Ofek, G., Laub, L., Louder, M.K., Doria-Rose, N.A., Longo, N.S., Imamichi, H., Bailer, R.T., Chakrabarti, B., Sharma, S.K., et al. (2012). Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. Nature *491*, 406–412.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J Mol Graph *14*, 33–8–27–8.

Humphris, E.L., and Kortemme, T. (2007). Design of multi-specificity in protein interfaces. *3*, e164.

Humphris, E.L., and Kortemme, T. (2008). Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. Structure *16*, 1777–1788.

Impagliazzo, A., Milder, F., Kuipers, H., Wagner, M.V., Zhu, X., Hoffman, R.M.B., van Meersbergen, R., Huizingh, J., Wanningen, P., Verspuij, J., et al. (2015). A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. Science *349*, 1301–1306.

Ionescu, R.M., Vlasak, J., Price, C., and Kirchmeier, M. (2008). Contribution of Variable Domains to the Stability of Humanized IgG1 Monoclonal Antibodies. Journal of Pharmaceutical Sciences *97*, 1414–1426.

Isobe, M., Huebner, K., Erikson, J., Peterson, R.C., Bollum, F.J., Chang, L.M., and Croce, C.M. (1985). Chromosome localization of the gene for human terminal deoxynucleotidyltransferase to region 10q23-q25. Proceedings of the National Academy of Sciences *82*, 5836–5840.

James, L.C., Roversi, P., and Tawfik, D.S. (2003). Antibody multispecificity mediated by conformational diversity. Science *299*, 1362–1367.

Jardine, J.G., Kulp, D.W., Havenar-Daughton, C., Sarkar, A., Briney, B., Sok, D., Sesterhenn, F., Ereño-Orbea, J., Kalyuzhniy, O., Deresa, I., et al. (2016). HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. Science *351*, 1458–1463.

Jardine, J.G., Ota, T., Sok, D., Pauthner, M., Kulp, D.W., Kalyuzhniy, O., Skog, P.D., Thinnes, T.C., Bhullar, D., Briney, B., et al. (2015). Priming a broadly neutralizing antibody response to HIV-1 using a germline-targeting immunogen. Science *349*, 156–161.

Jardine, J., Julien, J.-P., Menis, S., Ota, T., Kalyuzhniy, O., McGuire, A., Sok, D., Huang, P.-S., Macpherson, S., Jones, M., et al. (2013). Rational HIV immunogen design to target specific germline B cell receptors. Science *340*, 711–716.

Jefferson, T., Jones, M.A., Doshi, P., Del Mar, C.B., Hama, R., Thompson, M.J., Spencer, E.A., Onakpoya, I., Mahtani, K.R., Nunan, D., et al. (2014). Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children. Cochrane Database Syst Rev *19*, CD008965.

Jeliazkov, J.R., Sljoka, A., Kuroda, D., Tsuchimura, N., Katoh, N., Tsumoto, K., and Gray, J.J. (2018). Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. Front Immunol *9*, 413.

Jolly, C.J., Wagner, S.D., Rada, C., Klix, N., Milstein, C., and Neuberger, M.S. (1996). The targeting of somatic hypermutation. Semin. Immunol. *8*, 159–168.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics *79*, 926–935.

Kaas, Q., Ruiz, M., and Lefranc, M.-P. (2004). IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res. *32*, D208–D210.

Kadam, R.U., and Wilson, I.A. (2018). A small-molecule fragment that emulates binding of receptor and broadly neutralizing antibodies to influenza A hemagglutinin. Proc. Natl. Acad. Sci. U.S.a.

Kadam, R.U., Juraszek, J., Brandenburg, B., Buyck, C., Schepens, W.B.G., Kesteleyn, B., Stoops, B., Vreeken, R.J., Vermond, J., Goutier, W., et al. (2017). Potent peptidic fusion inhibitors of influenza virus. Science *358*, 496–502.

Kamisetty, H., Ghosh, B., Langmead, C.J., and Bailey-Kellogg, C. (2015). Learning sequence determinants of protein:protein interaction specificity with sparse graphical models. J. Comput. Biol. *22*, 474–486.

Kapp, G.T., Liu, S., Stein, A., Wong, D.T., Reményi, A., Yeh, B.J., Fraser, J.S., Taunton, J., Lim, W.A., and Kortemme, T. (2012). Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. Proc. Natl. Acad. Sci. U.S.a. *109*, 5277–5282.

Khan, T.A., Friedensohn, S., Gorter de Vries, A.R., Straszewski, J., Ruscheweyh, H.-J., and Reddy, S.T. (2016). Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. Sci Adv *2*, e1501371–e1501371.

King, N.P., Sheffler, W., Sawaya, M.R., Vollmar, B.S., Sumida, J.P., André, I., Gonen, T., Yeates, T.O., and Baker, D. (2012). Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science *336*, 1171–1174.

Klein, E.Y., Serohijos, A.W.R., Choi, J.-M., Shakhnovich, E.I., and Pekosz, A. (2014). Influenza A H1N1 pandemic strain evolution--divergence and the potential for antigenic drift variants. PLoS ONE *9*, e93632.

Klein, F., Diskin, R., Scheid, J.F., Gaebler, C., Mouquet, H., Georgiev, I.S., Pancera, M., Zhou, T., Incesu, R.-B., Fu, B.Z., et al. (2013). Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. Cell *153*, 126–138.

Kortemme, T., Joachimiak, L.A., Bullock, A.N., Schuler, A.D., Stoddard, B.L., and Baker, D. (2004). Computational redesign of protein-protein interaction specificity. Nat. Struct. Mol. Biol. *11*, 371–379.

Kosuri, S., and Church, G.M. (2014). Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods *11*, 499–507.

Kramer, A., Keitel, T., Winkler, K., Stöcklein, W., Höhne, W., and Schneider-Mergener, J. (1997). Molecular basis for the binding promiscuity of an anti-p24 (HIV-1) monoclonal antibody. Cell *91*, 799–809.

Krammer, F., and Palese, P. (2013). Influenza virus hemagglutinin stalk-based antibodies and vaccines. Curr Opin Virol *3*, 521–530.

Krause, J.C., and Crowe, J.E. (2014). Committing the Oldest Sins in the Newest Kind of Ways-Antibodies Targeting the Influenza Virus Type A Hemagglutinin Globular Head. Microbiol Spectr *2*.

Krause, J.C., Tsibane, T., Tumpey, T.M., Huffman, C.J., Albrecht, R., Blum, D.L., Ramos, I., Fernandez-Sesma, A., Edwards, K.M., García-Sastre, A., et al. (2012). Human monoclonal antibodies to pandemic 1957 H2N2 and pandemic 1968 H3N2 influenza viruses. J. Virol. *86*, 6334–6340.

Krause, J.C., Tsibane, T., Tumpey, T.M., Huffman, C.J., Basler, C.F., and Crowe, J.E. (2011). A broadly neutralizing human monoclonal antibody that recognizes a conserved, novel epitope on the globular head of the influenza H1N1 virus hemagglutinin. J. Virol. *85*, 10905–10908.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. U.S.a. *97*, 10383–10388.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science *302*, 1364–1368.

Kumar, B., Asha, K., Khanna, M., Ronsard, L., Meseko, C.A., and Sanicas, M. (2018). The emerging influenza virus threat: status and new prospects for its therapy and control. Arch. Virol. *163*, 831–844.

Lapidoth, G.D., Baran, D., Pszolla, G.M., Norn, C., Alon, A., Tyka, M.D., and Fleishman, S.J. (2015). AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. Proteins *83*, 1385–1406.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Lazaridis, T., and Karplus, M. (1999). Effective energy function for proteins in solution. Proteins *35*, 133–152.

Leaver-Fay, A., Froning, K.J., Atwell, S., Aldaz, H., Pustilnik, A., Lu, F., Huang, F., Yuan, R., Hassanali, S., Chamberlain, A.K., et al. (2016). Computationally Designed Bispecific Antibodies using Negative State Repertoires. Structure *24*, 641–651.

Leaver-Fay, A., Jacak, R., Stranges, P.B., and Kuhlman, B. (2011a). A generic program for multistate protein design. *6*, e20937.

Leaver-Fay, A., O'Meara, M.J., Tyka, M., Jacak, R., Song, Y., Kellogg, E.H., Thompson, J., Davis, I.W., Pache, R.A., Lyskov, S., et al. (2013). Scientific benchmarks for guiding macromolecular energy function improvement. Meth. Enzymol. *523*, 109–143.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011b). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Meth. Enzymol. *487*, 545–574.

Lee, P.S., Ohshima, N., Stanfield, R.L., Yu, W., Iba, Y., Okuno, Y., Kurosawa, Y., and Wilson, I.A. (2014). Receptor mimicry by antibody F045-092 facilitates universal binding to the H3 subtype of influenza virus. Nat Commun *5*, 3614.

Lee, P.S., Yoshida, R., Ekiert, D.C., Sakai, N., Suzuki, Y., Takada, A., and Wilson, I.A. (2012). Heterosubtypic antibody recognition of the influenza virus hemagglutinin receptor binding site enhanced by avidity. Proc. Natl. Acad. Sci. U.S.a. *109*, 17040–17045.

Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clément, O., Chaume, D., and Lefranc, G. (2005). IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. *33*, D593–D597.

Lehmann, A., Wixted, J.H.F., Shapovalov, M.V., Roder, H., Dunbrack, R.L., and Robinson, M.K. (2015). Stability engineering of anti-EGFR scFv antibodies by rational design of a lambda-to-kappa swap of the VL framework using a structure-guided approach. MAbs *7*, 1058–1071.

Levi, M., Sällberg, M., Rudén, U., Herlyn, D., Maruyama, H., Wigzell, H., Marks, J., and Wahren, B. (1993). A complementarity-determining region synthetic peptide acts as a miniantibody and neutralizes human immunodeficiency virus type 1 in vitro. Proceedings of the National Academy of Sciences *90*, 4374–4378.

Levinthal, C. (1969). How to fold graciously. Mossbaun Spectroscopy in Biological Systems Proceedings 1–3.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. *104*, 59–107.

Levitt, M., and Warshel, A. (1975). Computer simulation of protein folding. Nature *253*, 694–698.

Lewis, S.M., Wu, X., Pustilnik, A., Sereno, A., Huang, F., Rick, H.L., Guntas, G., Leaver-Fay, A., Smith, E.M., Ho, C., et al. (2014). Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface. Nat. Biotechnol. *32*, 191–198.

Liao, H.-X., Chen, X., Munshaw, S., Zhang, R., Marshall, D.J., Vandergrift, N., Whitesides, J.F., Lu, X., Yu, J.-S., Hwang, K.-K., et al. (2011). Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated. J. Exp. Med. *208*, 2237–2249.

Liao, H.-X., Lynch, R., Zhou, T., Gao, F., Alam, S.M., Boyd, S.D., Fire, A.Z., Roskin, K.M., Schramm, C.A., Zhang, Z., et al. (2013). Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature *496*, 469–476.

Lin, T., Wang, G., Li, A., Zhang, Q., Wu, C., Zhang, R., Cai, Q., Song, W., and Yuen, K.-Y. (2009). The hemagglutinin structure of an avian H1N1 influenza A virus. Virology *392*, 73–81.

Lippow, S.M., Wittrup, K.D., and Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. Nat. Biotechnol. *25*, 1171–1176.

Liu, J., Stevens, D.J., Haire, L.F., Walker, P.A., Coombs, P.J., Russell, R.J., Gamblin, S.J., and Skehel, J.J. (2009). Structures of receptor complexes formed by hemagglutinins from the Asian Influenza pandemic of 1957. Proc. Natl. Acad. Sci. U.S.a. *106*, 17175–17180.

Liu, M., Yang, G., Wiehe, K., Nicely, N.I., Vandergrift, N.A., Rountree, W., Bonsignori, M., Alam, S.M., Gao, J., Haynes, B.F., et al. (2015). Polyreactivity and autoreactivity among HIV-1 antibodies. J. Virol. *89*, 784–798.

Mandell, D.J., and Kortemme, T. (2009). Backbone flexibility in computational protein design. Curr. Opin. Biotechnol. *20*, 420–428.

Markham, A. (2018). Ibalizumab: First Global Approval. Drugs *78*, 781–785.

Martyna, G.J., Tobias, D.J., and Klein, M.L. (1994). Constant pressure molecular dynamics algorithms. The Journal of Chemical Physics *101*, 4177–4189.

Marvin, J.S., and Lowman, H.B. (2003). Redesigning an antibody fragment for faster association with its antigen. Biochemistry *42*, 7077–7083.

Marze, N.A., Jeliazkov, J.R., Roy Burman, S.S., Boyken, S.E., DiMaio, F., and Gray, J.J. (2016). Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28-35. Proteins *85*, 479–486.

Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K.I., Hayashida, H., Miyata, T., and Honjo, T. (1998). The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J. Exp. Med. *188*, 2151–2162.

McCarthy, K.R., Watanabe, A., Kuraoka, M., Do, K.T., McGee, C.E., Sempowski, G.D., Kepler, T.B., Schmidt, A.G., Kelsoe, G., and Harrison, S.C. (2018). Memory B Cells that Cross-React with Group 1 and Group 2 Influenza A Viruses Are Abundant in Adult Human Repertoires. Immunity *48*, 174–184.e179.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J Appl Crystallogr *40*, 658–674.

McGuire, A.T., Glenn, J.A., Lippy, A., and Stamatatos, L. (2014). Diverse recombinant HIV-1 Envs fail to activate B cells expressing the germline B cell receptors of the broadly neutralizing anti-HIV-1 antibodies PG9 and 447-52D. J. Virol. *88*, 2645–2657.

McLean, G.R., Nakouzi, A., Casadevall, A., and Green, N.S. (2000). Human and murine immunoglobulin expression vector cassettes. Mol. Immunol. *37*, 837–845.

Midelfort, K.S., Hernandez, H.H., Lippow, S.M., Tidor, B., Drennan, C.L., and Wittrup, K.D. (2004). Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. J. Mol. Biol. *343*, 685–701.

Mikell, I., Sather, D.N., Kalams, S.A., Altfeld, M., Alter, G., and Stamatatos, L. (2011). Characteristics of the earliest cross-neutralizing antibody response to HIV-1. PLoS Pathog *7*, e1001251.

Miklos, A.E., Kluwe, C., Der, B.S., Pai, S., Sircar, A., Hughes, R.A., Berrondo, M., Xu, J., Codrea, V., Buckley, P.E., et al. (2012). Structure-based design of supercharged, highly thermoresistant antibodies. Chem. Biol. *19*, 449–455.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. (1992). Stereochemical quality of protein structure coordinates. Proteins *12*, 345–364.

Mouquet, H., Scheid, J.F., Zoller, M.J., Krogsgaard, M., Ott, R.G., Shukair, S., Artyomov, M.N., Pietzsch, J., Connors, M., Pereyra, F., et al. (2010). Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. Nature *467*, 591–595.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell *102*, 553–563.

Murphy, K., Janeway, C.A., Travers, P., and Wallport, M. (2012). Janeway's Immunobiology (Garland Science).

Murshudov, G.N., Skubák, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F., and Vagin, A.A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr. D Biol. Crystallogr. *67*, 355–367.

Nakajima, S., Nakajima, K., and Kendal, A.P. (1983). Identification of the binding sites to monoclonal antibodies on A/USSR/90/77 (H1N1) hemagglutinin and their involvement in antigenic drift in H1N1 influenza viruses. Virology *131*, 116–127.

Nelson, J.D., Brunel, F.M., Jensen, R., Crooks, E.T., Cardoso, R.M.F., Wang, M., Hessell, A., Wilson, I.A., Binley, J.M., Dawson, P.E., et al. (2007). An affinity-enhanced neutralizing antibody against the membrane-proximal external region of human immunodeficiency virus type 1 gp41 recognizes an epitope between those of 2F5 and 4E10. J. Virol. *81*, 4033–4043.

Nivón, L.G., Moretti, R., and Baker, D. (2013). A Pareto-optimal refinement method for protein design scaffolds. *8*, e59004.

O'Meara, M.J., Leaver-Fay, A., Tyka, M.D., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., et al. (2015). Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput *11*, 609–622.

Oettinger, M.A., Schatz, D.G., Gorka, C., and Baltimore, D. (1990). RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science *248*, 1517–1523.

Ofek, G., Guenaga, F.J., Schief, W.R., Skinner, J., Baker, D., Wyatt, R., and Kwong, P.D. (2010). Elicitation of structure-specific antibodies by epitope scaffolds. Proc. Natl. Acad. Sci. U.S.a. *107*, 17880–17887.

Okuno, Y., Isegawa, Y., Sasao, F., and Ueda, S. (1993). A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. J. Virol. *67*, 2552–2558.

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

Owens, A.E., de Paola, I., Hansen, W.A., Liu, Y.-W., Khare, S.D., and Fasan, R. (2017). Design and Evolution of a Macrocyclic Peptide Inhibitor of the Sonic Hedgehog/Patched Interaction. J. Am. Chem. Soc. *139*, 12559–12568.

Pappas, L., Foglierini, M., Piccoli, L., Kallewaard, N.L., Turrini, F., Silacci, C., Fernandez-Rodriguez, B., Agatic, G., Giacchetto-Sasselli, I., Pellicciotta, G., et al. (2014). Rapid development of broadly influenza neutralizing antibodies through redundant mutations. Nature *516*, 418–422.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Pejchal, R., Gach, J.S., Brunel, F.M., Cardoso, R.M., Stanfield, R.L., Dawson, P.E., Burton, D.R., Zwick, M.B., and Wilson, I.A. (2009). A conformational switch in human immunodeficiency virus gp41 revealed by the structures of overlapping epitopes recognized by neutralizing antibodies. J. Virol. *83*, 8451–8462.

Petrova, V.N., and Russell, C.A. (2018). The evolution of seasonal influenza viruses. Nature Publishing Group *16*, 47–60.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802.

Pitisuttithum, P., Gilbert, P., Gurwith, M., Heyward, W., Martin, M., van Griensven, F., Hu, D., Tappero, J.W., Choopanya, K., Bangkok Vaccine Evaluation Group (2006). Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. J. Infect. Dis. *194*, 1661–1671.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *33*, D501–D504.

Rajewsky, K., Förster, I., and Cumano, A. (1987). Evolutionary and somatic selection of the antibody repertoire in the mouse. Science *238*, 1088–1094.

Reeves, J.D., and Doms, R.W. (2002). Human immunodeficiency virus type 2. J. Gen. Virol. *83*, 1253–1265.

Rocklin, G.J., Chidyausiku, T.M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V.K., Chevalier, A., et al. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. Science *357*, 168–175.

Roehr, J.T., Dieterich, C., and Reinert, K. (2017). Flexbar 3.0 - SIMD and multicore parallelization. Bioinformatics *33*, 2941–2942.

Rolfes, M.A., Foppa, I.M., Garg, S., and Flannery, B. (2016). Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States. Https://Www.Cdc.Gov/Flu/About/Disease/2015-16.Htm.

Rubinstein, N.D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J.M., and Pupko, T. (2008). Computational characterization of B-cell epitopes. Mol. Immunol. *45*, 3477–3489.

Russell, C.J. (2014). Acid-induced membrane fusion by the hemagglutinin protein and its role in influenza virus biology. Curr. Top. Microbiol. Immunol. *385*, 93–116.

Russell, R.J., Kerry, P.S., Stevens, D.J., Steinhauer, D.A., Martin, S.R., Gamblin, S.J., and Skehel, J.J. (2008). Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion. Proc. Natl. Acad. Sci. U.S.a. *105*, 17736–17741.

Sahini, L., Tempczyk-Russell, A., and Agarwal, R. (2010). Large-scale sequence analysis of hemagglutinin of influenza A virus identifies conserved regions suitable for targeting an anti-viral response. PLoS ONE *5*, e9268.

Sandelin, A., and Wasserman, W.W. (2004). Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J. Mol. Biol. *338*, 207–215.

Saphire, E.O., Parren, P.W., Pantophlet, R., Zwick, M.B., Morris, G.M., Rudd, P.M., Dwek, R.A., Stanfield, R.L., Burton, D.R., and Wilson, I.A. (2001). Crystal structure of a neutralizing human IGG against HIV-1: a template for vaccine design. Science *293*, 1155–1159.

Sather, D.N., Armann, J., Ching, L.K., Mavrantoni, A., Sellhorn, G., Caldwell, Z., Yu, X., Wood, B., Self, S., Kalams, S., et al. (2009). Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. J. Virol. *83*, 757–769.

Schatz, D.G., Oettinger, M.A., and Baltimore, D. (1989). The V(D)J recombination activating gene, RAG-1. Cell *59*, 1035–1048.

Scheid, J.F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T.Y.K., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., et al. (2011). Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science *333*, 1633–1637.

Schmidt, A.G., Therkelsen, M.D., Stewart, S., Kepler, T.B., Liao, H.-X., Moody, M.A., Haynes, B.F., and Harrison, S.C. (2015). Viral receptor-binding site antibodies with diverse germline origins. Cell *161*, 1026–1034.

Schmidt, A.G., Xu, H., Khan, A.R., O'Donnell, T., Khurana, S., King, L.R., Manischewitz, J., Golding, H., Suphaphiphat, P., Carfi, A., et al. (2013). Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. Proc. Natl. Acad. Sci. U.S.a. *110*, 264–269.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Schrodinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.7.

Sethi, D.K., Agarwal, A., Manivel, V., Rao, K.V.S., and Salunke, D.M. (2006). Differential Epitope Positioning within the Germline Antibody Paratope Enhances Promiscuity in the Primary Immune Response. Immunity *24*, 429–438.

Sevy, A.M., and Meiler, J. (2014). Antibodies: Computer-Aided Prediction of Structure and Design of Function. Microbiol Spectr *2*.

Sevy, A.M., Jacobs, T.M., Crowe, J.E., and Meiler, J. (2015). Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. PLoS Comput. Biol. *11*, e1004300.

Sevy, A.M., Panda, S., Crowe, J.E., Meiler, J., and Vorobeychik, Y. (2018). Integrating linear optimization with structural modeling to increase HIV neutralization breadth. PLoS Comput. Biol. *14*, e1005999.

Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., et al. (2010). Atomic-level characterization of the structural dynamics of proteins. Science *330*, 341–346.

Shifman, J.M., and Mayo, S.L. (2002). Modulating calmodulin binding specificity through computational protein design. J. Mol. Biol. *323*, 417–423.

Shih, H.H., Tu, C., Cao, W., Klein, A., Ramsey, R., Fennell, B.J., Lambert, M., Ní Shúilleabháin, D., Autin, B., Kouranova, E., et al. (2012). An ultra-specific avian antibody to phosphorylated tau protein reveals a unique mechanism for phosphoepitope recognition. J. Biol. Chem. *287*, 44425–44434.

Shingai, M., Nishimura, Y., Klein, F., Mouquet, H., Donau, O.K., Plishka, R., Buckler-White, A., Seaman, M., Piatak, M., Lifson, J.D., et al. (2013). Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. Nature *503*, 277–280.

Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St Clair, J.L., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., et al. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science *329*, 309–313.

Simonelli, L., Pedotti, M., Beltramello, M., Livoti, E., Calzolai, L., Sallusto, F., Lanzavecchia, A., and Varani, L. (2013). Rational engineering of a human anti-dengue antibody through experimentally validated computational docking. *8*, e55561.

Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins *Suppl 3*, 171–176.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. *213*, 859–883.

Skehel, J.J., and Wiley, D.C. (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. *69*, 531–569.

Smith, C.A., and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol. *380*, 742–756.

Smith, C.A., and Kortemme, T. (2011). Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *6*, e20451.

Song, L., Sun, Z.-Y.J., Coleman, K.E., Zwick, M.B., Gach, J.S., Wang, J.-H., Reinherz, E.L., Wagner, G., and Kim, M. (2009). Broadly neutralizing anti-HIV-1 antibodies disrupt a hinge-related function of gp41 at the membrane interface. Proc. Natl. Acad. Sci. U.S.a. *106*, 9057–9062.

Song, Y., DiMaio, F., Wang, R.Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013). High-resolution comparative modeling with RosettaCM. Structure *21*, 1735–1742.

Soto, C., Bombardi, R.G., Sinkovits, R.S., Branchizio, A., Kose, N., Matta, P., Gilchuk, P., Finn, J.A., Sevy, A.M., Mallal, S., et al. (2018a). High frequency of shared clonotypes in human B and T cell receptor repertoires. Submitted.

Soto, C., Finn, J.A., Bombardi, R.G., Willis, J.R., Götz, A.W., Sinkovits, R.S., Branchizio, A., and Crowe, J.E., Jr. (2018b). PyIR: a scalable wrapper for processing millions of immunoglobulin and T cell receptor sequences using IgBLAST. Submitted.

Stanfield, R.L., Julien, J.-P., Pejchal, R., Gach, J.S., Zwick, M.B., and Wilson, I.A. (2011). Structure-based design of a protein immunogen that displays an HIV-1 gp41 neutralizing epitope. J. Mol. Biol. *414*, 460–476.

Stein, A., and Kortemme, T. (2013). Improvements to robotics-inspired conformational sampling in rosetta. *8*, e63090.

Stranges, P.B., and Kuhlman, B. (2013). A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. Protein Science *22*, 74–82.

Strauch, E.-M., Bernard, S.M., La, D., Bohn, A.J., Lee, P.S., Anderson, C.E., Nieusma, T., Holstein, C.A., Garcia, N.K., Hooper, K.A., et al. (2017). Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. Nat. Biotechnol. *35*, 667–671.

Strebel, K. (2013). HIV accessory proteins versus host restriction factors. Curr Opin Virol *3*, 692–699.

Sui, J., Hwang, W.C., Perez, S., Wei, G., Aird, D., Chen, L.-M., Santelli, E., Stec, B., Cadwell, G., Ali, M., et al. (2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nat. Struct. Mol. Biol. *16*, 265–273.

Sui, J., Sheehan, J., Hwang, W.C., Bankston, L.A., Burchett, S.K., Huang, C.-Y., Liddington, R.C., Beigel, J.H., and Marasco, W.A. (2011). Wide prevalence of heterosubtypic broadly neutralizing human anti-influenza A antibodies. Clin. Infect. Dis. *52*, 1003–1009.

Sun, M.G.F., Seo, M.-H., Nim, S., Corbi-Verge, C., and Kim, P.M. (2016). Protein engineering by highly parallel screening of computationally designed variants. Sci Adv *2*, e1600692–e1600692.

Tapryal, S., Gaur, V., Kaur, K.J., and Salunke, D.M. (2013). Structural Evaluation of a Mimicry-Recognizing Paratope: Plasticity in Antigen-Antibody Interactions Manifests in Molecular Mimicry. J. Immunol. *191*, 456–463.

Taubenberger, J.K., and Morens, D.M. (2006). 1918 Influenza: the mother of all pandemics. Emerging Infect. Dis. *12*, 15–22.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*, D158–D169.

Thomas, J.W. (1993). V region diversity in human anti-insulin antibodies. Preferential use of a VHIII gene subset. The Journal of Immunology *150*, 1375–1382.

Thornburg, N.J., Nannemann, D.P., Blum, D.L., Belser, J.A., Tumpey, T.M., Deshpande, S., Fritz, G.A., Sapparapu, G., Krause, J.C., Lee, J.H., et al. (2013). Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. J. Clin. Invest. *123*, 4405–4409.

Tomaras, G.D., Yates, N.L., Liu, P., Qin, L., Fouda, G.G., Chavez, L.L., deCamp, A.C., Parks, R.J., Ashley, V.C., Lucas, J.T., et al. (2008). Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. J. Virol. *82*, 12449–12463.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

UNAIDS (2017). Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic.

Underwood, P.A. (1982). Mapping of antigenic changes in the haemagglutinin of Hong Kong influenza (H3N2) strains using a large panel of monoclonal antibodies. J. Gen. Virol. *62 (Pt 1)*, 153–169.

van Dongen, J.J.M., Langerak, A.W., Brüggemann, M., Evans, P.A.S., Hummel, M., Lavender, F.L., Delabesse, E., Davi, F., Schuuring, E., García-Sanz, R., et al. (2003). Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia *17*, 2257–2317.

van Rosmalen, M., Janssen, B.M.G., Hendrikse, N.M., van der Linden, A.J., Pieters, P.A., Wanders, D., de Greef, T.F.A., and Merkx, M. (2017). Affinity Maturation of a Cyclic Peptide Handle for Therapeutic Antibodies Using Deep Mutational Scanning. J. Biol. Chem. *292*, 1477–1489.

Vermeer, A.W., and Norde, W. (2000). The thermal stability of immunoglobulin: unfolding and aggregation of a multi-domain protein. Biophys. J. *78*, 394–404.

Victora, G.D., and Nussenzweig, M.C. (2012). Germinal Centers. Annu. Rev. Immunol. *30*, 429–457.

Walker, L.M., Phogat, S.K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J.L., Wrin, T., Simek, M.D., Fling, S., Mitcham, J.L., et al. (2009). Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science *326*, 285–289.

Wang, B., DeKosky, B.J., Timm, M.R., Lee, J., Normandin, E., Misasi, J., Kong, R., McDaniel, J.R., Delidakis, G., Leigh, K.E., et al. (2018). Functional interrogation and mining of natively paired human VH:VLantibody repertoires. Nat. Biotechnol. *36*, 152–155.

Wang, F., Sen, S., Zhang, Y., Ahmad, I., Zhu, X., Wilson, I.A., Smider, V.V., Magliery, T.J., and Schultz, P.G. (2013). Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation. Proc. Natl. Acad. Sci. U.S.a. *110*, 4261–4266.

Wang, S., Liu, M., Zeng, D., Qiu, W., Ma, P., Yu, Y., Chang, H., and Sun, Z. (2014). Increasing stability of antibody via antibody engineering: stability engineering on an anti-hVEGF. Proteins *82*, 2620–2630.

Ward, A.B., and Wilson, I.A. (2015). Insights into the trimeric HIV-1 envelope glycoprotein structure. Trends Biochem. Sci. *40*, 101–107.

Whittle, J.R.R., Zhang, R., Khurana, S., King, L.R., Manischewitz, J., Golding, H., Dormitzer, P.R., Haynes, B.F., Walter, E.B., Moody, M.A., et al. (2011). Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. Proc. Natl. Acad. Sci. U.S.a. *108*, 14216–14221.

Wilen, C.B., Tilton, J.C., and Doms, R.W. (2012). HIV: cell binding and entry. Cold Spring Harb Perspect Med *2*, a006866–a006866.

Wiley, D.C., Wilson, I.A., and Skehel, J.J. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature *289*, 373–378.

Williams, W.B., Liao, H.-X., Moody, M.A., Kepler, T.B., Alam, S.M., Gao, F., Wiehe, K., Trama, A.M., Jones, K., Zhang, R., et al. (2015). Diversion of HIV-1 vaccine-induced immunity by gp41-microbiota cross-reactive antibodies. Science *349*, aab1253–aab1253.

Willis, J.R., Briney, B.S., DeLuca, S.L., Crowe, J.E., and Meiler, J. (2013). Human germline antibody gene segments encode polyspecific antibodies. PLoS Comput. Biol. *9*, e1003045.

Willis, J.R., Finn, J.A., Briney, B., Sapparapu, G., Singh, V., King, H., LaBranche, C.C., Montefiori, D.C., Meiler, J., and Crowe, J.E. (2016). Long antibody HCDR3s from HIV-naïve donors presented on a PG9 neutralizing antibody background mediate HIV neutralization. Proc. Natl. Acad. Sci. U.S.a. *113*, 4446–4451.

Willis, J.R., Sapparapu, G., Murrell, S., Julien, J.-P., Singh, V., King, H.G., Xia, Y., Pickens, J.A., LaBranche, C.C., Slaughter, J.C., et al. (2015). Redesigned HIV antibodies exhibit enhanced neutralizing potency and breadth. J. Clin. Invest. *125*, 2523–2531.

Winarski, K.L., Thornburg, N.J., Yu, Y., Sapparapu, G., Crowe, J.E., and Spiller, B.W. (2015). Vaccine-elicited antibody that neutralizes H5N1 influenza and variants binds the receptor site and polymorphic sites. Proc. Natl. Acad. Sci. U.S.a. *112*, 9346–9351.

Wong, T.M., Allen, J.D., Bebin-Blackwell, A.-G., Carter, D.M., Alefantis, T., DiNapoli, J., Kleanthous, H., and Ross, T.M. (2017). Computationally Optimized Broadly Reactive Hemagglutinin Elicits Hemagglutination Inhibition Antibodies against a Panel of H3N2 Influenza Virus Cocirculating Variants. J. Virol. *91*.

Wrammert, J., Koutsonanos, D., Li, G.-M., Edupuganti, S., Sui, J., Morrissey, M., McCausland, M., Skountzou, I., Hornig, M., Lipkin, W.I., et al. (2011). Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. J. Exp. Med. *208*, 181–193.

Wu, H., Pfarr, D.S., Johnson, S., Brewah, Y.A., Woods, R.M., Patel, N.K., White, W.I., Young, J.F., and Kiener, P.A. (2007). Development of motavizumab, an ultra-potent antibody for the prevention of respiratory syncytial virus infection in the upper and lower respiratory tract. J. Mol. Biol. *368*, 652–665.

Wu, N.C., and Wilson, I.A. (2017). A Perspective on the Structural and Functional Constraints for Immune Evasion: Insights from Influenza Virus. J. Mol. Biol. *429*, 2694–2709.

Wu, N.C., Grande, G., Turner, H.L., Ward, A.B., Xie, J., Lerner, R.A., and Wilson, I.A. (2017). In vitro evolution of an influenza broadly neutralizing antibody is modulated by hemagglutinin receptor specificity. Nat Commun *8*, 15371.

Wu, N.C., Thompson, A.J., Xie, J., Lin, C.-W., Nycholat, C.M., Zhu, X., Lerner, R.A., Paulson, J.C., and Wilson, I.A. (2018). A complex epistatic network limits the mutational reversibility in the influenza hemagglutinin receptor-binding site. Nat Commun *9*, 1264.

Wu, X., Yang, Z.-Y., Li, Y., Hogerkorp, C.-M., Schief, W.R., Seaman, M.S., Zhou, T., Schmidt, S.D., Wu, L., Xu, L., et al. (2010). Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. Science *329*, 856–861.

Wu, X., Zhang, Z., Schramm, C.A., Joyce, M.G., Do Kwon, Y., Zhou, T., Sheng, Z., Zhang, B., O'Dell, S., McKee, K., et al. (2015). Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. Cell *161*, 470–485.

Xiao, X., Chen, W., Feng, Y., Zhu, Z., Prabakaran, P., Wang, Y., Zhang, M.-Y., Longo, N.S., and Dimitrov, D.S. (2009). Germline-like predecessors of broadly neutralizing antibodies lack measurable binding to HIV-1 envelope glycoproteins: implications for evasion of immune responses and design of vaccine immunogens. Biochem. Biophys. Res. Commun. *390*, 404–409.

Xu, H., Schmidt, A.G., O'Donnell, T., Therkelsen, M.D., Kepler, T.B., Moody, M.A., Haynes, B.F., Liao, H.-X., Harrison, S.C., and Shaw, D.E. (2015). Key mutations stabilize antigen-binding conformation during affinity maturation of a broadly neutralizing influenza antibody lineage. Proteins *83*, 771–780.

Xu, R., McBride, R., Nycholat, C.M., Paulson, J.C., and Wilson, I.A. (2011). Structural Characterization of the Hemagglutinin Receptor Specificity from the 2009 H1N1 Influenza Pandemic. J. Virol. *86*, 982–990.

Xu, R., Ekiert, D.C., Krause, J.C., Hai, R., Crowe, J.E., and Wilson, I.A. (2010). Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science *328*, 357–360.

Xu, R., Krause, J.C., McBride, R., Paulson, J.C., Crowe, J.E., and Wilson, I.A. (2013). A recurring motif for antibody recognition of the receptor- binding site of influenza hemagglutinin. Nat. Struct. Mol. Biol. *20*, 363–370.

Xu, R., Zhu, X., McBride, R., Nycholat, C.M., Yu, W., Paulson, J.C., and Wilson, I.A. (2012). Functional balance of the hemagglutinin and neuraminidase activities accompanies the emergence of the 2009 H1N1 influenza pandemic. J. Virol. *86*, 9221–9232.

Yang, C.-Y., Renfrew, P.D., Olsen, A.J., Zhang, M., Yuvienco, C., Bonneau, R., and Montclare, J.K. (2014a). Improved stability and half-life of fluorinated phosphotriesterase using rosetta. Chembiochem *15*, 1761–1764.

Yang, D., Singh, A., Wu, H., and Kroe-Barrett, R. (2016). Comparison of biosensor platforms in the evaluation of high affinity antibody-antigen binding kinetics. Analytical Biochemistry *508*, 78–96.

Yang, H., Carney, P., and Stevens, J. (2010). Structure and Receptor binding properties of a pandemic H1N1 virus hemagglutinin. PLoS Curr *2*, RRN1152.

Yang, H., Chang, J.C., Guo, Z., Carney, P.J., Shore, D.A., Donis, R.O., Cox, N.J., Villanueva, J.M., Klimov, A.I., and Stevens, J. (2014b). Structural stability of influenza A(H1N1)pdm09 virus hemagglutinins. J. Virol. *88*, 4828–4838.

Yanover, C., Fromer, M., and Shifman, J.M. (2007). Dead-end elimination for multistate protein design. J. Comput. Chem. *28*, 2122–2129.

Yassine, H.M., Boyington, J.C., McTamney, P.M., Wei, C.-J., Kanekiyo, M., Kong, W.-P., Gallagher, J.R., Wang, L., Zhang, Y., Joyce, M.G., et al. (2015). Hemagglutinin-stem nanoparticles generate heterosubtypic influenza protection. Nat. Med. *21*, 1065–1070.

Yin, J., Beuscher, A.E., Andryski, S.E., Stevens, R.C., and Schultz, P.G. (2003). Structural plasticity and the evolution of antibody affinity and specificity. J. Mol. Biol. *330*, 651–656.

Zhang, Y., Aevermann, B.D., Anderson, T.K., Burke, D.F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C.N., Lee, A.J., et al. (2017). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. Nucleic Acids Res. *45*, D466–D474.

Zhou, T., Georgiev, I., Wu, X., Yang, Z.-Y., Dai, K., Finzi, A., Kwon, Y.D., Scheid, J.F., Shi, W., Xu, L., et al. (2010). Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. Science *329*, 811–817.

Zhou, T., Zhu, J., Wu, X., Moquin, S., Zhang, B., Acharya, P., Georgiev, I.S., Altae-Tran, H.R., Chuang, G.-Y., Joyce, M.G., et al. (2013). Multidonor Analysis Reveals Structural Elements, Genetic Determinants, and Maturation Pathway for HIV-1 Neutralization by VRC01-Class Antibodies. Immunity *39*, 245–258.

Zwick, M.B., Labrijn, A.F., Wang, M., Spenlehauer, C., Saphire, E.O., Binley, J.M., Moore, J.P., Stiegler, G., Katinger, H., Burton, D.R., et al. (2001). Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. J. Virol. *75*, 10892–10905.

APPENDIX A.

Protocol Capture for Chapter II

Introduction

The following is a detailed description of how to run and analyze the results from restrained convergence in multi-specificity design (RECON). RECON is used for multi-specificity design, to minimize the energy of multiple protein complexes simultaneously. This is useful in many different contexts, for example to design a single sequence that can bind multiple proteins with high affinity, or a sequence that stabilizes a protein in multiple conformations. In the benchmark cases we selected, the goal was to optimize the sequence of proteins with multiple binding partners to increase affinity for all binding partners. Below is a description of how to perform multi-specificity design of a protein, FYN kinase, which has been crystallized with two binding partners, and optimize the sequence to have low energy in complex with its two partners.

RECON is run completely within ROSETTASCRIPTS, as a combination of movers written specifically for the purpose of multistate design. This offers the benefit of making all other movers available within ROSETTASCRIPTS compatible with a multi-specificity protocol, i.e. backbone minimization, rigid body docking, atom pair constraints, etc. **All ROSETTA commands were run with version 8641cc1735a37dff08c3f1857bbe3035908f7f04.** All analysis scripts are available for download at https://github.com/sevya/msd_analysis_scripts. Note: scripts provided in the analysis directory are dependent upon each other, and when moved from this directory may not function properly.

PDB Preparation

First, the PDB structures were downloaded from the RCSB and manually inspected to remove all but one asymmetric unit. In this case, the PDB IDs of the FYN kinase structure of interest are 1AVZ and 1M27. Structures can be processed manually or with the clean_pdbs.py script, located in /path/to/Rosetta/tools/protein_tools/scripts/clean_pdb.py. This script will download the specific chains of your structure and remove all non-proteinogenic molecules, which makes the structure compatible with ROSETTA. The syntax for this command is:

```
clean_pdb.py 1avz ABC
clean_pdb.py 1m27 ABC
```

In this case chain C in both 1avz and 1m27 is the FYN kinase that will be designed. However 1m27 has an extra leading valine residue at the N-terminus that is not present in 1avz. To simplify the protocol this residue was removed in PyMol before proceeding - this residue can also be removed using a text editor. Next, the chains in each structure were reordered to put the one protein common to both structures, FYN, as the first chain, chains were renamed to A, B, and C, and each chain was renumbered starting from one. This simplifies the protocol by giving the input structure a common format. The renumbering can be done manually or with the script reorder_pdb_chains.py, which takes as input your desired chain order, desired chain ids, and the input and output PDBs. This script simply moves the order of the chains to the desired order and renames the chains, while also renumbering residues from 1 to N. Note that this does not change the coordinates of any atoms, only the order in the PDB file and the chain identifier. The command for this is:

```
reorder_pdb_chains.py --new_chain_order C,A,B --new_chain_ids
A,B,C 1avz.pdb 1avz_renum.pdb
reorder_pdb_chains.py --new_chain_order C,A,B --new_chain_ids
A,B,C 1m27.pdb 1m27_renum.pdb
```

Next 50 relaxed models were created from each of the two starting PDBs, using the following commands, XML scripts and flags. Of the 50 relaxed models I selected the lowest energy model for the design process. The flags I use control the memory usage when ROSETTA is building side chain rotamers (linmem_ig), the number of extra rotamers to include in the library (ex1/2, use_input_sc), the number of models to make (nstruct), and a designation to include all side chain atoms (fullatom). For more information on ROSETTA relax and available options see https://www.rosettacommons.org/docs/latest/prepare-pdb-for-rosetta-with-relax.html. Below are the commands used to create relaxed models.

```
/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgc
crelease @relax.flags -s 1avz_renum.pdb -parser:protocol
relax.xml
relax.xml:
<ROSETTASCRIPTS>
  <SCOREFXNS>
    </SCOREFXNS>
    <TASKOPERATIONS>
        <InitializeFromCommandline name=ifcl />
        <RestrictToRepacking name=rtr />
    </TASKOPERATIONS>
```

```
    <FILTERS>
    </FILTERS>
    <MOVERS>
        <FastRelax name=relax repeats=8 task_operations=ifcl,rtr
min_type=lbfgs_armijo_nonmonotone/>
    </MOVERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
<Add mover=relax />
    </PROTOCOLS>
</ROSETTASCRIPTS>

relax.flags:
-database /path/to/Rosetta/main/database
-linmem_ig 10
-ex1
-ex2
-in:file:fullatom
-out:file:fullatom
-use_input_sc
-nstruct 50
```

Input files

Once the input structures have been processed, the input files needed for RECON can be generated. First residue files (resfiles) are needed that specify the designable and repackable residues for both

of my complexes. Residues that are designable can be substituted with any other amino acid, whereas ones that are repackable can only be substituted with different rotational isomers (rotamers) of the same amino acid. For more information on resfile syntax and logic see: https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/d1/d97/resfiles.html. In this case all residues on chain A that are at the interface of chain A and chains B+C were chosen for design. Since the two complexes have different binding partners they may have overlapping but not identical interface residues – in this case I selected only interface residues common to both complexes. In addition I want to repack any residues on chains B+C that are at the interface. A residue file is needed for each complex that specifies which residues are to be designed and repacked. The number of designable residues must be the same between copmlexes, but repackable residues can be unique to each complex. The following script and flags will generate these files:

```
generate_interface_files.py --side1 A --side2 BC --design-side 1
--repack --output 1avz 1avz_relaxed.pdb
generate_interface_files.py --side1 A --side2 BC --design-side 1
--repack --output 1m27 1m27_relaxed.pdb
```

This identifies all residues at the interface between chains A and B+C, specifies side 1 as the one with designable residues, and signals to repack any residues at the opposing side of the interface. It also specifies a name for the output file, which will be followed by the extension .resfile. After generating residue files, to ensure that both complexes are designing the same number of residues it's important to manually remove residues on the A chain that are at the interface of one complex but not the other. The resfiles used in the benchmark are shown below for reference:

1avz.resfile:

```
NATRO
start
12 A ALLAA EX 1 EX 2
13 A ALLAA EX 1 EX 2
14 A ALLAA EX 1 EX 2
15 A ALLAA EX 1 EX 2
16 A ALLAA EX 1 EX 2
35 A ALLAA EX 1 EX 2
48 A ALLAA EX 1 EX 2
1 C NATAA EX 1 EX 2
2 C NATAA EX 1 EX 2
3 C NATAA EX 1 EX 2
4 C NATAA EX 1 EX 2
5 C NATAA EX 1 EX 2
6 C NATAA EX 1 EX 2
7 C NATAA EX 1 EX 2
10 C NATAA EX 1 EX 2
12 C NATAA EX 1 EX 2
13 C NATAA EX 1 EX 2
16 C NATAA EX 1 EX 2
17 C NATAA EX 1 EX 2
20 C NATAA EX 1 EX 2
47 C NATAA EX 1 EX 2
48 C NATAA EX 1 EX 2
49 C NATAA EX 1 EX 2
50 C NATAA EX 1 EX 2
```

1m27.resfile:

```
NATRO
start
12 A ALLAA EX 1 EX 2
13 A ALLAA EX 1 EX 2
14 A ALLAA EX 1 EX 2
15 A ALLAA EX 1 EX 2
16 A ALLAA EX 1 EX 2
35 A ALLAA EX 1 EX 2
48 A ALLAA EX 1 EX 2
61 B NATAA EX 1 EX 2
63 B NATAA EX 1 EX 2
75 B NATAA EX 1 EX 2
76 B NATAA EX 1 EX 2
77 B NATAA EX 1 EX 2
78 B NATAA EX 1 EX 2
79 B NATAA EX 1 EX 2
82 B NATAA EX 1 EX 2
83 B NATAA EX 1 EX 2
85 B NATAA EX 1 EX 2
86 B NATAA EX 1 EX 2
```

RECON Script

The following script contains the RECON fixed backbone protocol:

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
```

```
                <tal weights=talaris2013.wts >
                        <Reweight scoretype=res_type_constraint
weight=1.0 />
                </tal>
        </SCOREFXNS>
        <TASKOPERATIONS>
                <InitializeFromCommandline name=ifcl />
                <RestrictToRepacking name=rtr />
        </TASKOPERATIONS>
        <MOVERS>
                <PackRotamersMover name=design scorefxn=tal
task_operations=ifcl />
                <MSDMover name=msd1 design_mover=design
constraint_weight=0.5 resfiles=1avz.resfile,1m27.resfile debug=1
/>
                <MSDMover name=msd2 design_mover=design
constraint_weight=1 resfiles=1avz.resfile,1m27.resfile debug=1
/>
                <MSDMover name=msd3 design_mover=design
constraint_weight=1.5 resfiles=1avz.resfile,1m27.resfile debug=1
/>
                <MSDMover name=msd4 design_mover=design
constraint_weight=2 resfiles=1avz.resfile,1m27.resfile debug=1
/>
                <FindConsensusSequence name=finish scorefxn=tal
resfiles=1avz.resfile,1m27.resfile />
        </MOVERS>
```

```
        <FILTERS>
        </FILTERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
                <Add mover=msd1 />
                <Add mover=msd2 />
                <Add mover=msd3 />
                <Add mover=msd4 />
                <Add mover=finish />
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

In this case the design mover used is a PackRotamersMover, which is given to each MSDMover as a submover. Note that the design mover is never actually called – it is called within the MSDMover. The four MSDMovers also specify a weight for residue constraints, which are ramped throughout the protocol, and a debug flag for extra output messages. The resfiles tag uses the files generated in the previous step to tell the MSDMover which residues should be linked together in multistate design. The resfiles don't need to have designable residues at the same positions (i.e. position 1 on protein 1 can correspond to a position 10 on protein 2), but they must have the same number of total designable residues. **Note: RECON matches resfiles to structures by input order. It is critical that PDBs are specified on the command line in the same order as resfiles in the XML file.** FindConsensusSequence is the greedy selection protocol to ensure that a single multi-specific sequence results from RECON. It checks at each position specified in the resfiles if the two input PDBs have a different amino acid, and if they do it places each of the candidate

amino acids onto all states, packs rotamers and checks the sum of energy across states. Whichever of the candidates results in the lowest energy across all states is accepted as the final identity.

A flags file is also needed to specify ROSETTA options – the following are the flags used in the benchmark:

```
-in:file:fullatom
-out:file:fullatom
-database /path/to/Rosetta/main/database/
-linmem_ig 10
-ex1
-ex2
-nstruct 100
-run:msd_job_dist
-run:msd_randomize
```

The only flags specific to the RECON protocol are the last two. Run:msd_job_dist is needed for the JobDistributor to be able to give multiple input poses to a mover at the same time, which is needed for multi-specificity design. This protocol will fail and throw an error message without this flag. Run:msd_randomize randomizes the order of input poses before applying each mover. This is not completely necessary for multi-specificity design but is recommended, the reason being that there is slightly different behavior depending on the order in which PDBs are input. By randomizing the order before you keep this from biasing your results. More information on the other flags can be found at https://www.rosettacommons.org/docs/wiki/full-options-list.

Running RECON

Now that the setup is complete RECON can be performed with the following command line:

```
/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgc
crelease @msd.flags -s 1avz.pdb 1m27.pdb -parser:protocol
fix_bb.xml -scorefile fix_bb.fasc
```

This will generate 100 fixed backbone designs using RECON. For my backbone minimized designs the same options, input files, and commands were used, with the only difference being my xml:

```
<ROSETTASCRIPTS>
     <SCOREFXNS>
          <tal weights=talaris2013.wts >
               <Reweight scoretype=res_type_constraint
weight=1.0 />
          </tal>
     </SCOREFXNS>
     <TASKOPERATIONS>
          <InitializeFromCommandline name=ifcl />
          <RestrictToRepacking name=rtr />
          <RestrictToInterfaceVector name=rtiv chain1_num=1
chain2_num=2,3 CB_dist_cutoff=10.0 nearby_atom_cutoff=5.5
vector_angle_cutoff=75 vector_dist_cutoff=9.0 />
     </TASKOPERATIONS>
     <MOVERS>
```

```
        <PackRotamersMover name=design scorefxn=tal
task_operations=ifcl />
        <MSDMover name=msd1 design_mover=design
constraint_weight=0.5 resfiles=1avz.resfile,1m27.resfile debug=1
/>
        <MSDMover name=msd2 design_mover=design
constraint_weight=1 resfiles=1avz.resfile,1m27.resfile debug=1
/>
        <MSDMover name=msd3 design_mover=design
constraint_weight=1.5 resfiles=1avz.resfile,1m27.resfile debug=1
/>
        <MSDMover name=msd4 design_mover=design
constraint_weight=2 resfiles=1avz.resfile,1m27.resfile debug=1/>
        <FindConsensusSequence name=finish scorefxn=tal
resfiles=1avz.resfile,1m27.resfile />
        <TaskAwareMinMover name=min tolerance=0.001
task_operations=rtiv type=lbfgs_armijo_nonmonotone chi=1 bb=1
jump=1 scorefxn=talaris2013 />
        <FastRelax name=relax scorefxn=talaris2013
task_operations=ifcl,rtr,rtiv repeats=1 />
    </MOVERS>
    <FILTERS>
    </FILTERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
        <Add mover=msd1 />
```

```
            <Add mover=min />
            <Add mover=msd2 />
            <Add mover=min />
            <Add mover=msd3 />
            <Add mover=min />
            <Add mover=msd4 />
            <Add mover=min />
            <Add mover=finish />
            <Add mover=relax />
      </PROTOCOLS>
</ROSETTASCRIPTS>
```

In addition structures generated using backrub motions were generated using the same options, input files, and commands, with the following XML:

```
<ROSETTASCRIPTS>
      <SCOREFXNS>
            <tal weights=talaris2013.wts >
                  <Reweight scoretype=res_type_constraint
weight=1.0 />
            </tal>
      </SCOREFXNS>
      <TASKOPERATIONS>
            <InitializeFromCommandline name=ifcl />
      </TASKOPERATIONS>
      <MOVERS>
```

```
          <PackRotamersMover name=design scorefxn=tal
task_operations=ifcl />
          <MSDMover name=msd1 design_mover=design
constraint_weight=0.5 resfiles=%%resfiles%% debug=1 />
          <MSDMover name=msd2 design_mover=design
constraint_weight=1 resfiles=%%resfiles%% debug=1 />
          <MSDMover name=msd3 design_mover=design
constraint_weight=1.5 resfiles=%%resfiles%% debug=1 />
          <MSDMover name=msd4 design_mover=design
constraint_weight=2 resfiles=%%resfiles%% debug=1/>
          <FindConsensusSequence name=finish scorefxn=tal
resfiles=%%resfiles%% />
          <BackrubDD name=brub moves=5000 >
               <span begin=1 end=57 />
          </BackrubDD>
     </MOVERS>
     <FILTERS>
     </FILTERS>
     <APPLY_TO_POSE>
     </APPLY_TO_POSE>
     <PROTOCOLS>
          <Add mover=msd1 />
          <Add mover=brub />
          <Add mover=msd2 />
          <Add mover=brub />
          <Add mover=msd3 />
          <Add mover=brub />
```

```
            <Add mover=msd4 />
            <Add mover=brub />
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

MPI_MSD File Preparation

To run MPI_MSD the same designable and repackable residues were used, and files were reformatted for this application. Full documentation on the MPI_MSD application is available at https://www.rosettacommons.org/docs/latest/mpi-msd.html. Briefly, the necessary input files are described below. An entity resfile is needed that specifies the residues to be designed (FYN.entres), a correspondence file that maps designable residues to an index (FYN.corr), and secondary resfiles that specify which additional residues are to be repacked (1avz.2res, 1m27.2res). The residues included in these files are derived from the interface residues I used in RECON. In addition a fitness file is needed that specifies the fitness function used (fitness.daf), and state files for each input pdb (1avz.state, 1m27.state).

The contents of the input files used in the benchmark are shown below:

```
FYN.entres:
7
ALLAA EX 1 EX 2
start
1 A ALLAA EX 1 EX 2
2 A ALLAA EX 1 EX 2
3 A ALLAA EX 1 EX 2
4 A ALLAA EX 1 EX 2
```

5 A ALLAA EX 1 EX 2

6 A ALLAA EX 1 EX 2

7 A ALLAA EX 1 EX 2

FYN.corr:

1 12 A

2 13 A

3 14 A

4 15 A

5 16 A

6 35 A

7 48 A

1avz.2res:

NATRO EX 1 EX 2

start

1 C NATAA EX 1 EX 2

2 C NATAA EX 1 EX 2

3 C NATAA EX 1 EX 2

4 C NATAA EX 1 EX 2

5 C NATAA EX 1 EX 2

6 C NATAA EX 1 EX 2

7 C NATAA EX 1 EX 2

10 C NATAA EX 1 EX 2

12 C NATAA EX 1 EX 2

13 C NATAA EX 1 EX 2

16 C NATAA EX 1 EX 2

```
17 C NATAA EX 1 EX 2
20 C NATAA EX 1 EX 2
47 C NATAA EX 1 EX 2
48 C NATAA EX 1 EX 2
49 C NATAA EX 1 EX 2
50 C NATAA EX 1 EX 2

1m27.2res:
NATRO EX 1 EX 2
start
61 B NATAA EX 1 EX 2
63 B NATAA EX 1 EX 2
75 B NATAA EX 1 EX 2
76 B NATAA EX 1 EX 2
77 B NATAA EX 1 EX 2
78 B NATAA EX 1 EX 2
79 B NATAA EX 1 EX 2
82 B NATAA EX 1 EX 2
83 B NATAA EX 1 EX 2
85 B NATAA EX 1 EX 2
86 B NATAA EX 1 EX 2

fitness.daf:
STATE_VECTOR avz 1avz.state
STATE_VECTOR m27 1m27.state
FITNESS vmin(avz) + vmin(m27)
```

```
1avz.state:
1avz_relax_2_0010.pdb FYN_min.corr 1avz.2res

1m27.state:
1m27_relax_2_0003.pdb FYN_min.corr 1m27.2res
```

Running MPI_MSD

MPI_MSD can be run with the following command:

```
mpiexec -n 12
/path/to/Rosetta/main/source/bin/mpi_msd.mpi.linuxgccrelease -
database /path/to/Rosetta/main/database/ -entity_resfile
FYN.entres -fitness_file fitness.daf -ms::pop_size 100 -
ms::generations 105 -ms::numresults 100 -no_his_his_pairE -
ms::fraction_by_recombination .02 -msd::double_lazy_ig_mem_limit
100 -ex1 -ex2
```

This runs the application on 12 processors and generates 100 output files.

Design analysis

To perform design analysis structures are first sorted by the fitness of all designs, which is the sum of energy of my input proteins. I analyzed the top ten designs for each of these three methods for fitness, sequence recovery, and similarity to evolutionary sequence profile. After identifying the top ten designs I used the Weblogo server to generate sequence logos, and the deep_analysis script as a wrapper to calculate amino acid frequencies at each position and make my sequence logo.

Deep analysis takes as input a resfile to identify which residues should be compared - however, note that a separate resfile should be made for only designable residues for this purpose (FYN_analysis.resfile), otherwise it will output a sequence logo for all designable and repackable residues. The contents of this resfile are shown below:

```
FYN_analysis.resfile:
start
12 A ALLAA EX 1 EX 2
13 A ALLAA EX 1 EX 2
14 A ALLAA EX 1 EX 2
15 A ALLAA EX 1 EX 2
16 A ALLAA EX 1 EX 2
35 A ALLAA EX 1 EX 2
48 A ALLAA EX 1 EX 2
```

Note: deep analysis does not link residues between complexes like RECON. It's most useful to analyze each input complex separately. However, since the result of my design run will be two complexes with exactly the same sequence at all designable positions, it's only strictly necessary to analyze sequences from one of the complexes. An example command for this script is the following:

```
deep_analysis --prefix 1avz_fixbb_ --native 1avz_renum.pdb --
stack_width 30 --seq --format png --labels sequence_numbers --
res FYN_analysis.resfile -s d *pdb --path /path/to/weblogo
```

This will output a sequence logo, as well as a .tab file that contains all amino acid frequencies at all positions. From this file you can convert amino acid frequencies into a bitscore (which is equal to $p_i \times \log_2(20 \times p_i)$) and calculate the native sequence recovery (defined as the bitscore of the native amino acid divided by total bitscore at a position).

Evolutionary Sequence Profiles

To generate an evolutionary sequence profile for each protein PSIBlast was run with the following command:

```
psiblast -query fyn.fasta -db non_redundant_database.db -
num_iterations 2 -out out.txt -out_pssm fyn_pssm.txt -
out_ascii_pssm fyn_ascii_pssm.txt
```

The ASCII PSSM contains amino acid frequencies for all positions in the FASTA file. I filtered by 1) residues that were specified in my resfile, and 2) residues that were mutated in the top ten models produced by any of the three design protocols. This evolutionary PSSM was then compared to the design PSSM for each design method. To do this I calculated a squared difference matrix between the two PSSMs, and summed the difference over all amino acids at a given position. At each position, I subtracted this value from 2 and normalized by a factor of 2 to yield a percent similarity score. I then averaged the percent similarity over all positions to generate an overall percent similarity score.

APPENDIX B.

Protocol Capture for Chapter III

Introduction

The following is a protocol capture of how to run the parallelized RECON multistate design method. It will review an example of designing the influenza antibody C05 against a panel of influenza antigens and evaluating the models. For simplicity's sake we have condensed the protocol to a set of five antigens rather than the 13 discussed in the chapter.

All Rosetta commands for this publication were run with version 3e41de71be009712db5ba0f3b0cd1080a1603181, from March 2016. Note that all analysis scripts will only function properly if they are in the correct directory as provided.

All materials from this protocol capture can be downloaded from https://github.com/sevya/parallelized_RECON_protocol_capture

Dependencies:

Several scripts require the use of Python 2.7 as well as the Biopython package (https://github.com/biopython/biopython.github.io/). We recommend installing all necessary packages before beginning this protocol.

Structure preparation

First, the C05 Fab structure (PDB ID 4fnl) was downloaded from the Protein Data Bank (PDB; www.rcsb.org) and manually processed in PyMol. All waters were removed from the structure and

non-protein residues were also removed, and all chains but H and L were deleted. The antibody was processed to remove the constant domain of the Fab fragment – this was done to reduce the total size of the system and simulation time. In this case, residues 114 – 214 on chain H and residues 108 – 213 on chain L were removed. The complex was then saved to a PDB file. Next the structures of the antigenic proteins of interest were downloaded from the PDB. In this example we will use five H1 structures as a test case – PDB IDs 1rvx, 1ruy, 4hkx, 3ubq, and 4lxv. These five structures were downloaded from the PDB and waters and non-protein residues were removed. In addition for each antigen all chains except for one HA1 monomer were deleted. The HA1 subunit was also truncated to the head domain, based on the start and end points of structure 4hkx. The HA1 subunits were renamed so that the chain ID is A, so that the chain IDs would be uniform between different complexes, using the following command:

```
alter 1rvx, chain='A'
```

This command was repeater for all five HAs. Mock co-complexes of C05 in complex with each of these antigens were created by aligning to the known co-crystal structure of C05 in complex with an H3 antigen from PDB ID 4fp8. After downloading the 4fp8 structure and deleting all but one copy in the asymmetric unit, the C05 Fab structure was aligned to the antibody in the 4fp8 co-complex and the antigens were aligned to the antigen in the 4fp8 co-complex. Alignments were done using the following commands:

```
super 4fnl, 4fp8
super 1rvx, 4fp8
super 1ruy, 4fp8
```

```
super 4hkx, 4fp8
super 3ubq, 4fp8
super 4lxv, 4fp8
```

Next, mock co-complexes were created and saved to use in multistate design. We used the create command in PyMol to create an object for each complex, which then can be save and processed in ROSETTA. An example command is shown below:

```
create C05_1rvx, 4fnl or 1rvx
save C05_1rvx.pdb, C05_1rvx
```

This command was repeated for all antigens and the new complexes were saved. Also save the C05 Fab structure for use later using the command:

```
save C05.pdb, 4fnl
```

These complexes were renumbered using a python script. The following script both renumbers the PDB chains and reorders the chains, so that chains H and L come first in the file. Note that this does not change any of the atomic coordinates in the structure, only reorders them. Run this command for all antigens used in the panel, as well as for the C05 apo structure.

```
python reorder_pdb_chains.py --new_chain_order H,L,A C05_1rvx.pdb
C05_1rvx_renum.pdb
```

```
python reorder_pdb_chains.py --new_chain_order H,L C05.pdb C05_renum.pdb
```

Refinement of input structures

After preparing and saving the complexes to be used for design, we first want to refine them to prevent small clashes from introducing artifacts into the score function. We will use Rosetta relax with restraints to the backbone coordinates to do a subtle refinement. The restraints are placed on all Cα atoms with a standard deviation of 1 Å. We use the following command, XML and options file to run the restrained relax. We must also make the output folder for our models to go in.

```
mkdir C05_templates_relaxed
```

```
/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgcc
release @relax.options -s C05_1rvx_renum.pdb C05_1ruy_renum.pdb
C05_4hkx_renum.pdb C05_3ubq_renum.pdb C05_4lxv_renum.pdb
```

```
relax.options:
-in:file:fullatom
-out:file:fullatom
-database /path/to/Rosetta/main/database
-out:pdb_gz
-ex1
-use_input_sc
-nstruct 1
```

```
-out:path:pdb C05_templates_relaxed/
-parser:protocol relax_cst.xml

relax_cst.xml:
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="talaris_rwt"
weights="talaris2013.wts" >
               <Reweight scoretype="coordinate_constraint"
weight="1.0" />
          </ScoreFunction>
     </SCOREFXNS>
     <FILTERS>
     </FILTERS>
     <TASKOPERATIONS>
          <InitializeFromCommandline name="ifcl"/>
          <RestrictToInterfaceVector name="rtiv"
chain1_num="1,2" chain2_num="3" CB_dist_cutoff="10.0"
nearby_atom_cutoff="5.5" vector_angle_cutoff="75"
vector_dist_cutoff="9.0" />
          <RestrictToRepacking name="rtr"/>
     </TASKOPERATIONS>
     <MOVERS>
          <FastRelax name="relax" task_operations="ifcl,rtr"
scorefxn="talaris_rwt" />
```

```
        <ddG name="ddg" per_residue_ddg="0" repack_unbound="1"
chain_num="3" task_operations="rtiv,ifcl,rtr"
scorefxn="talaris2013" />
        <AtomCoordinateCstMover name="cst" />
        <VirtualRoot name="root" />
    </MOVERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
        <Add mover_name="root" />
        <Add mover_name="cst" />
        <Add mover_name="relax"/>
        <Add mover_name="ddg" />
    </PROTOCOLS>
    <OUTPUT scorefxn="talaris2013" />
</ROSETTASCRIPTS>
```

Multistate design

After refining the structure we are ready to run multistate design. To define the residues which will be included in design, we used a script to calculate residues within a 5 Å cutoff of the antigen. This script calculates residues in contact with the antigen using the 4hkx complex as a template. In addition it will calculate residues on the antigen which are in contact with the antibody and designate these residues for repacking. Note that we use the original mock co-complex to calculate contact residues, not the refined structure. The script is shown below:

```
python define_interface.py --side1 HL --side2 A --design-side 1 --repack --output C05 C05_4hkx_renum.pdb
```

In addition we will generate a repacking only resfile that will identify the same residues as the previous resfile, but will designate all residues as repack only.

```
python define_interface.py --side1 HL --side2 A --design-side 1 --repack --native --output C05_repack C05_4hkx_renum.pdb
```

At this point the files are prepared for multistate design of the five complexes. First we will make an output folder for the models to go into. Then we will run multistate design with the following command.

```
mkdir models/
```

```
mpiexec -n 5 \
/path/to/Rosetta/main/source/bin/rosetta_scripts.mpi.linuxgccrelease @msd.options -l templates.list
```

```
templates.list:
C05_templates_relaxed/C05_1rvx_renum_0001.pdb.gz
C05_templates_relaxed/C05_1ruy_renum_0001.pdb.gz
C05_templates_relaxed/C05_4hkx_renum_0001.pdb.gz
C05_templates_relaxed/C05_3ubq_renum_0001.pdb.gz
C05_templates_relaxed/C05_4lxv_renum_0001.pdb.gz
```

```
msd.options:
-in:file:fullatom
-out:file:fullatom
-database /path/to/Rosetta/main/database/
-out:pdb_gz
-use_input_sc
-ex1
-run:msd_job_dist
-nstruct 50
-mute protocols.simple_moves.GenericMonteCarloMover
-parser:protocol msd_brub.xml
-scorefile msd.fasc
-out:path:pdb models
-nstruct 100
-out:suffix _msd_rlx

msd_brub.xml
<ROSETTASCRIPTS>
     <SCOREFXNS>
          <ScoreFunction name="talaris_rwt"
weights="talaris2013_cst.wts" />
     </SCOREFXNS>
     <TASKOPERATIONS>
          <InitializeFromCommandline name="ifcl" />
          <RestrictToRepacking name="rtr" />
```

```
          <RestrictToInterfaceVector name="rtiv"
chain1_num="1,2" chain2_num="3" CB_dist_cutoff="10.0"
nearby_atom_cutoff="9.0" vector_angle_cutoff="75"
vector_dist_cutoff="9.0" />
          <ReadResfile name="repackable"
filename="C05_repack.resfile" />
     </TASKOPERATIONS>
     <MOVERS>
          <Backrub name="backrub_man" pivot_residues="106-118"
/>
          <GenericMonteCarlo name="backrub"
mover_name="backrub_man" scorefxn_name="talaris2013"
trials="500" temperature="0.8" recover_low="1" />

          <PackRotamersMover name="design"
scorefxn="talaris_rwt" task_operations="ifcl" />

          <MSDMover name="msd1" design_mover="design"
post_mover="backrub" constraint_weight="0.5"
resfiles="C05.resfile" debug="0" />
          <MSDMover name="msd2" design_mover="design"
post_mover="backrub" constraint_weight="1"
resfiles="C05.resfile" debug="0" />
          <MSDMover name="msd3" design_mover="design"
post_mover="backrub" constraint_weight="1.5"
resfiles="C05.resfile" debug="0" />
```

```
        <MSDMover name="msd4" design_mover="design"
post_mover="backrub" constraint_weight="2"
resfiles="C05.resfile" debug="0" />

        <FindConsensusSequence name="finish"
scorefxn="talaris2013" resfiles="C05.resfile" debug="1"
task_operations="ifcl,repackable" repack_one_res="1" />

        <FastRelax name="rlx" task_operations="ifcl,rtr,rtiv"
scorefxn="talaris_rwt" />

        <FavorSequenceProfile name="fnr" pdbname="C05_H.pdb"
weight="0.25" scaling="prob" matrix="IDENTITY" />
        <ClearConstraintsMover name="clear_cst" />

        <InterfaceAnalyzerMover name="ddg"
scorefxn="talaris2013" packstat="0" pack_input="0"
pack_separated="1" fixedchains="H,L" />

        <AtomCoordinateCstMover name="cst" coord_dev="1.0" />
        <VirtualRoot name="root" removable="1" />
        <VirtualRoot name="rmroot" remove="1" />
    </MOVERS>
    <FILTERS>
        <FitnessFilter name="fitness" output_to_scorefile="1"
/>
    </FILTERS>
```

```
    <PROTOCOLS>
        <Add mover="fnr" />
        <Add mover="msd1" />
        <Add mover="msd2" />

        <Add mover="msd3" />
        <Add mover="msd4" />
        <Add mover="finish" />

        <Add mover="root" />
        <Add mover="cst" />
        <Add mover="rlx" />
        <Add mover="rmroot" />

        <Add mover="ddg" />

        <Add filter="fitness" />

    </PROTOCOLS>
    <OUTPUT scorefxn="talaris2013" />
</ROSETTASCRIPTS>
```

This protocol will run in parallel over 5 processors and generate 100 output decoys. The protocol will run multistate design followed by a restrained relax to refine the designed complexes, and a step to calculate the fitness, which is calculated as the sum of energy over all input states.

To analyze the sequence profile of designs we used the WebLogo tool. The following script will create a fasta file with the sequences of all designs, as well as a sequence logo summarizing the results. Note that we only need to analyze the designs from one state, since analyzing the designs from each state would be redundant. For example, the designed sequences from trajectory 1 will be identical in the 1rvx complex, 1ruy complex, etc. Here we will only analyze sequences in the 1rvx complex.

```
design_analysis.py --native C05_renum.pdb --format eps --resfile
C05.resfile --multiproc --units probability
models/C05_1rvx*pdb.gz
```

RosettaCM comparative modeling

In addition, in this chapter we describe design against a panel of modeled HAs generated by RosettaCM comparative modeling. To do this we first downloaded HA sequences from the Influenza Research Database (www.fludb.org). We curated these sequences to identify those that comprised full-length HA sequences, aligned them to the truncated head domain from PDB ID 4hkx and truncated the sequences to the same start and end residue. We also removed redundant sequences to create a unique set of seuqences. We clustered these sequences at 95% to reduce the panel to a more tractable size using the CD-HIT software with the following command:

```
cd-hit -i all_H1_unique_head_unique_clean.fasta -o
95_pct_cluster -c 0.95
```

To illustrate the homology modeling protocol we will use the strain H1 A_Uruguay_23_2009 as an example. We used the 13 H1 antigens from the original panel as template sequences – these can be found in H1_templates.fasta. We used the following python script to generate files for RosettaCM in an automated fashion.

```
python generate_files.py A_Uruguay_23_2009
```

This file will generate Grishin alignments for the top five templates by sequence identity, plus a file called template_identity.txt identifying these templates. You must thread your target sequence over the template structure for each of the top five templates. Use the following command to do so:

```
/path/to/Rosetta/main/source/bin/partial_thread.default.linuxgcc
release -in:file:fasta A_Uruguay_23_2009.fasta -in:file:alignment
A_Uruguay_23_2009_3m6s.grishin -in:file:template_pdb 3m6s.pdb

mv 3m6s.pdb.pdb A_Uruguay_23_2009_on_3m6s.pdb
```

Repeat this command for all five templates. Next we recommend using the Robetta server to generate fragments for use in RosettaCM modeling. After downloading 3mer and 9mer fragment files you are ready to run RosettaCM. Run the following command to generate your models. They will be saved to a silent file format to save space – you can easily extract the top scoring models after homology modeling.

```
/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgc
crelease @rosetta_cm.options -scorefile A_Uruguay_23_2009.fasc -
out:file:silent A_Uruguay_23_2009.silent
```

APPENDIX C.

Protocol Capture for Chapter IV

Introduction

The following is a protocol capture of how to run the BROAD multistate design method. It will go through an example of making models of an antibody-antigen complex of VRC23 in complex with HIV gp120, generating a training set of randomly mutated antibodies, fitting the structural models to a classifier, and using integer linear programming to search in sequence space for a broadly binding antibody.

All ROSETTA commands for this publication were run with version 52d173bb0f823b30c009662efb2eb7e635176fc4, from Jan 2016. Note that all analysis scripts will only function properly if they are in the correct directory as provided.

All materials from this protocol capture can be downloaded from https://github.com/sevya/broad_protocol_capture

Dependencies:

Several scripts require the use of Python 2.7 as well as the Biopython package (https://github.com/biopython/biopython.github.io/). The scikit learn package is also required to fit the classification model (https://github.com/scikit-learn/scikit-learn). Lastly IBM cplex is required for solving the integer linear program (https://www.ibm.com/analytics/data-science/prescriptive-analytics/cplex-optimizer). We recommend installing all necessary packages before beginning this protocol.

PDB preparation

First, PDB structure 4j6r was downloaded from the RCSB and manually processed in PyMol. All waters were removed from the structure and non-protein residues were also removed. The antibody was processed to remove the constant domain of the Fab fragment – this was done to reduce the total size of the system and simulation time. In this case, residues 114 – 214 on chain H and residues 108 – 214 on chain L were removed. The complex was then saved and renumbered using a python script. The following script both renumbers the PDB chains and reorders the chains, so that chains H and L come first in the file. Note that this does not change any of the atomic coordinates in the structure, only reorders them.

```
python reorder_pdb_chains.py --new_chain_order H,L,G 4j6r.pdb
4j6r_renum.pdb
```

To select binding sites for the antibody and antigen, we selected residues within a 4 Å cutoff of the opposing chain using PyMol. We used the following commands to do so:

```
select ab_binding_site, byres( chain H+L within 4 of chain G )
select ag_binding_site, byres( chain G within 4 of chain H+L )
```

After expanding the selection to several neighboring residues to allow for contiguous stretches of residues and exclude single residue stretches, we used the following specifications for binding sites (in PDB numbering, from 1-N):

Antibody site:

45-62 EWMGWIKPERGAVSYAPQ

71-74 RDLY

101-105 RDASW

Antigen site:

160-167 NITNNAKI

222-228 SGGDLEI

276-281 NMWQRA

306-313 TRDGGKDN

324-327 GDMR

Generating input files

Next, we used a python script to generate ROSETTA residue files (resfiles) for each virus that will be modeled. This script uses the provided multiple sequence alignment of the viral panel, titled 180_viruses_plus_4j6r.aln, to make a resfile to create each virus. The usage is shown below:

```
python2.7 make_viruses.py
```

This will create a directory called virus_resfiles, and within it a set of 180 resfiles. It will also create a file called viral_variants.fasta, with the concatenated binding sites of all viruses in the set.

Next, we used a similar script to make randomly mutated antibodies. The default setting is to make 5 random substitutions and 500 total antibody variants, but these can be specified in the script:

```
python2.7 make_antibody_variants.py 5 500
```

Generating a training set with ROSETTA

At this point the files are prepared for ROSETTA modeling of the complexes to generate a training set for SVM training and integer linear programming optimization. The input ROSETTA XML file, options file, and command are provided in the protocol capture folder. The following command will generate 50 models for one antibody-antigen pair:

```
/path_to_rosetta/Rosetta/main/source/bin/rosetta_scripts.default
.linuxgccrelease @relax_training.options -s 4j6r_renum.pdb -
parser:protocol relax_training.xml -out:suffix _$abres"_"$virres
-parser:script_vars abres=$abres -parser:script_vars
virres=$virres
```

where abres has been set to an antibody variant resfile (e.g. ab-001) and virres has been set to a virus resfile (e.g. CAAN-A2).

Training the Model

Training involves a) classification to predict binding and b) regression to predict stability scores. To perform training, we attach the following data: the set of the 30 virus sequences is in the file v_30.txt, the training set antibody and virus pairs and the corresponding scores are in train_set_ab_30.txt, train_set_v_30.txt, train_set_scores_binding_30.txt and train_set_scores_stability_30.txt respectively.

The following program creates feature sets from the training antibody and virus files for a given training size and then trains the machine learning models to predict binding and stability.

```
python train_models.py trainsize
```

The program outputs the coefficient and intercept terms for each machine learning model (classification and regression) and saves them to file. The variable 'trainsize' denotes the number of virus sequences chosen for training (on the corresponding antibody-virus pairwise datapoints). To execute the above program, scikit-learn needs to be installed on the system.

The following command learns prediction models on all the data:

```
python train_models.py 30
```

The above program learns a linear model in the default setting. To learn a non-linear model, the function scikit_programs_classification can be modified to choose the rbf kernel parameter. The parameters are set to the values mentioned in the chapter. However, a 10-fold cross validation can be performed to optimize the parameter settings for a specific dataset.

The subset of 20 sequences is in v_20.txt. The training set antibody and virus pairs and the corresponding scores are in train_set_ab_20.txt, train_set_v_20.txt, train_set_scores_binding_20.txt and train_set_scores_stability_20.txt respectively. The following command learns prediction models on the subset of the data:

```
python train_models.py 20
```

The features used for training are saved in classification_features_trainsize.txt and regression_features_trainsize.txt. The parameters are saved in the format coefficients_trainsize_model.txt and intercept_trainsize_model.txt. Model maps to 1 for classification and 2 for regression.

Integer Linear Program

The following program reads the saved coefficients and intercepts, writes the integer linear program using cplex, solves it and writes the optimized antibody sequence to file. It needs IBM cplex to be installed on the system.

```
python solve_ILP.py trainsize
```

This writes the optimized antibody in BM_trainsize.txt (breadth maximized).

Evaluating Optimized Antibodies

Finally, the optimized antibodies can be evaluated using the saved model trained on all data. The following program computes the predicted breadth and the predicted stability score of the optimized antibody, and writes these numbers to file as predicted_breadth.txt and predicted_stability.txt.

```
python evaluate_optimized_antibody.py trainsize
```

Running RECON multistate design

To run RECON multistate design, you must first create an antibody-antigen complex for all the viral variants in the panel, using the crystal structure of VRC23 in PDB ID 4j6r as a template. Before running multistate design, we first threaded over the sequence of each viral variant and refined the complexes using a relax protocol with constraints to the starting backbone coordinates, to prevent too much movement of the protein backbone. We generated one relaxed model for each co-complex to use for design. Using the resfiles generated earlier, we use the following command to create the co-complexes:

```
/path_to_rosetta/Rosetta/main/source/bin/rosetta_scripts.default
.linuxgccrelease @make_templates.options -s 4j6r_renum.pdb -
parser:protocol make_templates.xml -out:suffix _$virres -
parser:script_vars virres=$virres
```

where virres has been set to a virus resfile (e.g. CAAN-A2).

After generating the refined co-complexes we then performed RECON multistate design. The following command was used to run multistate design:

```
mpiexec -n 180
/path_to_rosetta/Rosetta/main/source/bin/rosetta_scripts.mpi.lin
uxgccrelease @msd.options -parser:protocol msd_brub.xml
```

The multistate design was run on a computing cluster with 180 processors allocated to the job. The residues allowed to design were the same residues in the paratope binding site as described previously, and the residues in the epitope binding site were allowed to repack. Backrub movements of the residues in the binding site of both the antibody and virus were performed in between rounds of design.

After generating multistate design models the top ten designs by overall fitness were selected and used to move forward to re-modeling in ROSETTA to enable a direct comparison with the BROAD generated sequences. Overall fitness is defined as the sum of the ROSETTA energy of all complexes included in multistate design. The following command was used to identify the top ten models in overall fitness, and to make a sequence logo for the top ten models:

```
grep CAAN msd.fasc | sort -nk19 | awk '{print $NF.pdb.gz}' |
head -10 > msd_top10.list
python2.7 design_analysis.py --prefix msd_top10.list --res
4j6r.resfile --native 4j6r_renum.pdb `cat msd_top10.list`
```

Evaluating breadth from ROSETTA models

After determining the best ten sequences predicted by both BROAD and multistate design, we subjected them to a more thorough modeling protocol to see if the sequences retain the predicted increases in breadth.

```
/path_to_rosetta/Rosetta/main/source/bin/rosetta_scripts.default
.linuxgccrelease @relax_test.options -s 4j6r_renum.pdb –
parser:protocol relax_test.xml -out:suffix _$abres"_"$virres -
parser:script_vars abres=$abres -parser:script_vars
virres=$virres
```

We next calculated the predicted breadth over the entire viral panel to determine which had greater predicted breadth. To do this we used the output of the testing set relaxation and measured the binding energy and score of the lowest scoring model for each antibody-antigen pair. We used the following script to do this analysis:

```
python2.7 compile_results.py relax_test.fasc
```

This script will calculate breadth over the whole panel for the native antibody and designed antibodies, and will output scatter plots showing the binding energy for each gp120 viral variant, before and after design, to see if design improved binding energy for this viral protein. A similar plot is also output showing the change in score instead of binding energy.

APPENDIX D.

Protocol Capture for Chapter V

Introduction

The following is a protocol capture of how to run the P3SM modeling protocol discussed in Chapter V of this thesis. The sequences used in this study as well as all scripts needed for analysis are deposited along with this thesis.

All Rosetta commands for this publication were run with ROSETTA 3.8, version 816aebf79eac0fd4f103e61e017905d8d234fa4d, from May 2017. Note that all analysis scripts will only function properly if they are in the correct directory as provided.

Dependencies:

Several scripts require the use of Python 2.7 as well as the Biopython package (https://github.com/biopython/biopython.github.io/). We recommend installing all necessary packages before beginning this protocol.

Structure preparation

First, the co-crystal structures of both Z13e1 and 641 I-9 (PDB IDs 3fn0 and 4yk4, respectively) were downloaded from the Protein Data Bank (PDB; www.rcsb.org) and manually processed in PyMol. All waters were removed from the structure and non-protein residues were also removed. For structure 4yk4 all chains but A, B and C were deleted. The chains were changed to A, H, and L, using the following command in PyMol:

```
alter 4yk4 and chain B, chain='L'
alter 4yk4 and chain C, chain='H'
```

For structure 3fn0 the antigen chain was renamed to chain A, with the following command:

```
alter 3fn0 and chain P, chain='H'
```

The antibody was then processed to remove the constant domain of the Fab fragment – this was done to reduce the total size of the system and simulation time. In the case of 4yk4, residues 126 – 225 on chain H and residues 108 – 211 on chain L were removed. In the case of 3fn0, residues 114 – 225 on chain H and residues 108 – 211 on chain L were removed. The complex was then saved to a PDB file. Next a Python script was used to reorder and renumber the antibody chains to simplify the modeling protocol. The reordering script was run as below:

```
python reorder_pdb_chains.py -new_chain_order H,L,A 4yk4.pdb 4yk4_renum.pdb
python reorder_pdb_chains.py -new_chain_order H,L,A 3fn0.pdb 3fn0_renum.pdb
```

Generating training set for P3SM

After preparing and saving the complexes to be used for modeling, we next need to thread the 500 random sequences over the CDRH3 loops and model them to in order to train the P3SM. I used the following script to randomly select 500 sequences and generate resfiles for threading and modeling:

```
python Scripts/make_resfiles.py Run115+RID_VH459_19_clean.csv
```

We next need to thread each of the 500 sequences over both the Z13e1 and 641 I-9 complex and refine using ROSETTA relax. Also included in the input files are resfiles encoding the wild-type sequence of both of these templates.

Now that the resfiles are prepared we will run the threading step. Below are the input files and XML scripts to run the protocol, and a ROSETTA command to thread the sequences over both complexes and generate relaxed models:

thread_relax.xml:

```
<ROSETTASCRIPTS>
     <SCOREFXNS>
          <ScoreFunction name="talaris2014_cst"
weights="talaris2014_cst.wts" />
     </SCOREFXNS>
     <TASKOPERATIONS>
          <InitializeFromCommandline name="ifcl" />
          <RestrictToRepacking name="rtr" />
          <ReadResfile name="mutate"
filename="resfiles/%%resfile%%.resfile" />
     </TASKOPERATIONS>
     <MOVERS>
          <PackRotamersMover name="thread"
scorefxn="talaris2014_cst" task_operations="ifcl,mutate" />
```

```
          <FastRelax name="rlx" task_operations="ifcl,rtr"
scorefxn="talaris2014_cst" />
          <AtomCoordinateCstMover name="cst" coord_dev="1.0" />
          <VirtualRoot name="root" removable="1" />
          <VirtualRoot name="rmroot" remove="1" />
          <InterfaceAnalyzerMover name="ddg"
scorefxn="talaris2014" packstat="0" pack_input="0"
pack_separated="1" fixedchains="H,L" />
     </MOVERS>
     <FILTERS>
     </FILTERS>
     <PROTOCOLS>
          <Add mover="thread" />

          <Add mover="root" />
          <Add mover="cst" />
          <Add mover="rlx" />
          <Add mover="rmroot" />
          <Add mover="ddg" />

     </PROTOCOLS>
     <OUTPUT scorefxn="talaris2014" />
</ROSETTASCRIPTS>

thread_relax.flags:
-in:file:fullatom
-out:file:fullatom
```

```
-database /path/to/Rosetta/main/database
-out:pdb_gz
-ex1
-use_input_sc
-nstruct 10
-out:path:pdb models/
-parser:protocol thread_relax.xml
-scorefile thread_relax.fasc

Commands to run ROSETTA relax:
mkdir models

/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxcl
angrelease @thread_relax.flags -out:suffix _thread_rlx_$resfile
-parser:script_vars resfile=$resfile -s 3fn0_renum.pdb

/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxcl
angrelease @thread_relax.flags -out:suffix _thread_rlx_$resfile
-parser:script_vars resfile=$resfile -s 4yk4_renum.pdb
```

This command will use variable substitution to substitute `$resfile` for the resfile you are currently modeling. You will have to loop through all 502 resfiles (500 randomly selected + two native sequences) and generate models for each of these resfiles separately. I recommend doing this step in parallel on a computing cluster to expedite the simulations.

After the modeling simulations have finished you need to fit the ridge regression model to generate P3SM weights for each amino acid at each position. Below is a Python script that will generate these weights. This script requires a list of amino acids to fit the coefficients and a list of models, examples of which are shown below.

```
cdr3.res:
96 H
97 H
98 H
99 H
100 H
101 H
102 H
103 H
104 H
105 H
106 H
107 H
108 H
109 H
110 H
111 H
112 H
113 H
114 H
```

```
ls models/3fn0* > 3fn0_models.txt
ls models/4yk4* > 4yk4_models.txt

python REU_to_pssm_linear.py --targets 3fn0_models.txt --res
cdr3.res --multi --out 3fn0.pssm

python REU_to_pssm_linear.py --targets 4yk4_models.txt --res
cdr3.res --multi --out 4yk4.pssm
```

Multistate design

The last part of this protocol capture will address how to run RECON multistate design on these modeled sequences. Once I have identified which of the modeled sequences I'm interested in redesigning, I can redesign them to increase affinity for both antigenic targets. I will use the following scripts and flags files to run RECON design:

```
multistate_design.xml:
<ROSETTASCRIPTS>
     <SCOREFXNS>
          <talaris_cst weights="talaris2014_cst.wts" />
     </SCOREFXNS>
     <TASKOPERATIONS>
          <InitializeFromCommandline name="ifcl" />
          <RestrictToRepacking name="rtr" />
          <RestrictToInterfaceVector name="rtiv"
chain1_num="1,2" chain2_num="3" CB_dist_cutoff="10.0"
```

```
nearby_atom_cutoff="9.0" vector_angle_cutoff="75"
vector_dist_cutoff="9.0" />
     </TASKOPERATIONS>
     <MOVERS>
          <Backrub name="backrub_man" pivot_residues="96-114" />
          <GenericMonteCarlo name="backrub"
mover_name="backrub_man" scorefxn_name="talaris2014"
trials="500" temperature="0.8" recover_low="1" />

          <PackRotamersMover name="design"
scorefxn="talaris_cst" task_operations="ifcl" />

          <MSDMover name="msd1" design_mover="design"
post_mover="backrub" constraint_weight="0.5"
resfiles="4yk4.resfile,3fn0.resfile" />
          <MSDMover name="msd2" design_mover="design"
post_mover="backrub" constraint_weight="1"
resfiles="4yk4.resfile,3fn0.resfile" />
          <MSDMover name="msd3" design_mover="design"
post_mover="backrub" constraint_weight="1.5"
resfiles="4yk4.resfile,3fn0.resfile" />
          <MSDMover name="msd4" design_mover="design"
post_mover="backrub" constraint_weight="2"
resfiles="4yk4.resfile,3fn0.resfile" />
```

```
        <FindConsensusSequence name="finish"
scorefxn="talaris2014" resfiles="4yk4.resfile,3fn0.resfile"
debug="0" task_operations="ifcl" repack_one_res="0" />

        <FastRelax name="rlx" task_operations="ifcl,rtr,rtiv"
scorefxn="talaris_cst" />

        <FavorSequenceProfile name="fnr" use_starting="1"
weight="0.25" scaling="prob" matrix="IDENTITY" />
        <ClearConstraintsMover name="clear_cst" />

        <InterfaceAnalyzerMover name="ddg"
scorefxn="talaris2014" packstat="0" pack_input="0"
pack_separated="1" fixedchains="H,L" />

        <AtomCoordinateCstMover name="cst" coord_dev="1.0" />
        <VirtualRoot name="root" removable="1" />
        <VirtualRoot name="rmroot" remove="1" />
    </MOVERS>
    <FILTERS>
        <EnergyPerResidue name="per_res_filter"
scorefxn="talaris2014" resnums="96-114" energy_cutoff="50"/>
        <CalculatorFilter name="cdrh3_filter" equation="19 *
x" >
            <Var name="x" filter="per_res_filter" />
        </CalculatorFilter>
    </FILTERS>
```

```
    <PROTOCOLS>
        <Add mover="fnr" />
        <Add mover="msd1" />
        <Add mover="msd2" />

        <Add mover="msd3" />
        <Add mover="msd4" />
        <Add mover="finish" />

        <Add mover="root" />
        <Add mover="cst" />
        <Add mover="rlx" />
        <Add mover="rmroot" />
        <Add mover="ddg" />

        <Add filter="cdrh3_filter" />
    </PROTOCOLS>
    <OUTPUT scorefxn="talaris2014" />
</ROSETTASCRIPTS>

multistate_design.flags:
-in:file:fullatom
-out:file:fullatom
-database /path/to/Rosetta/main/database/
-use_input_sc
-ex1
-run:msd_job_dist
```

```
-nstruct 10
-mute protocols.simple_moves.GenericMonteCarloMover
-parser:protocol multistate_design.xml
-out:path:pdb models

multistate_design command:
mpiexec -n 2
/path/to/Rosetta/main/source/bin/rosetta_scripts.mpi.linuxgccrel
ease @msd.flags -out:suffix _msd -scorefile
multistate_design.fasc -s $4YK4PDB $3FN0PDB
```

This command also uses variable substitution, where you will need to define variables `4YK4PDB` and `3FN0PDB` with the names of the PDBs you want to design.

APPENDIX E.

Protocol Capture for Chapter VI

Introduction

The following is a protocol capture of designing cyclic peptides based on antibody CDRH3 loops. It will review the design of CDRH3 loop from anti-influenza antibody C05.

All Rosetta commands for this publication were run with version 3e41de71be009712db5ba0f3b0cd1080a1603181, from March 2016.

All materials from this protocol capture can be downloaded from https://github.com/sevya/cyclic_peptide_protocol_capture

Dependencies:

Several scripts require the use of Python 2.7 as well as the Biopython package (https://github.com/biopython/biopython.github.io/). We recommend installing all necessary packages before beginning this protocol. Note that all analysis scripts will only function properly if they are in the correct directory as provided.

Structure preparation

First, the C05 Fab structure (PDB ID 4fnl) was downloaded from the Protein Data Bank (PDB; www.rcsb.org) and manually processed in PyMol. All waters were removed from the structure and non-protein residues were also removed. All residues except for one copy of the CDRH3 from

chain H (residues 93-102, sequence AKHMSMQQVVSAGWERADLVGDAFDV) were deleted. The loop was then saved to a PDB file with the following command:

```
save C05_H3.pdb
```

The CDRH3 peptide was then renumbered using a python script to convert the numbering to start from 1 and ignore insertion codes. The renumbering script was run with the following command:

```
/path/to/Rosetta/tools/protein_tools/scripts/pdb_renumber.py
C05_H3.pdb C05_H3_renum.pdb
```

Next the PeptideStubMover functionality in Rosetta was used to add a cysteine to the N- and C-termini of the CDRH3 peptide. The following command will add these cysteines:

```
/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgc
crelease -parser:protocol add_disulfide.xml -s C05_H3_renum.pdb -
out:prefix disulfide_ -out:no_nstruct_label -extra_res_fa
CYD.params
```

GeneralizedKIC Loop Closure

Now the peptide is ready for loop modeling simulations. We used Generalized Kinematic Closure (GeneralizedKIC) to close the loop and perturb the φ and ψ angles. Full documentation of the GeneralizedKIC protocol can be found at:

https://www.rosettacommons.org/docs/latest/scripting_documentation/RosettaScripts/composite_protocols/generalized_kic/GeneralizedKIC. Briefly, we first declare a bond between the cysteines at residues 1 and 28, and set the degrees of freedom to be used in loop modeling. We set residue 13 at the tip of the CDRH3 loop to be the anchor point of loop modeling, and the remaining residue to be mobile degrees of freedom. We next add a perturber that will modify the closed loop by perturbing the φ and ψ angles of all residues in the loop by a value drawn from Gaussian distribution centered at 15 degrees. Finally we add a single round of ROSETTA relax to add side chains and refine the structure before evaluating the energy. The GeneralizedKIC protocol will generate 20 solutions after loop closure, perturbation, and relaxation, and the lowest energy solution is reported as the final decoy. This entire protocol is repeated to generate 1,000 final output decoys. To create the models output directory and run the GeneralizedKIC protocol use the following command:

```
mkdir models

/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgc
crelease @close_relax.flags -s disulfide_C05_H3_renum.pdb
```

Analysis

We will use a python script to analyze the folded peptide models by calculating ROSETTA score and Cα RMSD for each of the models. To do so we will run the following command:

```
python
/path/to/Rosetta/tools/protein_tools/scripts/score_vs_rmsd.py --
table native_sc_vs_rmsd.tsv --native disulfide_C05_H3_renum.pdb -
-CA --term total models/disulfide*pdb

python plot_score_vs_rmsd.py native_sc_vs_rmsd.tsv
```

This will make a plot of score vs RMSD for all models. In addition it will give a funnel discrimination score that is used to assess how well models converge on the native conformation. This score is derived from Conway *et al.*[1] Overall a lower score (more negative) indicates that the structures are converging well.

Peptide sequence redesign

Next we want to see if we can redesign the peptide for greater stability and convergence on the active conformation. As an example we will take the lowest RMSD structure and run fixed backbone ROSETTADESIGN to optimize the sequence. In a production run we recommend to design more than one structure – in the manuscript we redesigned all peptides under 2 Å.

```
mkdir redesign/
sort -nk3 native_sc_vs_rmsd.tsv | head -2
```

[1] Patrick Conway et al., "Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling.," *Protein Science* 23, no. 1 (January 2014): 47–55, doi:10.1002/pro.2389.

Next copy the lowest energy model into the redesign folder. We provide an example structure for the purpose of this protocol capture. We will run 10 iterations of fixed backbone design and use the lowest scoring design. The resfile we use to guide design will allow all residues to be mutated to anything except for cysteine (ALLAAxC) and will disallow design on the N- and C-termini cysteines.

```
cp models/disulfide_C05_H3_renum_close_relax_0050.pdb redesign/
```

```
/path/to/Rosetta/main/source/bin/fixbb.default.linuxgccrelease -s disulfide_C05_H3_renum_close_relax_0050.pdb -nstruct 10 -out:prefix redesign_ -ex1 -use_input_sc -resfile redesign.resfile
```

```
redesign.resfile:
NATAA
start
2 - 27 A ALLAAxC
```

After making the peptide designs we will analyze the lowest scoring design to see if it improves the folding funnel. We will first use python scripts to convert the sequence from the PDB into a fasta file, then create a resfile from the fasta file to mutate our folding template.

```
/path/to/Rosetta/tools/protein_tools/scripts/get_fasta_from_pdb.py redesign_disulfide_C05_H3_renum_close_relax_0050_0009.pdb A redesign.fasta
```

```
python fasta_to_resfile.py redesign.fasta
```

This will create a resfile called `redesign_disulfide_C05_H3_renum_close_relax_0050_0009.resfile` that we will use to mutate our peptide template. Navigate back to the starting directory, create the template for folding and run the folding simulations:

```
cd ..
```

```
/path/to/Rosetta/main/source/bin/fixbb.default.linuxgccrelease -s disulfide_C05_H3_renum.pdb -out:prefix d1_ -resfile redesign/redesign_disulfide_C05_H3_renum_close_relax_0050_0009.resfile -out:no_nstruct_label -use_input_sc
```

```
/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgccrelease @close_relax.flags -s d1_disulfide_C05_H3_renum.pdb
```

After the design folding decoys are finished we can analyze the score and RMSD and compare to the native sequence:

```
python /path/to/Rosetta/tools/protein_tools/scripts/score_vs_rmsd.py --table d1_sc_vs_rmsd.tsv --native disulfide_C05_H3_renum.pdb --CA --term total models/d1*pdb
```

```
python plot_score_vs_rmsd.py native_sc_vs_rmsd.tsv
d1_sc_vs_rmsd.tsv
```

Binding affinity measurement

In the design process to this point we haven't accounted for the effect that a mutation may have on binding to the antigen. After we have identified a candidate peptide we next want to make sure that the mutations that stabilize the peptide do not interfere with an interaction hotspot. To do this we will model the redesigned sequence in the context of the antibody-antigen interface. We will thread the redesigned sequence over the co-crystal structure, perform a subtle refinement using ROSETTA relax with backbone constraints, and measure the binding energy.

First make a new subdirectory called `ddg_measure` to place all the new files and navigate to this directory. Next we need to download and prepare the structure of the C05 co-crystal structure (PDB ID 4fp8). Download the structure and use PyMol to delete all chains except for A+H. Remove all waters and non-protein residues from the structure. Then remove all residues on the antibody except for the CDRH3 (residues 93-102, sequence AKHMSMQQVVSAGWERADLVGDAFDV). Save this structure as C05_H3_Ag.pdb

Next we will renumber the structure to make sure the numbering is uniform. Repeat the same command from earlier in the protocol to renumber the structure:

```
/path/to/Rosetta/tools/protein_tools/scripts/pdb_renumber.py
C05_H3_Ag.pdb C05_H3_Ag_renum.pdb
```

Once the structure is prepared we will run the relaxation on both the native sequence and redesigned sequence. As an example the sequence of one of the peptides provided in this chapter (d1) is provided, along with a native resfile to measure the binding energy of the wild-type loop. Run the constrained relaxation with these two resfiles to create the models:

```
mkdir models
```

```
/path/to/Rosetta/main/source/bin/rosetta_scripts.default.linuxgc
crelease @relax.flags -out:suffix _rlx_d1 -parser:script_vars
resfile=d1
```

NOTE: the XML format was updated for Rosetta version 3.8, released in February 2017. It is not backwards compatible, so if you are running a version ≥ 3.8, you will need to convert the script with the following command:

```
/path/to/rosetta-
3.8/Rosetta/tools/xsd_xrw/rewrite_rosetta_script.py --input
relax_cst.xml –output relax_cst.xml
```

This will generate ten models for each of the sequences. To analyze the results use the following command to output the score and binding energy in the bound state:

```
sort -nk2 relax.fasc | grep rlx_native | head -1 | awk '{print $2"
"$6}'
sort -nk2 relax.fasc | grep rlx_d1 | head -1 | awk '{print $2"
"$6}'
```