USING PROTEIN FLEXIBILITY TO MODEL VIRAL GLYCOPROTEIN MUTATION TOLERANCES AND SITES OF VULNERABILITY

By

Marion Frances Sauer

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

June 30, 2020

Nashville, Tennessee

Approved:

Hassane Mchaourab, Ph.D.

Jens Meiler, Ph.D.

James Crowe, Jr., M.D.

Ivelin Georgiev, Ph.D.

Kristen Ogden, Ph.D.

Cinque Soto, Ph.D.

ACKNOWLEDGMENTS

I would like to thank the dynamic and interdisciplinary environment the Meiler and Crowe Labs has provided to support me during my graduate career. This first extends to my thesis advisors Dr. Jens Meiler and Dr. James Crowe, Jr., whose academic and financial support granted me the independence and academic freedom to pursue the ideas presented in this thesis. I most likely would have never been granted the same freedom in any other academic environment without their support, especially given my original training as neurologist and cancer biologist.

The Meiler and Crowe Labs are ever-changing in terms of the people that drive the research and the ideas that are presented, but what remains constant is the enthusiasm and academic support each person contributes to the science pursued in either lab. Without the sum total of knowledge presented at lab meetings and the many spaces in between, I would not have absorbed the interdisciplinary training necessary to complete this thesis. I would especially like to thank Dr. Axel W. Fischer, Dr. Alex M. Sevy, Dr. Amanda M. Duran, and Diego del Alamo for their instructive discussions and time spent answering my questions as well as providing insights into my particular project(s). This also extends to members of the Crowe-Meiler Interface, including Dr. Cristina Martina, Dr. Clara Schoeder, Dr. Jessica Finn, Dr. Nina Bozhanova, and Dr. Ben Mueller, whose questions, support, and approachableness helped guide my scientific development. Last, and certainly not least, I would like to thank Heather Darling, Andrea Jordan, and Merissa Mayo, whose exceptional administrative skills made many of my days much less of a headache, especially Heather with her help planning the lab retreat.

I would also like to thank my committee members, Drs. Hassane S. Mchaourab, Ivelin Georgiev, Cinque Soto, and Kristen Ogden in their endeavor to improve my scientific merit, experimental design, and most of all clarity. I truly appreciate the suggestions and questions each of you have had along the way — and the time trying to understand what I intend to say.

My time as a graduate student involved a great deal of learning about myself as well as the work I present in this thesis. In the end, I can say it has been the best period of my life so far, and will ever be grateful for this experience. This could not have been possible without the support of my family and friends during the many life changes that led up to and occured during my graduate career. To my parents, Mavis and Jim, and my siblings, Jenny and John, I will be ever most grateful for the natural curiosity and drive to understand all things complex you helped cultivate. And to the two people that make my life simply happy, Corrinne and Axel, who have made this journey all worthwhile.

TABLE OF CONTENTS

Page

Chapter

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

SUMMARY

The principal focus of this dissertation was to interrogate how structural flexibility influences mutational tolerance and promotes sites of vulnerability within viral glycoproteins. As discussed in chapter I on page 1, bioinformatic and epitope mapping approaches have been successful, but are reactive, in determining the mutation preferences and commonly targeted B-cell epitopes of viral fusion proteins. The primary motivation of this thesis is to describe methods, particularly computational methods, that are proactive in predicting mutational tolerances and B-cell epitopes by assuming that the conformational rearrangements viral fusion glycoproteins undergo are one of the major fitness selection pressures that drive the evolution, especially the conservation, of viral fusion glycoproteins.

Chapter I on page 1 provides a summary of the different class types of fusion proteins and the underlying physical laws that govern the entropy-driven process of viral fusion. Since the body of this work seeks to improve upon current methods to determine mutation preferences and conformational epitopes, chapter I on page 1 also provides an overview of the most commonly used techniques used to either screen mutational preferences or determine the biophysical properties of antibody-antigen interactions, including computational methods such as protein design. Chapter II on page 16 describes a benchmark in which the mutation preferences of eight highly flexible proteins were determined by performing either RECON multi-state design or single-state design on a set of discrete conformations of each protein to estimate the local physicochemical changes needed to assume multiple, low-energy conformations. This chapter focused on two topics — first, the similarity between the designed mutation preferences and natural homologs' sequence diversity, and second, the relationship of sequence conservation and stability to different aspects of protein flexibility. To address the latter topic, a new conformational *dis*similarity metric was introduced, termed contact proximity deviation, which quantified the relative changes in neighboring contacts of each residue experienced within an ensemble. Although this chapter did not specifically address the prediction of viral mutation preferences, the benchmark included two Class I fusion proteins, influenza HA_2 and RSV F protein, and one Class II fusion protein, dengue virus Envelope (DV E) protein to demonstrate that the combined inclusion of at least the pre-fusion and post-fusion protein backbone during design limited mutation tolerances particularly for residues that substantially alter their proximity to neighboring residues. Chapter III on page 42 discusses how the contact proximity deviation metric introduced in chapter II on page 16 and residue relative free energy score might be used as predictors of conformational B-cell epitopes, given that sites of vulnerability often overlap with sites of local conformational change that occur during viral fusion. This chapter focuses primarily on Fab-mediated neutralization and/or protection. Lastly, chapter IV on page 50 discusses the outcome and limitations of the computational methods used in this dissertation to evaluate the dependency of mutational tolerance and epitope locations on the conformational rearrangments necessary to propel attachment and fusion. This chapter also discusses potential future directions that addresses the limitations of the methods discussed in earlier chapters that may be useful in determining how viral fusion glycoprotein flexibility constrains sequence tolerance which in turn gives rise to conserved sites of vulnerability that may be more readily targeted by antibodies.

CHAPTER I

INTRODUCTION

Enveloped viruses — such as influenza, respiratory syncitial virus (RSV), dengue, Ebola, and human immunodeficiency virus (HIV) — commence host infection and subsequent rounds through the attachment and then penetration of a target host cell, known as cell fusion. Cell fusion is mediated by one or more viral surface glycoproteins coating the the mature viral capsid envelop. Although the sequence and three-dimensional structure of the folded viral glycoprotein is incredibly varied among enveloped viruses, all glycoproteins share the common mechanism of coercing two membranes to fuse together. Prevention of viral infection is often mediated by the detection of viral glycoproteins through antibody binding to the surface on the viral glycoprotein, known as an epitope. Recurrent recognition of an epitope through antibody binding requires that the binding surface of the viral antigen remain relatively the same — viral fitness relies on the selection of acquired neutral or beneficial mutations that promote evasion of antibody detection while preserving virion infectivity during successive rounds of infection. In the case of viral glycoprotein fitness, the initiation of infection is dependent on the surface glycoproteins' ability to attach to and fuse a host cell, requiring a substantial conformational change to facilitate the mechanical opening of the host cell. The conservation of physicochemical properties necessary for successful fusion places constraints on viral sequence mutation preferences to conserve aspects of its sequence and/or structure. Thus, the fusion-dependent mutation preferences possibly enhance the likelihood of a common epitope for antibody recognition.

Identification of common epitopes constrained by viral fusion glycoprotein conformational rearrangments requires describing the underlying physical relationship between sequence conservation and protein flexibility. This thesis describes the means to interrogate the dependency of sequence and structural flexibility, which can be applied in the identification of conformation-specific epitopes. In the introduction, a brief review of viral fusion glycoproteins is include to describe the current understanding of the mechanisms of viral fusion and sequence conservation, as well as the experimental methods used to identify common epitopes. Additionally, a more comprehensive review is provided describing the current technologies and underlying principles used to model mutation preferences and conformational flexibility.

I.1. Overview of viral fusion glycoproteins

Viral fusion occurs with the forceful merging of two separate lipid bilayers through the process of hemifusion, where a small region of the outer lipid bilayer is amalgamated to form a single layer while the inner lipid bilayers of each original bilayer remain intact, creating a "hemifusion intermediate" that is widened to form a narrow fusion pore through which the viral genetic components can be delivered into the host cell for infection. In general, viral fusion proteins facilitate the formation of the hemifusion intermediate through assuming a pre-fusion conformation on the exposed virion surface, then following attachment and uptake into the endoplasmic reticulum (ER), some "priming" and/or "triggering" event releases a newly-exposed hydrophobic "fusion peptide" that inserts into the host cell as a transmembrane anchor.[1] As the virion is transported through the ER and trans-Golgi network (TGN), acidification of the ER and TGN lumen promotes the irreversable conformatinonal rearrangment of the fusogenic domain of the viral glycoprotein to its post-fusion conformation. Despite the shared mechanism of forming a hemifusion intermediate, the mechanochemical processes that drive hemifusion employed by viral fusion glycoproteins are currently classified into three distinct groups, which are defined by the priming or triggering event(s) that expose the fusion peptide and the general conformational rearrangments necessary to transition from the pre- to post-fusion state.[2]

I.1.1. Class I viral fusion glycoproteins

The first class of viral fusion glycoproteins contains several well characterized human pathogens, including influenza, HIV-1, and RSV, which use a "hidden dagger" approach to penetrate the host cell membrane. Class I fusion proteins are first expressed on the virion surface as an uncleaved precursor trimer of monomers, such as the influenza HA_0, HIV-1 gp160 Envelop protein (Env), and the RSV F_0 proteins. Priming by proteolytic cleavage can occur either prior or after receptor binding, but is necessary to cleave the precursor into its mature form as a trimer of dimers. In the case of influenza type A HA, binding to sialyc acid precedes the furin cleavage of the HA_0 into the receptor binding domain HA_1 and the fusogenic stem domain HA_2, which are linked by a covalent disulfide bond. Cleavage releases the N-terminus of HA_2, priming the mature form of HA for fusion.[3] For HIV-1 and the related simian immunodeficiency virus (SIV), a furin-like protease must cleave the trimeric gp160 Env into the receptor binding domain gp120 and the fusogenic domain gp41 to increase binding affinity to the CD4 receptor. Upon attachment to CD4, the gp120:gp41 trimer is primed for attachment to the co-receptor CCR5 through a conformational rearrangment of the gp120 domain to expose the co-receptor binding surface.[4,5] Priming of RSV F_0 is a bit more complicated; RSV F contains two furin-like basic cleavage motifs, where the F_0 precursor must be cleaved at both the F1 and F2 cleavage sites for the removal of the p27 fusion peptide.[6,7] However, the timing of cleavage of each site has been disputed, where it was first proposed that F1 cleavage must occur for RSV F egress,[8] although subsequent studies have shown that F1 cleavage does not alter RSV F expression patterns but may confer more rapid cleavage of the F2 site.[9,10] Regardless of exact timing of cleavage, the removal of the fusion peptide primes the RSV F trimer for fusion upon receptor binding. In general, priming by proteolytic cleavage of class I fusion proteins facilitates a local conformational rearrangment within the receptor binding domain, or N-terminal peptide fragment of the fusion precursor, to position the fusion protein for the subsequent triggering step of the fusion process.

The term triggering is distinct from priming, in that some physicochemical interaction allows for the fusogenic region, or C-terminal fragment of the precursor, to undergo a substantial conformational rearrangement that draws for the N- and C-terminal regions of the fusogenic domain into close proximity, thus forcing the viral and host cell membranes into close proximity for the completion of hemifusion. During this triggering event, Class I fusion proteins form what is called a "trimer-of-hairpins" intermediate conformation, whereby the N-terminus of each fusogenic domain rearranges into a three-helix bundle, or hairpin, and inserts as a trimer-of-hairpins into the host membrane via the exposed hydrophobic fusion peptide.[11] For HA, the gradual accumulation of protons from the acidic TGN lumen via negatively charged residues, aspartic acid (Asp) and glutamic acid (Glu) with side chain pKa values of 3.9 and 4.3, respectively, reaches a critical charge to trigger the spontaneous formation of the trimer-of-hairpins hemifusion intermediate.[12,13] The HIV-1 and RSV F triggering event occurs upon (co-)receptor binding and does not require acidification. The fully-cleaved pre-fusion RSV F conformation, once the attachment RSV G protein is bound to its host receptor, spontaneously undergoes a large conformational rearrangment, where the fusogenic F_2 domain rearranges into into an extended six-helix bundle that embeds into the host target cell to spontaneously initiate hemifusion.[14] HIV-1 binding to its co-receptor CCR5 triggers the formation of a six-helix bundle within the C-terminal fusion peptide proximal region of gp41.[15,16]

Completion of hemifusion is a result of the irreversable stabilization of the folded-back trimer-of-hairpins. Stabilization of the HA_2 stem region requires that the C-terminal loop region S5 drag the N-terminal HA_2 B-loop region, which rearranges to form the hairpin helix, into the newly formed fusion pore to force the fusion process to completion.[1] Fusion pore formation by HIV-1 Env is thought to occur following the change in relative angle of the hinge point connecting the membrane proximal external region (MPER) and C-terminal transmembrane region of gp41, which forces completion of hemifusion and the stabilization of the fusion peptide proximal region.[16] The complete formation of a nascent viral fusion pore of Class I fusion proteins is thought to occur when more than one fusion proteins complete hemifusion in close proximity to each other.[17]

I.1.2. Class II viral fusion glycoproteins

The second class of viral fusion glycoproteins includes those of flavivirus and alphaviruses, with the most well structurally characterized system being that of fusogenic DV E protein and its chaperone, dengue virus Matrix (DV M) protein.[1,18–20] Both flaviviruses and alphaviruses have symmetric icosahedral coats consisting of the exposed fusion protein, designated E for flaviviruses and E1 for alphaviruses with an underlying layer consisting of a chaperone protein, prM or pE2, respectively. Unique to Class II fusion proteins, the fusion protein E or E1 forms a continuous polypeptide chain with its chaperone protein, *i.e.* prM and pE2. The uncleaved polypeptide forms a dimer of either E:prM or E1:pE2, with one monomer laying antiparallel to to the other, so that the dimer lays parallel with respect to the viral membrane.[18,21] Like Class I, Class II fusion proteins require both a priming and triggering event to initiate fusion. However, Class II fusion proteins are distinguishable in that there are two cleavage events of the protector chaperone, not the fusion protein itself. The first cleavage event occurs in the TGN prior to egress which separates the polypeptide chain into a heterodimer consisting of the fusion protein and its immature form of the chaperone. In the case of alphaviruses, the cleavage of pE2 results in the formation of an E2:E3 dimer.[21] Although the first cleavage event does not result in a conformational change of the fusion E or E1 protein, it does allow for a substantial rearrangment within the M or E2:E3 protein that results in the transformation of the virion surface from a spiky appearance to a smooth appearance.[22,23] For both flaviviruses and alphaviruses, this cleavage event maintains stable shielding of the fusogenic loop region with E or E1, preventing premature fusion. After exposure to the neutral extracellular environment, however, a second cleavage event occurs resulting in either the dissociation of the flavivirus E:M dimer or the alphavirsus E3 subunit, priming the mature viral capsid for hemifusion and destabilizing the protective shielding of the fusion loop region within E or E1.[20] This renders the virion as infectious.[22] At least for DV E and the alphavirus Semliki Forest virus E1, receptor binding occurs via the C-terminal Domain III, after which the virion is taken up by an endosome.[19,23,24] Upon acification of the endosome, the monomer undergoes a series of conformational changes, so that an E or E1 homotrimer inserts each of its three fusion loop regions into the host cell membrane, resulting in a perpendicular, permanent (re)orientation of the trimer relative to the viral membrane. After complete acidification, the three fusion loop regions within a trimer is brought into close proximity to each of the three transmembrane regions. Finally, completion of hemifusion and fusion pore formation is putatively thought to occur when the formation of a five-trimer cluster forms, which creates a ring of embedded loop regions promote a rapid formation of the fusion pore.[24]

I.1.3. Class III viral fusion glycoproteins

The viral capsid coat of Class III fusion proteins form uniform distribution of protein homotrimers covering the viral outer membrane, with each monomer consisting of nested domains.[25,26] The class III fusion mechanism has been associated with viruses within the Rhabdoviridae and Herpesviridae families, where only the vesticular stomatitis virus (VSV) glycoprotein (G) has been structurally characterized in its pre- and post-fusion states.[27,28] VSV G does not require priming for activity; upon receptor binding and uptake into an acidic endocytotic vesicle environment, the low pH triggers the spontaneous conformational rearrangment of each VSV G monomer so that the fusion domain of VSV G undergoes a 180° rotation.[29] The lack of requirement of priming allows this conformational rearrangment to be completely reversible, unlike a Class I or II fusion.[30,31] The reversible fusion mechanism is achieved solely with the acceptance or loss of a proton by five key acidic Asp and Glu residues.[30] The loss of negative charge within the amino acid side chain results in a destabilization of the prefusion conformation and a stabilization of the post-fusion conformation, whereas the gain of negative charge reverses the stability of each conformation. Even though the pre-fusion conformations of other Class III proteins are not known, given the high conservation of both sequence and structure of other Class III fusion proteins, including Epstein Barr virus Gb, Herpes simplex virus-1 Gb, and *Autographa californica* multiple nucleopolyhedrovirus GP64, it can be inferred that other Class III fusion proteins utilize a similar fusion mechanism as VSV G protein.

I.2. Glycoproteins are metastable facilitators of host cell entry

Like any physical system, the physical properties that define viral glycoproteins and their ability to mechanically open a host cell via fusion comply with the laws of thermodynamics. A protein's structure is best described by the second law of thermodynamics — a physical system will tend to reach its minimum Gibbs free energy as it reaches its thermodynamic equilibrium.[32–34] As discussed in the previous sections, viral glycoproteins are structurally dynamic during the fusion process. Depending on a viral glycoprotein's state, *e.g.* uncleaved, primed, triggered, or after completion of fusion, viral glycoproteins must reach a thermodynamic equilibrium consistent with its physicochemical environment during each phase of the fusion process.

For many Class I and Class II viral glycoproteins, the post-fusion conformation has been shown to be its most stable state regardless of its physiologic pH.[35–38] Based on Anfinsen's dogma, the optimal three-dimensional structure, or native state, of a protein is one that adopts its lowest Gibbs free energy from its one dimensional amino acid sequence.[34] Given that a viral glycoprotein does not alter its sequence as it progresses through a single fusion event, a viral glycoprotein will tend to adopt its post-fusion conformation to achieve its native state. However, viral fusion is dependent on its fusion machinery to interact with the correct host cell receptor(s) and fuse its membrane with the host cell membrane in a step-wise fashion. Therefore, viral glycoprotein function is dependent on its ability to limit its rate of folding into the native, post-fusion conformation by folding into less stable, but functionally relevant conformations, including its pre-fusion and intermediate conformations. Similarly, proteins like serpins[39] or α-lytic protease,[40] are known to form metastable, active conformations that provide kinetic barriers to the formation of the native, inactive conformation, where the formation of metastable conformations has been shown to be an effective general strategy to regulate function.[41–43]

The mechanism for preventing an unfolded viral glycoprotein to directly adopt its native conformation is not well understood and has not been a subject of study for quite some time.[35] However, given that fusion proteins require a triggering, and often also a priming, step to convert its precursor and metastable pre-fusion conformations into its native, post-fusion form, it can be inferred that each conformation presides in a relative Gibbs free energy minimum that is dependent on its physicochemical environment. In other words, the likelihood of a certain conformation being observed within a population of fusion proteins is is dependent upon its thermodynamic equilibrium relative to other energy minima. This can be described by its relative Gibbs free energy, or ΔG, defined as $\Delta G = \Delta H - T\Delta S$, where ΔH is the relative enthalpy, T is the temperature, and ΔS is the relative entropy of the system. Assuming that each local energy minimum represents a system at thermodynamic equilibrium where the thermal and chemical potential are constant, the probability of a given conformation, p_i, is proportional its contribution to the mean energy of the system of all conformations within an ensemble, stated as the Boltzmann relation in equation (I.1). Therefore, if the probability of a certain protein conformation being within an ensemble is very high, the mean free energy of a local energy minimum approximates the free energy of that conformation, *i.e.*, that conformation is highly favored over the other conformations of the same sequence for that particular thermodynamic equilibrium.

$$p_i = \frac{1}{Z} e^{\frac{\epsilon_i}{kT}} = \frac{e^{\frac{\epsilon_i}{kT}}}{\sum_i^N e^{\frac{\epsilon_j}{kT}}} \tag{I.1}$$

where:

Z = partition function
i = state
ϵ = energy of state i
k = Boltzmann's constant
T = temperature of the system
N = total number of states within the system

Relating back to viral fusion protein function, in particular Class I and Class II function, viral glycoproteins first assume an uncleaved precursor conformation within the acidic endocytotic environment of an infected cell. In the case of HA, RSV F, or DV E proteins, all are expressed and folded as a single polypeptide. Folding of the nascent polypeptide chain into a stable, three-dimensional structure is driven by the sum of intramolecular interactions, including hydrogen bond network, ionic, van der Waals, and hydrophobic interactions, with hydrophobic interactions being the predominant driving force for protein folding.[32] This is because, although the conversion of the unfolded to folded state results in $\Delta S_{protein} < 0$, the burial of hydrophobic residues increases the entropy of the solvent, so that $\Delta G_{system} < 0$ and protein folding is a favorable, or spontaneous, event. Indeed, pulse-chase analysis of HA_0, DV E, and Zika E proteins revealed that the precursor monomers are translated, folded, and self-assembled into trimers or dimers, respectively, independent of other structural protein expression or pH,[44,45] suggesting that the formation of precursor fusion proteins is an energetically favorable process.

Egress of viral particles, as discussed earlier, may or may not require priming. However, in each case, priming induces a conformational change by proteolytic cleavage. The release of either the N-terminus of the fusogenic domain from the C-terminus of the receptor domain, as in several Class I fusion proteins, or the release of the chaperone in Class II fusion proteins increases the entropy of the system due to the favorable increase in disorder order of the solvent, so that the conversion of the precursor to pre-fusion conformation has a $\Delta G_{system} < 0$ and is irreversible. The activation energy required for this conformational change is typically achieved through proteolytic cleavage itself — molecular dynamics (MD) simulations have shown that binding to furin, the most common enzyme involved in cleavage, lowers the ΔG with respect to the unbound furin conformation due to ligand binding and histidine (His) protonation and is sufficient to provide the activation energy required for cleavage.[46,47]

Once cleaved, however, the metastable pre-fusion conformation does not spontaneously rearrange into its post-fusion form. As discussed earlier, Class I-III fusion proteins require a triggering event that allows for the exposure of the hydrophobic fusogenic region and eventual conformational rearrangment necessary for hemifusion. This occurs either through acidification of the solvent or by binding to the proper host cell receptor. In the case of pH-dependent triggering, the increase in protonation of the acidic residue carbonyl group ($H_3O^+ + COO^- \longrightarrow H_2O + COOH$) or the τ N of the imidazole side chain in His ($H_3O^+ + N \longrightarrow H_2O + NH^+$) results in $\Delta G_{rxn} < 0$. This can be determined by the relationship between ΔG and the relative population sizes of the amino acid side chain states under acidic conditions, as shown in equation (I.2), where the relative increase in COOH and NH^+ compared to COO^- and N population sizes decreases the reaction quotient Q so that the reaction is favorable.

$$\Delta G_{rxn} = \Delta G^0 + RTlnQ \tag{I.2}$$

where:

ΔG^0 = free energy change of the reaction under standard conditions
R = gas constant
T = temperature of the reaction
Q = $\frac{[COOH][NH^+]}{[2\,H_3O^+][COO^-][N]}$

For the triggering event, whether due to acidification of the endocytotic vescicle lumen or due to receptor binding, $\Delta H \approx 0$ while $\Delta G < 0$, meaning that the triggering event predominantly increases the entropy of the system. However, due to the exposure of the hydrophobic fusogenic region, the solvent experiences a loss of entropy so that the entropy term $-T\Delta S_{solvent} > 0$. Burial of the hydrophobic fusogenic region results in a more favorable ΔG to ensure the conformational rearrangment of the pre-fusion to the post-fusion conformation. For Class I and Class II fusion proteins, burial of the hydrophobic fusogenic region coupled with large structural rearrangments is thought to be the rate limiting step to complete fusion, where the formation of either the folding over of trimer-of-hairpins or the trimerization of Class II fusion proteins overcomes the "hydration-force barrier"[48] as the viral membrane comes into close proximity to the host cell membrane.[1] However, for Class III proteins, the structural rearrangment is reversible due to the reversibility of protonation states, and therefore, the relative free energy of either the pre-fusion or post-fusion conformation in relation to the other is dependent only the acidity of the local environment, and therefore the relative free energy change necessary to drive fusion is solely dependent on the cumulative acidity of the fusion protein.

I.3. Conservation of glycoproteins among quasispecies is not uniform

Viral infection is thought to be initiated by not one, but a population of virions, all of which contain similar but not identical genomes. Collectively, this population of similarly infectious virions is known as a quasispecies.[49] The dynamic distribution of mutations within a quasispecies stems from the mechanisms that replicate its mutable genome, which can consist of either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). All viruses mentioned in this thesis are RNA viruses that lack a DNA polymerase and instead use either an RNA-dependent RNA polymerase or reverse transcriptase that lacks the same proofreading capabilities to maintain the fidelity of the viral genome.[50] RNA viral genomes also tend to be smaller and are replicated at faster rates, which highly correlate with higher mutation rates. It is estimated that RNA viruses accumulate 10^{-2} to 10^{-5} RNA base substitutions per site per year.[51] Even though RNA viruses have a high mutation rate, not all mutations can be translated into functional protein, *i.e.*, not all mutations are *tolerable*. Of the tolerated mutations, most have neutral effects on the fitness, or successful reproductive capabilities, of a viral sequence in relation to the wild-type, or original, sequence. The first attempt to assign relative fitness values to mutations was done using the VSV genome, which found that 90% of mutations resulted in reduced replicative fitness as compared to wild-type, and of that, 40% of mutations were lethal.[52] Subsequent mutational studies on tobacco etch virus, influenza A/WSN/33/H1N1, and poliovirus also found similar rates of fitness selection for tested mutations.[53–55] However, of these studies, only the latter performed whole genome deep-sequencing on serially-passaged poliovirus. By measuring successive selection of mutation fitness, synonymous mutations, or mutations that translate into the wild-type amino acid, were found to be significantly less likely to be lethal than non-synonymous mutations, especially for structural proteins.[55] Therefore, it appears that, despite the high mutational rate of RNA viruses, fitness selection limits the mutational tolerance of transmissable viruses.

The definition of a quasispecies is that they share a genome that is close, if not identical, to a consensus sequence.[50] The apparent high fitness penalty for assuming non-synonymous mutations within a quasispecies can be related to Anfinsen's dogma — if a virus's functional actors, *e.g.* proteins, are dependent on the translated genome sequence for the maintence of its fitness, then the acquisition of new sequence mutations is more likely change viral fitness. However, are the selection pressures on mutation tolerance uniform across the whole genome or not? Based on the whole-genome sequencing of poliovirus, the structural proteins were significantly less likely to have fewer mutation variance as compared to wild-type than the non-structural proteins.[55] However, whole-genome sequence alignments of dengue strains 1 – 4, show that the mean conservation rates of DV E protein within each strain are insignificantly lower than other strain-specific proteins, although certain sequences of DV E are conserved 100% across all strains.[56] This suggests that selection pressure for sequence conservation acts on individual amino acids rather than different viral protein types.

Specifically regarding viral fusion glycoproteins, there have been several studies showing that selection pressure to conserve amino acid identity is location specific. Deep mutational scanning of influenza hemagglutinin from strains A/WSN/1933 (H1N1) and A/Perth/2009 (H3N2) after serial passages showed that selection pressures to conserve amino acid sequence is posisition-specific for sites known to form disulfide bridges, the receptor binding domain, the cleavage site, and the HA_2 B and S5 regions that undergo large local conformational rearrangements during fusion.[57] Additionally, comparision of average mutation rates between influenza H1 and H3 subtypes revealed independent mutation tolerances of the H1 and H3 HA_1 and HA_2 domains as measured by Shannon entropy.[58] For H1, the fusogenic HA_2 domain was much more likely to be conserved than the HA_1 domain, whereas the opposite was true for H3 — however, substantial shifts in mutation preferences were specific to a limited few residues, where the greatest shift in preferences was found in residues with low conservation (Spearman's $p < 1 \times 10^{-18}$) and that are known to be clade specific (Spearman's $p < 1 \times 10^{-12}$). A similar deep mutational scanning approach was used to identify mutation tolerances of HIV Env of two

CXCR4-tropic strains from clade A, including BG505.W6M.C2.T332N and BF520.W14M.C2.[59] Again, sequencing of serially-passages revealed that pressure for sequence conservation was specific to sites known to form disulfide bonds as well as mediate CD4 or CCR5 binding. Moreover, limited mutation tolerances were specific to regions known to undergo conformational rearrangments during fusion, and yet distinct between strains. However, despite the distinct mutation preferences between strains, the solvent accessibility and number of maintained contacts were nearly identical for each conformation between strains.

Given the intensity in terms of time and cost of performing deep mutational scanning, mutation tolerances unfortunately have been determined for only Class I fusion proteins. However, in both studies, it was demonstrated that fitness selection of sequences is not uniform due to the relatively unequal distribution of sequence conservation. However, both studies identified high sequence conservation in regions critical for either the stability of a conformation — as indicated in the high cysteine (Cys) conservation, stability of receptor-virus binding interface — as indicated by the conservation critical for sialic acid or CD4/CCR5 receptor binding, or instability necessary to undergo a conformational change — as indicated in regions known to rearrange during fusion. Therefore, it can be extrapolated from these deep mutational scanning data that some portion of the fitness selection of mutational tolerances is due to the energetic requirements necessary to facilitate fusion constrain specific positions necessary to progress from one local energy minimum to another during fusion.

I.4. Fusion glycoproteins are common antigenic targets of broadly neutralizing antibodies

Influenza and RSV infections alone have been shown to be the underlying cause of over 2 million pneumonia deaths during an annual influenza season within just the United States (October - May).[60] From a more global perspective, as of 2018, 39.7 million people currently test positive for HIV infection (UNAIDS.org, 2019). In order to prevent further disease burden of viral infections, substantial efforts have been put into developing preventative therapies, including the development of vaccines. Many effective current vaccines specifically elicit an immune response against the fusion glycoprotein by means of antibody recognition of a specific sequence or structural motif, known as an epitope, on the fusion protein surface. One of the primary challenges of developing a vaccine, however, is eliciting a population-wide response for a long time period. As mentioned in Section 1.3, fusion glycoproteins rapidly mutate, albeit non-uniformly, which prevents its detection by antibody recognition. Broadly-neutralizing antibodiess (bnAbs) are less susceptible to fusion glycoprotein mutations in part due to recognition of more highly conserved epitopes within a circulating strain or serotype.

For influenza type A and B HA, there are 26 known bnAbs, and of these, 21 target the more highly conserved HA_2 domain.[61] Only one bnAb, CR9114, protects against both type A groups 1 and 2 as well as type B, and binds to an HA_2 epitope.[62] The development of protective immunizations against HIV have proven to be more challenging, since the human antibody response to HIV infection predominantly targets epitopes that are either shielded by diverse posttranslational glycan modifications or are within hypervariable loop regions that are more easily accessable but offer limited long-term protection.[63] However, 10% – 30% of HIV-infected individuals develop bnAbs responses[64–67] against four sites of vulnerability, or common epitopes, two of which are localized at the more highly conserved CD4 binding site and the MPER, which are known to undergo conformational rearrangments during fusion.[18] The only current prophylactic treatment against RSV is palivizumab (SYNAGIS ®, AstraZeneca), which targets a structural epitope on the RSV F protein that is conserved in both the RSV F pre- and postfusion conformations. More recent characterizations of bnAbs against RSV F protein have found that highly potent bnAbs specifically target the cleaved, pre-fusion conformation, which includes D25, 5C4, AM14, and AM22, which target a structural epitope that surface-accessible only in the pre-fusion conformation.[68–70] Another strongly-neutralizing antibody RB1 against both RSV types A and B was found to bind the pre- and post-fusion conformations, but preferentially bound to the pre-fusion conformation, as its disassociation constant (K_D) was much lower, with a K_D of 22 pM for the pre-fusion conformation versus its K_D of 1.35×10^5 pM for the post-fusion conformation. With the stabilization of the precursor, cleaved pre-fusion, and post-fusion conformations of RSV F as nanoparticle vaccines, previously determined bnAbs against RSV F were similarly screened for binding affinity to each RSV F conformation and were found to preferentially bind to the cleaved pre-fusion conformation.[6]

Given their potential large impact on human health, bnAbs against influenza, HIV, and RSV fusion proteins were discussed in the previous paragraph. However, recent emerging threats like Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV), Middle East Respiratory Syndrome-Coronavirus (MERS-CoV), Marburg, and Ebola have tested the response time necessary to identify and characterize effective vaccines. Critical to developing vaccines is understanding which viral target, *e.g.* an antibody epitope, most effectivly blocks infection and/or promotes an immune response necessary to clear infection. As in the case of the previously discussed bnAbs, which are or may be tested as vaccine candidates, a few common themes emerge — first, several antibodies that target fusion glycoproteins exhibit conformational selection for binding, and second, sites of vulnerability often overlap with regions that are both more highly conserved and are conformationally dynamic during fusion. By understanding the underlying mechanisms that drive, or at least promote fitness selection of sequence conservation within fusion glycoproteins, it may be possible to identify conserved sites of "most vulnerability" that are not only sequence dependent, but also conformational dependent. However, to do so, it is necessary to understand the basis of antigen recognition and the current methods used to determine antigen recognition.

I.4.1. The basis of antigen recognition and antigenic memory

The ability of an antibody to bind to wide array of foreign molecules, also known as antigens, is one of the key events that leads to a humoral adaptive immune response against infectious diseases. An antibody is a heterodimer of homodimers and is often cartoonishly represented as a 'Y' shape, with the top two short branches representing two variable fragment (Fv) regions, and the bottom branch representing the constant fragment (Fc) region. Each Fv and Fc region consists of a heavy and light chain, with each chain consisting of two immunoglobulin folds.[50] The N-terminal immunoglobulin fold of each Fv heavy and light chain contains three loops of highly-variable sequence and length, which are termed heavy-chain complementarity determining regions (HCDRs) and light-chain complementarity determining regions (LCDRs) respectively. When the Fv heavy and light chain form a heterodimer, the HCDR and LCDR come in close proximity to form the paratope, or region that forms the binding interface with an antigen. Due to the high degree of both sequence and conformational diversity, paratopes provide the structural basis for molecular recognition of a vast array of potential foreign antigens. This diversity arises due to two processes termed somatic recombination and somatic hypermutation.

Somatic recombination of antibodies describes the combinatorial assembly of three gene segments known as the variable (V), diversity (D), and joining (J) segments, otherwise known as V(D)J combination.[71] The heavy chain is encoded by one of each of the 43 V,[72] 27 D, and 6 J gene segments.[73]

Encoding of the light chain occurs at either the κ or λ gene locus and includes only a constant, V, and J gene — depending on the locus, there are 44κ V genes and 38λ V genes that provide the majority of genetic variation within the light chain.[74,75] V(D)J recombination allows for the creation of a very large germline antibody repertoire, but of the theoretical 3×10^{11} combinations, some are removed due to auto reactivity. With the introduction of junctional diversity by means of random insertion or deletion of nucleotides at imprecisely matched gene segment ends,[50] the genetic diversity of the germline antibody repertoire is massive.

Somatic hypermutation occurs through the activation of a B cell via binding of a B cell receptor to an antigen whereupon the expression of the enzyme, activation-induced cytidine deaminase, is upregulated to promote high mutation rates during B cell replication,[76] resulting in the genetic diversification of produced antibodies. The induced mutations tend to be localized in complementarity determining regions (CDRs),[77] which lead to a change in binding affinity and avidity of the newly synthesized antibodies' paratopes. The repetition of B cell activation and induction of high mutation rates during replication is termed affinity maturation, through which B cells are able to produce antibodies with very high affinity to an exposed antigen.[78]

The purpose of introducing the concepts of somatic recombination and somatic hypermutation is to give an appreciation of the breadth of molecular recognition antibody binding provides. However, the expression of a unique antibody and its ability to bind to a foreign antigen is spatially and temporally dependent — the right B cell must be present for antigen detection. The complex process of affinity maturation promotes a clonal response, whereby the original activated B cell promotes the expansion of memory B cells and antibody-secreting plasma cells to become clonal variants of the activated B cell.[50] This expansion increases the likelihood of antigen detection with a high affinity antibody during an infecton. After the infection is cleared, however, the rate of affinity maturation drops off, as there is no antigen left to promote further clonal expansion of plasma cells. Memory B cells, on the other hand, have been shown to persist in a state of quiescence within human peripheral blood for decades, albeit at very low concentrations,[79–82] and in the case a similar antigen is presented to the memory B cell, the memory B cell is activated to promote a rapid response to the re-emergence of the antigen. B cells can recognize either linear peptide or discontinuous structural epitopes, although it is thought that almost 90% of all B cell epitopes are structural epitopes.[83] Furthermore, B cell activation is dependent upon T_H cell linked recognition,[50] whereby an antigen-specific B cell must effectively concentrate the presented antigen on its cell surface for recognition by armed T_H cells. Therefore, activation of a memory B cell is dependent on antigen presentation that not only retains the same peptide sequence, but also conformation of the antigen that originally promoted its expansion, so that it can efficiently recognize and present a single antigen during linked recognition. Since the generation of vaccines is dependent on the maintenance of antigenic memory, including memory B cells, identification of conserved structural epitopes that promote memory B cell activation could be used as a strategy to design better vaccine candidates.

I.4.2. Methods to map conserved epitopes for vaccine design

Epitope mapping techniques provide structural insights into antibody-antigen interactions and possible antibody-mediated mechanisms of neutralization and/or protection. To accurately characterize an antibody-antigen interaction, epitope mapping must overcome several experimental challenges. First, since the majority of antibody-antigen interactions are conformation dependent, a good epitope mapping technique must be able to provide structural information of either a linear or discontinuous epitope. The epitope binding surface area is usually confined to a surface area containing roughly 20 contact residues, with typically only two to five amino acid "hotspots" that contribute to the majority of the energetic contributions necessary for binding.[84–86] Second, epitope targets like viral glycoproteins are conformationally complex in that they typically include multiple domains and/or oligomers as well as posttranslational modifications like glycans. To accurately map out all possible epitopes, the technique used should be able to determine these conformational complexities.

Structure determination methods including X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (EM) provide structurally detailed information of antibody-antigen interactions. The most commonly used and sometimes deemed most "accurate" approach is using X-ray crystallography, which provides atomic-resolution of the amino acid side chain configurations of paratope-epitope interactions. However, as its name implies, X-ray crystallography requires that the antibody-antigen complex is stable enough to form crystals. As mentioned in previously in Section 2 and the opening of Section 3, a substantial portion of bnAbs target the metastable cleaved pre-fusion conformation of Class I viral glycoproteins. Uses of X-ray crystallography for determining antibody-antigen interactions typically circumvent this challenge by crystallizing only the Fab and stabilized glycoprotein construct, which may include only subunites or monomers of the glycoprotein.[87–92] In cases where the complete oligomeric state of the glycofusion in complex with an antibody has been crystallized, either site-directed mutatagenesis was used to introduce disulfide bonds and/or antibody binding was required for additional stabilization of the viral fusion protein to form crystals.[68,93–97] These efforts to stabilize the full oligomeric form of fusion proteins is important for determining the relative orientation of the bound antibody as well as the fusion protein conformational state. However, stabilization of the RSV F protein in its pre-fusion exhibits up to 4.5 Å root mean square distance (RMSD) structural difference between the pre- and post-fusion conformations,[68,98] making it difficult to determine the structural heterogeneity induced by stabilizing mutations, antibody binding, or intrinic flexibility of the pre-fusion conformations. NMR and EM methods allow for the structural determination of proteins closer to their native state in solution, and therefore can be used to determine antibody-antigen complexes whose structural heterogeneity is prohibitive for X-ray crystallography. The disadvantage of NMR, however, is that NMR loses its sensitivity in measuring relaxation times for proteins larger than 25 kDa to 35 kDa, and relies on determination of multimeric proteins or protein complexes as separate subunits,[99] which makes the method intractable for epitope mapping of most fusion proteins. EM on the other hand, is well-suited for proteins and protein-complexes greater than 200 kDa and is increasingly used to determine antibody-antigen interactions by means of single particle EM.[100–103] Single particle EM does not require proteins to be deglycosylated or truncated, as opposed to X-ray crystallography, and therefore can be used to determine antibody-antigen interactions in their more native state.[104] In particular, when purified antibody-antigen complexes, or particles, are flash-frozen in vitreous ice, the particles are theoretically frozen in random orientations, providing a thorough two-dimensional sampling of all surfaces. The drawback of cryo-EM is that the contrast of an individual particle within vitreous ice is exceptionally low, and so during particle picking — the process of discerning a particle from the fixing medium within a collected image — the probability of filtering signal from noise is sometimes close to, or is, no better than picking random points within the image. Moreover, the creation of vitreous ice is hard to reproduce, and depending on the thickness of the ice, the vitreous ice can create imaging artifacts that distort the conformation of individual particles. Therefore, an exceptionally large number of particles, *e.g.* more than 1×10^5 particles, must be picked to prevent false positives during image reconstruction, but it can approach a high-resolution determination of antibody-antigen complexes at less than 3.0 Å, similar to X-ray crystallography.[105,106]

Negative-stain EM can be used in place of cryo-EM, but the use of uranyl acetate limits the resolution of reconstructed negative stain images to around 20 Å.[107] Nevertheless, this is still sufficient to determine the general binding orientation of an antibody-antigen complex.[108]

Structural determination methods require a relatively large amount of expressed protein, which is not tenable for all systems and are typically used as final validation of antibody-antigen complexes. Other methods, such as shotgun mutatgenesis, hydrogen-dueteritium exchange (HDX), yeast or phage display, and competition methods, do not provide structural insights to antibody-antigen complexes, but they do provide higher-throughput platforms by which sequence-dependent binding mechanisms can be interrogated. High-throughput shotgun mutagenesis was performed on DV E protein to determine residues critical for function[109] and mAb binding.[110] The comprehensive site-specific mutagenesis is very useful for determining interaction hotspots, but is difficult to extend to determining amino acid interaction networks that are critical for stability, especially for conformation-specific stability. Display methods, such as phage or yeast display, allow for a rapid generation of a large antibody repertoires which can be screened for specific binding interactions against stabilized antigen constructs.[111] HDX is very useful in determining residues within discontinuous epitopes, as it quantifies changes in solvent-accessible surface area (SASA) due to changes in binding surface area.[112] The most commonly used epitope mapping technique, the competition assay using surface plasmon resonance, assesses whether or not an antibody whose epitope is not determined binds to a similar binding surface as an antibody whose epitope is well-characterized.[113] This approach is very useful for screening panels of antibodies against stabilized constructs to determine commonly targeted sites of vulnerability as well as novel epitopes.

I.4.3. Limitations of epitope mapping of broadly neutralizing antibodies

Prior to determining B cell epitopes, the antibody and antigen must be obtained seperately. Most mAb are obtained, or isolated, via hybridoma screening, a process which typically takes 8 − 12 months to isolate a panel of neutralizing antibodies.[114] In the case of isolating human mAbs against a circulating virus, peripheral blood mononuclear cells (PBMCs) are isolated from donors following infection or immunization and immortalized by transformation into lymphoblastoid cells, which are then selected for survival and expansion. Afterwards, the transformed B cell cultures are screened for antigen-specific interactions by enzyme-linked immunosorbent assay (ELISA) and then fused with myeloid cells to generate hybridoma cells.[82] The hybridoma screening process poses several bottlenecks for selection of potential bnAbs. First, the efficiency of transformation is less than 50% even with improved transformation methods.[115] Second, ELISA screening of antibody-antigen complexes requires that the antigen used for screening is stably expressed, *e.g.* does not exhibit conformational heterogeneity during screening, or else the conformational epitope cannot be inferred from the ELISA screen. This later limitation is pertinent to the screening against conformation-specific epitopes of metastable conformations.

Advances in the understanding of the coformational dynamics of RSV F, HIV Env, and influenza HA proteins have led to the recent development of engineered, stabilized pre-fusion conformations and the isolation of highly-potent bnAbs and their high-resolution structural determination bound to novel epitope interfaces.[6,98,116–120] A common trait exhibited by these pre-fusion conformations is transient fluctuation, or breathing, that temporarily exposes conserved sites of vulnerability. Additionally, in two cases, binding of the bnAb to the temporarily exposed epitope induces a rapid disassociation of the fusion trimer,[98,120] suggesting that the mechanism of protection is in part due to the recognition of a discontinuous structural epitope of the fusion protein trimer. The screening of antibodies against antigen-specific targets is generally limited to only one conformation of viral glycoproteins, which negates the selection of alternative conformation-specific antibodies during the initial screening process. Furthermore, antibody screening does not always include screening against the native oligomeric state of each conformation, which may not fully screen for quarternary epitopes. These recent findings of "cryptic" epitopes suggest that there may be additional sites of vulnerability that have not been described yet, and that the antibody repertoire may contain other bnAbs that target transiently exposed viral glycoprotein surfaces that cannot be shielded by glycans. Additionally, these transiently exposed epitopes tend to be more highly conserved.[121] As mentioned in the deep mutational scanning of HA,[57] residues may be highly conserved within a specific subtype, but are distinct between between subtypes. Current epitope mapping strategies are time intensive and often rely on a single or small set of antigens with a limited number of sequences. Screening against the full scope of all possible mutations with a clade, let alone across multiple clades, would be cost and time prohibitive for each and every antibody, although would be useful in determining the breadth of antigen recognition and protection.

I.5. Predicting conserved epitopes requires knowledge of what drives sequence and structural conservation

As discussed in previous sections, the search for a universal vaccine is limited to reactively identifying common epitopes and is time-intensive. Experimental epitope mapping strategies cannot feasibly test for the complete sequence and conformational heterogeneity a quasispecies may exhibit during a round of infection. In the case of an outbreak of an emerging viral strain or pandemic that results in either a change in viral sequence or structure, current methods of identifying conserved antigenic targets struggle to provide a rapid response to help aid in the production of an effective vaccine. Computational methods, on the other hand, may assist in narrowing down potential antigenic targets by modeling the contraints that sequence and/or structural stability impose on viral glycoprotein fitness selection.

The body of this thesis discusses methods that assist in the prediction of viral glycoprotein mutation tolerance and sites of vulnerability. The second chapter uses computational protein design, more specifically multi-state protein design, to identify allowable mutations that maintain or improve the stability of a protein backbone ensemble. Additionally, the first chapter describes which metrics used to quantify protein flexibility also correlate with sequence conservation. The third chapter discusses current computational B cell epitope prediction methods and introduces how the metric described in Chapter 2, contact proximity deviation, can be used in conjunction with other previously known predictors of sites of vulnerability to predict conformation-dependent epitopes. However, this chapter is limited to predicting Fab-mediated antigen recognition, whereas the fourth chapter, provides a case example where Fab recognition of the same RSV F epitope may or may not confer protection due to the angle of the binding pose. Given that Chapter 2 and 3 rely on an understanding of computational protein modeling, the following section provides an overview of computational methods used to predict mutation tolerance and the limitations of these approaches.

I.5.1. An introduction to protein modeling and design approaches

The principle assumption of protein modeling is related to Anfinsen's thermodynamic hypothesis in that a protein's optimal conformation exists at its lowest relative free energy state and is dictated by its linear amino acid sequence.[34] However, estimation of a protein's native state is a non-trivial problem as first stated by Cyrus Levinthal — given that each peptide bond has an estimated 3^2 or 9 degrees of freedom, *i.e.* conformations, the number of conformations a single polypeptide chain can adopt quickly becomes astronomically large the longer the polypeptide chain, but must fold into a native state at biologically relevant timescales.[122] Computational protein modeling algorithms have tried replicate the parameters which drive protein folding *in silico* by defining the molecular forces that drive the rapid folding of a polypeptide chain into its native conformation.[123,124] These molecular forces that define the free energy of a conformation can be estimated by the sum of its hydrophobic, hydrogen bond, van der Waals, and ionic bond interactions within itself and with its surrounding molecules. The combined improvement in available computation power to tackle memory intensive problems and protein energy functions to better estimate protein structure and dynamics[112,125] has paved the way for the development of a plethora of computational protein structure prediction and design methods that address specific biological problems in finer detail.

Protein design specifically addresses the "inverse-folding problem" — provided a specific conformation, which linear amino sequences can adopt that conformation? Computational protein design methods typically employ either a *de novo* or template-based strategy. *De novo* strategies use a multiple sequence alignment of a target protein to its homologous sequences, which are used to predict common local secondary structure features, or fragments, that are then assembled into tertiary protein structural models and ranked using an all-atom energy scoring function.[126] The advantage of *de novo* protein design approaches is their ability to search the available conformational space of natural homologues to identify common structural similarities of diverse sequences.[127] The major pitfalls of such an approach, however, is that accurate fragment assembly relies on exhaustive sampling to assign accurate protein folds and that all fragment backbone geometries are represented within the fragment library. For proteins that adopt a single, native conformation, the use of fragment assembly has been shown to correctly approximate a global fold.[128] However, for proteins that adopt higher energy, metastable states of large proteins (such as RSV F in its pre-fusion conformation), accurate assignment fragments necessary to adopt the correct global fold are selected against during selection of sampled conformations unless provided experimental constraints, which may be difficult given the difficulty of obtaining structural data of metastable states.

Template-based protein design approaches also are limited by the availability of experimentally determined conformations, but do not have to assess whether or not a protein fold assembled *in silico* represents a protein in its native folded state. Instead, template-based protein design strategies assume that an experimentally determined structure represents its natively folded state, and are more commonly used to assess mutation tolerance. During template-based design modeling, one or more residues selected for design are evaluated by assigning an energy score using an energy scoring function, termed as the native or reference score, and then are removed from the backbone. Next, a rotamer library is used to randomly place an amino acid side chain configuration on the template backbone. Each "mutation" is assigned a score using the same energy scoring function used to evaluate the native residue, so that rotamers evaluated with a lower score than the reference score are selected as the designed sequence. At the end of a template-based design simulation, all favorably scored residues become the designed sequence corresponding with the template backbone. Most template-based protein design approaches consider the design of a single backbone template, which most likely underestimates the conformational space a single sequence occupies. To diminish the underestimation of available conformational flexibility of a single sequence, multi-state design approaches, such as REstrained CONvergence (RECON) multi-state design[129] as described in Chapter 2, use multiple discrete template backbones to select sequences that improve the stability not only a single template, but an ensemble of template backbones. Even so, the use of discrete, experimentally determined templates that depict local energy minima do not consider the local energetic constraints that higher energy states, such as transition states, place on the available sequence space of a backbone ensemble.

I.6. Significance and Innovation

As discussed in Sections I.4.2 and I.4.3, current structural methods have provided invalue models for mechanisms of antibody-mediated protection and/or neutralization by targeting viral fusion glycoproteins. However, these methods are limited addressing how changes in amino acid sequence affect antibody binding, at least on a scale equivalent to the number of mutations as quantified by deep mutational scanning or sequencing of viral glycoproteins. However, given that both deep mutational scanning and deep sequencing of viral glycoproteins have demonstrated that viral glycoproteins do not undergo similar mutation rates at all positions, especially within regions known to undergo conformational rearrangments during attachment and fusion, it is likely that these regions have limited mutational tolerances to maintain the similar changes in ΔG, in particular ΔS, to conserve the entropic changes necessary to facilitate the conformational changes required to complete hemifusion with the host target cell.

By assuming that viral glycoproteins must assume certain conformations during hemifusion, this thesis shows how multi-state design, can be used as an *in silico* approximiation of natural sequence variation, which is faster and much cheaper to perform than deep sequencing or deep mutational scanning. Although applied more generally to proteins that require substantial conformational changes necessary for function, Chapter 2 demonstrates that the simultaneous modeling of conformational and sequence space can approximate the natural selection of mutation preferences required to maintain a select conformational ensemble. Furthermore, Chapter 2 describes how local conformational flexibility, in particular amino acid side chain rearrangments, limit mutation tolerances. By describing a protein metric that is associated sequence convervaton and the relative free energy changes associated with conformational changes, the estimation of sites of vulnerability can be related to the physical requirement to conserve local sites of protein flexibility necessary to undergo the conformational changes necessary to complete fusion. This is in stark contract to other predictors of sites of vulnerability, such as surface accessibility or electrostatic potentials, which typically describe a single conformation.

CHAPTER II

MULTI-STATE DESIGN OF FLEXIBLE PROTEINS PREDICTS SEQUENCES OPTIMAL FOR CONFORMATIONAL CHANGE

This chapter is based on the publication "Multi-state design of flexible proteins predicts sequences optimal for conformational change". Marion F. Sauer contributed to the development of the benchmark, performance of experiments, analysis of the data, and writing the aritcle.

Computational protein design of an ensemble of conformations for one protein – i.e., multi-state design – determines the side chain identity by optimizing the energetic contributions of that side chain in each of the backbone conformations. Sampling the resulting large sequence-structure search space limits the number of conformations and the size of proteins in multi-state design algorithms. Here, we demonstrated that the REstrained CONvergence (RECON) algorithm can simultaneously evaluate the sequence of large proteins that undergo substantial conformational changes. Simultaneous optimization of side chain conformations across all conformations increased sequence conservation when compared to single-state designs in all cases. More importantly, the sampled sequence space of RECON designs resembled the evolutionary sequence space of flexible proteins, particularly when confined to predicting the mutational preferences of limited common ancestral descent, such as in the case of influenza type A hemagglutinin. Additionally, we found that sequence positions which require substantial changes in their local environment across an ensemble of conformations are more likely to be conserved and whose conservation rates are better simulated by the consideration of multiple local side chain environments during design. To quantify this rewiring of contacts at a certain position in sequence and structure, we introduced a new metric designated 'contact proximity deviation' that enumerates contact map changes. This measure allows mapping of global conformational changes into local side chain proximity adjustments, a property not captured by traditional global similarity metrics such as RMSD or local similarity metrics such as changes in ϕ and ψ angles.

II.1. Introduction

Computational protein design solves the so-called 'inverse folding problem' by identifying an amino acid sequence that is compatible with a given protein structure, i.e., backbone conformation and possibly interactions with partner biomolecules. This approach allows for the molecule to conduct its function in this single state. Protein function, however, often relies on the transition between multiple conformations – a protein must be thermodynamically stable in multiple conformations before it is capable of achieving a defined function. Thus, for a protein to conserve its function, we hypothesized that the conservation of protein flexibility limits the protein's sequence space to be consistent with the conformational changes needed for function. Determining functionally relevant sequence tolerance, or rather, the set of amino acid sequences that are allowable given a protein's function, therefore depends on identifying the set of amino acid sequences that is stable in each of the conformations needed.

Testing this hypothesis is complicated, as typically not all functionally relevant conformations have been determined experimentally. The picture gets even more complicated if we look not only at functionally relevant conformations that are by definition local free energy minima (*i.e.*, thermodynamics) but also include an analysis of the height of barriers connecting these states that determine the kinetics of interconversion.

Humphris-Narayanan and colleagues demonstrated that prediction of mutation preferences of HIV-1 protease and HIV-1 reverse transcriptase was improved up to 25% when structural ensembles were included during protein design, as opposed to design of a single conformation.[131,132] The structural ensembles used for this approach were generated by reverting all structural side chains to a consensus sequence and using Rosetta Backrub to introduce small local rotation about the $C_\alpha - C_\alpha$ axis of each of three-residue segments while maintaining ideal bond length, angle, and the starting χ_1 angle[133,134] at sites distributed throughout the protein known to acquire mutations. Next, protein design was performed on each backbone within the ensemble to select for mutations sequence that contributed to increased protein fold, dimer, and peptide stability. which was calculated as the lowest weighted sum of energy scores. They found that the substitution frequency of the consensus sequence, or profile, of the backbone ensemble better corresponded to the mutation frequencies observed within the Stanford HIV-1 Database[135] than the substitution frequencies obtained from design of an individual conformation. Additionally, they showed that the sequence profiles acquired with their Backrub ensembles were similar to the sequence profiles attained using an ensemble of experimentally-derived structural models. With such results, this approach succeeded in showing that representation of conformational plasticity during protein design better mimicked the mutational tolerances of HIV-1 protease and reverse transcriptase, which the authors attribute the requirement of small backbone changes to accommodate mutations from the starting sequence.

For certain proteins, sub-Angstrom perturbations of the peptide backbone are not sufficient to represent the conformational space consistent with their function. In the case of ubiquitin, Friedland and colleagues also used Rosetta Backrub to generate ubiquitin ensembles, but unlike in the previously mentioned method, they randomly inserted local rotations about the $C_\alpha - C_\alpha$ axis of two to twelve residues to diversify the conformational space of the generated ensembles, and then culled any generated models which did not agree with NMR residual dipolar couplings (RDCs), thus generating ubiquitin ensembles more similar to native-state solution dynamics.[136] Using these RDC-constrained ensembles for design, they demonstrated that the mutation profiles obtained from the collection of individually-designed poses were more consistent with sequences within the ubiquitin family, which was calculated from the sum of substitution costs of each designed pose from the aligned ubiquitin consensus sequence. In combination with the aforementioned study, these approaches demonstrated that a requirement for protein flexibility of a native-state ensemble substantially dictates the sequence space available for evolution.

A limitation of these approaches, however, is the assumption that the tolerated sequence space for a conformationally flexible protein can be determined by integrating over each SSD profile of each conformation within an ensemble, *i.e.*, enumerating the most energetically favorable amino acid for each position and each conformation. However, the most energetically favorable amino acid, as determined by the aforementioned methods, may be the most energetically favorable for a one or more single conformations, but may not be energetically tolerable at the same position in another conformation. For instance, in a certain conformation, the energetically most favorable amino acid might be the only allowed amino acid, with all others prohibited (imagine a tiny space where only glycine fits). At the same position in other conformations, there may be acceptable alternatives with more energetically favorable scores, but those residues could not be tolerated as an acceptable mutation in the aforementioned more constrained position. Thus, we hypothesized that MSD over all conformations relevant for function will yield a more accurate representation of the biologically relevant sequence space compatible with function.

Using a pre-defined scoring function, positive-state MSD approaches rank the stability of a sequence within an ensemble as the average stability when threaded over each state across an ensemble.[137] For most MSD approaches, replacement of the starting, or native, sequence occurs only if the designed mutation is lower than the native average score, meaning that, although a sequence may not lower the evaluated stability of a single conformation, it lowers the stability of an ensemble as a whole from the native ensemble. The Best Max-Marginal First (BMMF) algorithm was used to demonstrate that the MSD of 16 unique calmodulin-substrate complexes increased the similarity of the designed calmodulin binding site to evolutionary sequence profiles by two-fold, and increased in native sequence recovery from 52.5% for SSDs to 80% for the 16-state design scenario (8). Challenges for applying MSD methods like the BMMF algorithm, however, are the efficiency of the search algorithm, large memory requirements, and extended computational time needed. MSD methods up to now have been limited to designing a small number of amino acid positions across all states, with the largest number of simultaneously evaluated design positions being 27 designed positions across 60 states using the MSD FASTER algorithm.[129,138–140]

We sought to study large proteins that undergo conformational rearrangements that include domain or hinge displacements of greater than a few Å in RMSD. We expected that the tolerated sequence space must be restricted in some regions to allow for substantial 'rewiring' of contact networks when transitioning from one state to another. The tolerated sequence space of these types of conformational changes is not limited to local regions, such as protein-protein interfaces, but instead distributed over the entire amino acid sequence. Thus, an MSD approach that seeks to explore such sequence spaces needs to include the entire protein. The RECON algorithm was used previously to estimate the sequence tolerance within protein-protein interfaces. However, already at that time this approach proved to be more computationally efficient than the generic Rosetta MSD algorithm.[129] With the addition of a message passage interface (MPI), RECON MSD can combine the SSD efficiency of evaluating the sequence tolerance of a full-length protein with the MSD capability of evaluating the fitness function of a sequence across multiple conformations.[141]

Highly flexible viral glycoproteins, such as the influenza A HA protein and its stem domain (HA_2), undergo conformational rearrangements of greater than 30 Å and have been shown to be conserved in sequence greater than 90% percent across subtypes.[142] For other highly flexible proteins, such as calmodulin, kinases, and voltage-gated sensory channels, regions known to mediate conformational change can be conserved up to 100% across phylogenies, suggesting that a limited set of sequences is suitable for select conformation transitions.[143–145] Here we use the Rosetta RECON MSD algorithm to demonstrate that the sequence space consistent with all experimentally determined conformations of a protein approximates sequence profiles observed in evolution.

II.2. Results

It is our aim to demonstrate that simultaneous evaluation of sequence space across an ensemble of conformations improves the correspondence of the designed sequences to an evolutionary sequence profile by considering the constraints that local and global protein flexibility impose on rotamer placement. For this benchmark, we perform RECON MSD and compare the designed profiles to SSD

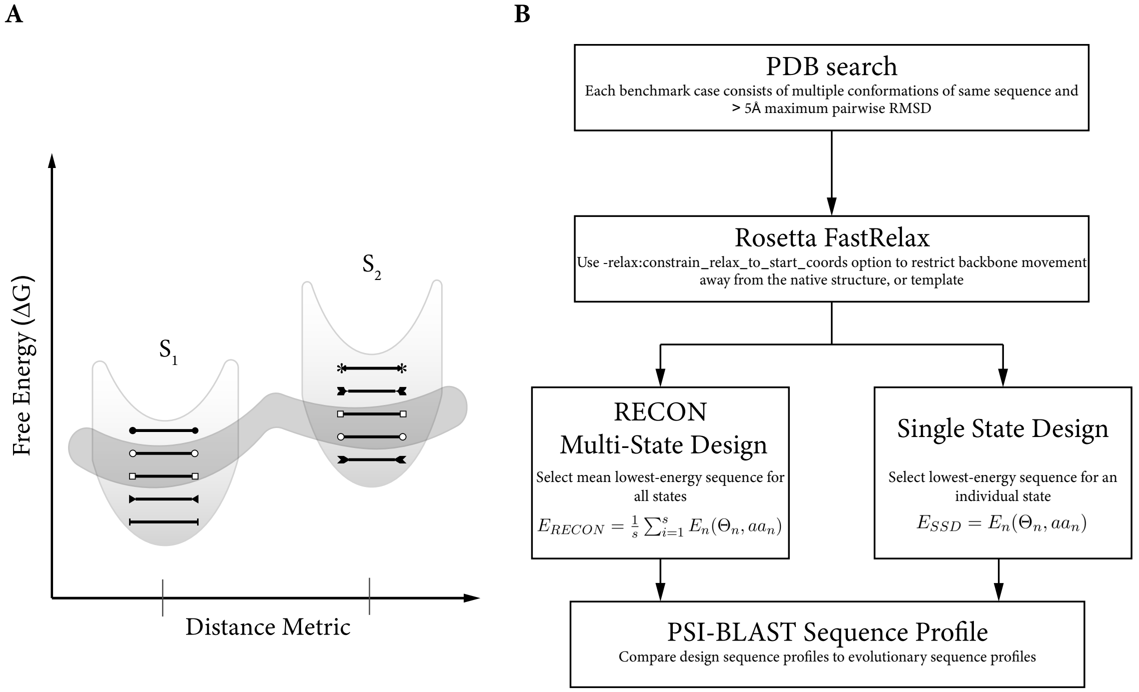

A
B
Free Energy (ΔG)
Distance Metric
S_1
S_2
PDB search
Each benchmark case consists of multiple conformations of same sequence and > 5Å maximum pairwise RMSD
Rosetta FastRelax
Use -relax:constrain_relax_to_start_coords option to restrict backbone movement away from the native structure, or template
RECON Multi-State Design
Select mean lowest-energy sequence for all states
$E_{RECON} = \frac{1}{s}\sum_{i=1}^{s} E_n(\Theta_n, aa_n)$
Single State Design
Select lowest-energy sequence for an individual state
$E_{SSD} = E_n(\Theta_n, aa_n)$
PSI-BLAST Sequence Profile
Compare design sequence profiles to evolutionary sequence profiles

Figure II.1.: Graphical representation of hypothesis and experimental design. *(A) Schematic of sequence space and the impact of flexibility on sequence tolerance. S_1 and S_2 represent two unique conformations of the same residue length separated by some RMSD that populate two local energy minima. Black lines with end caps represent unique sequences that are energetically most favorable for a single conformation. The dark shaded area encircles sequences that are energetically favorable for both conformations. Here we illustrate that by using multiple conformations during protein design, we identify sequences that are energetically suitable for conformational flexibility, yet are not necessarily the most stable sequence for any given conformation. Additionally, the requirement to adopt multiple conformations constrains the number of suitable sequences (B) Flow chart of benchmark design.*

and PSI-BLAST profiles to quantify the similarity between designs and evolutionary sequence profiles (figure II.1).

II.2.1. Compilation of a benchmark set of eight proteins

We selected proteins with multiple known conformations of identical sequence from the PDBFlex database.[146] The benchmark included eight proteins, requiring that each benchmark case have at least two published conformations with an RMSD greater than 5 Å, and an identical sequence greater than 100 amino acids in length (Table 1). We omitted duplicate conformations, which we define as conformations with and RMSD of less than 0.5 Å, to avoid design bias towards similar conformations. In addition, we used a resolution cutoff of 5 Å with the requirement that greater than 75% of the included models within each design ensemble were determined at a resolution of better than 3 Å. We also omitted any models with longer sequence gaps or missing densities. For structural models with chain breaks that had missing density for only one or two consecutive residues (PDB IDs 1OK8, 3C5X, and 3C6E of the DV E protein monomer) we added the missing densities with the Rosetta loop modeling application.[147] All structural models were gently relaxed with a restraint to start coordinates to remove any energetic frustrations frequent in models derived from low-resolution experimental structures.

II.2.2. Metrics to measure amplitudes of local and global conformational change

Quantification of protein flexibility commonly relies on the structural comparison of two structural models, whether that be through the similarity of equivalent atoms in three-dimensional space, calculated as RMSD, or by the similarity of equivalent ϕ and ψ backbone dihedral angles, calculated as dihedral angle root mean square deviation (RMSD$_{\text{da}}$) of a C_α atom..[148] RMSD is used frequently as a global metric used to describe the overall similarity of two conformations of the same protein and has been a powerful metric to quantify overall structural similarity. RMSD$_{\text{da}}$, on the other hand, is used to describe local backbone displacements and is well-established, for example, to compare loop conformations. The disadvantage of both metrics is that they do not capture whether or not a particular residue is reconfigured in its interactions with neighboring amino acids. However, we hypothesize that such a metric of local rewiring driven by a global conformational space will best correlate with restrictions in sequence space introduced through conformational flexibility. Thus, we settled on three metrics that capture the structural dissimilarity of a protein ensemble in terms of its maximum global structural dissimilarity, local backbone dissimilarity, or contact map dissimilarity: 1) The maximum pairwise RMSD of all atom coordinates of two superimposed structures within a set of n superimposed structures was used as a metric to describe the maximal global conformational change an ensemble undergoes (Panel A in figure II.2 on page 22). To allow for comparison of RMSD values between benchmark cases that involve proteins of different size, we used RMSD100, a RMSD value normalized to protein of length 100 amino acids.[149] 2) Residue ϕ and ψ RMSD$_{\text{da}}$ was used as a local metric of similarity (Panel B in figure II.2 on page 22). This metric will directly identify hinge regions between moving domains. 3) Lastly, we designed a metric that captures changes in the contact map computed as $C_\beta - C_\beta$ distance variation. This metric captures local changes in the environment of a residue by including non-local tertiary contacts in the analysis. Thus, it is designed to capture the local and global changes of the physicochemical environment of a residue and thus defines which amino acids are tolerated in a certain position (Panels C and D in figure II.2 on page 22). For a complete description of each metric, see section II.5 on page 37.

II.2.3. RECON MSD samples sequence profiles that are more similar to evolutionary observed sequence profiles when compared to SSD

We first examined the correspondence of native sequence recovery determined by MSD versus SSD designed sequences to conservation rates within natural homologues. To accomplish this goal, we performed either RECON MSD or SSD on each set of protein conformations, allowing for the substitution of the native residue to all twenty amino acids and ignoring the presence of any disulfide bonds present in the native model. Designed sequence profiles of each conformation were generated using ten designed sequences, which were selected from either the ten lowest-scoring designed ensembles, as in the case of RECON MSD, or conformations for SSD. The mean total score of all conformations designed during the same RECON MSD run was used to sort and select the ten lowest-scoring designed ensembles. From the ten mean lowest-scoring ensembles, the set of ten models of each conformation was used for analysis. For SSD, each conformation was designed independently, and therefore the ten lowest-scoring models of each conformation were used for analysis. Therefore, the distinction between

PROTEIN	PDB ID OF STATES USED WITHIN EACH ENSEMBLE	DETERMINATION METHOD	DESIGNED POSITIONS	AVERAGE PAIRWISE RMSD OF DESIGNED STATES
5'-NUCLEOTIDASE	1HPU 1O18 1IOD, chain A 1IOD, chain B 4WWL	X-ray X-ray X-ray X-ray X-ray	523	5.18 ± 1.12
ADENYLATE KINASE	1AKE 4AKE	X-ray X-ray	214	7.19
CAGL	3ZCJ 4CII 4YVM	X-ray X-ray X-ray	169	27.0 ± 22.2
CALMODULIN	1A29 1CFC 1CFD 1CFF 1CKK 1CLL 1CM1 1CM4 1G4Y 1LIN 1MUX 1NIW 1NWD 2F2P 2N8J 2WEL 3EWT 3EWV 4DJC 4HEX	X-ray NMR NMR NMR NMR X-ray X-ray X-ray X-ray X-ray NMR X-ray NMR X-ray NMR X-ray X-ray X-ray X-ray X-ray	169	27.0 ± 22.2
DENGUE VIRUS ENVELOPE PROTEIN (MONOMER)	1OAN 1OK8 3C5X 3C6E 3J27 3J2P	X-ray X-ray X-ray X-ray Cryo-EM Cryo-EM	394	6.63 ± 2.89
GROEL SUBUNIT	1AON, chain A 1AON, chain B 2C7E 3WVL 4AB3 4KI8	X-ray X-ray Cryo-EM X-ray Cryo-EM X-ray	523	9.06 ± 1.13
INFLUENZA HEMAGGLUTININ STEM (TRIMER)	1QU1 1HTM 2HMG 3EYM	X-ray X-ray X-ray X-ray	344	23.7 ± 17.4
RESPIRATORY SYNCYTIAL VIRUS FUSION PROTEIN (TRIMER)	3RKI 3RRR 4MMS 4ZYP	X-ray X-ray X-ray X-ray	1252	29.9 ± 22.0

Table II.1.: Proteins used in conformation-dependent sequence tolerance benchmark. *For a complete description of Protein Data Bank (PDB) identification and sequence information included in the benchmark, see S1 Table.*

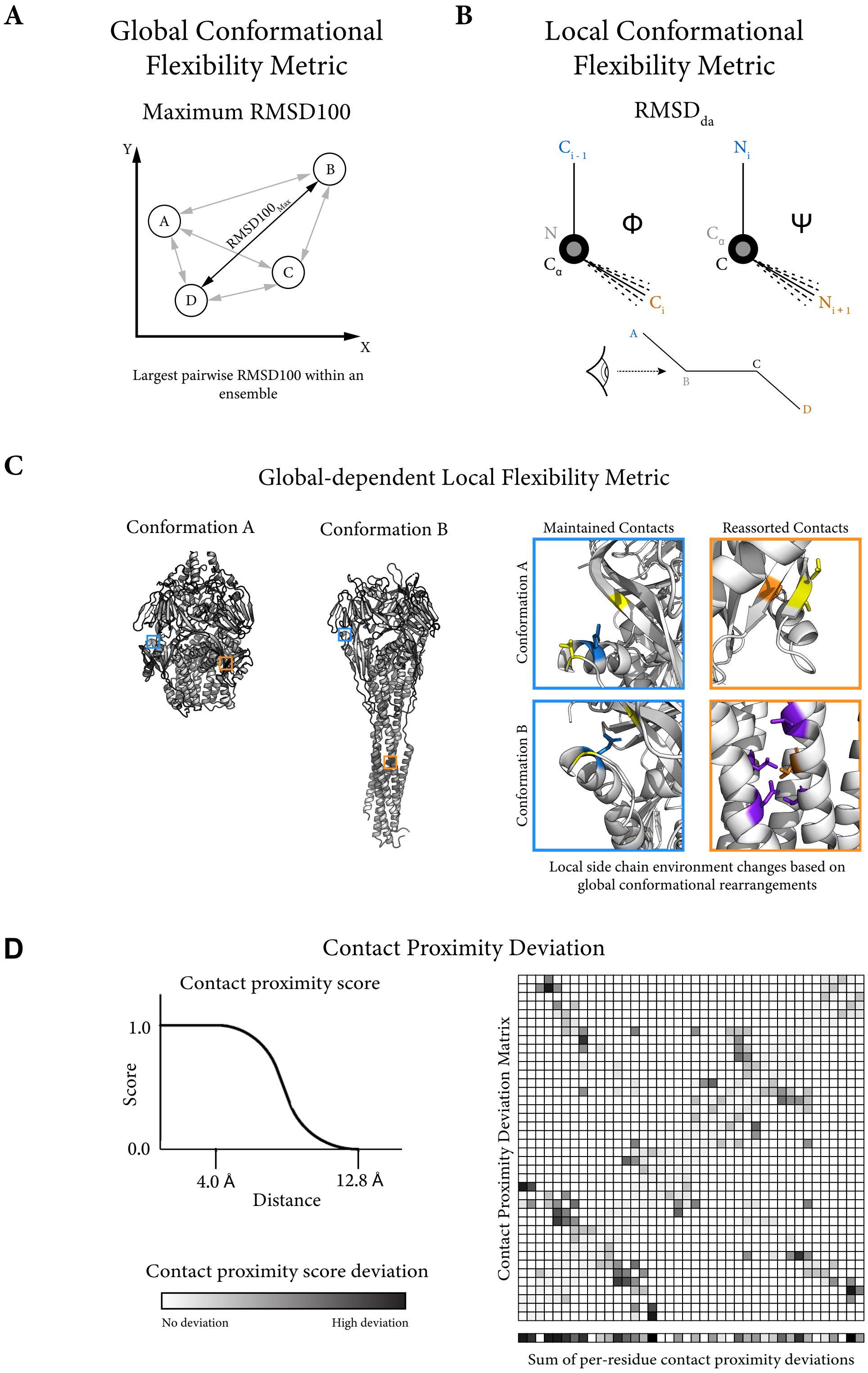

A
Global Conformational Flexibility Metric
Maximum RMSD100
Y
X
A
B
C
D
RMSD100 Max
Largest pairwise RMSD100 within an ensemble
B
Local Conformational Flexibility Metric
RMSDda
C i - 1
N
Cα
Φ
Ci
Ni
Cα
C
Ψ
N i + 1
A
B
C
D
C
Global-dependent Local Flexibility Metric
Conformation A
Conformation B
Maintained Contacts
Reassorted Contacts
Conformation A
Conformation B
Local side chain environment changes based on global conformational rearrangements
D
Contact Proximity Deviation
Contact proximity score
1.0
0.0
Score
4.0 Å
12.8 Å
Distance
Contact proximity score deviation
No deviation
High deviation
Contact Proximity Deviation Matrix
Sum of per-residue contact proximity deviations

Figure II.2.: *Metrics used to quantify large-scale, or global, conformational flexibility. Caption continued on next page.*

Figure II.2.: Metrics used to quantify conformational flexibility. *(A) Illustration of maximum RMSD100, the metric used to quantify large-scale, or global, conformational flexibility. For simplicity, we only represent RMSD on a two-dimensional plane, where the x and y axes represent the difference in distance of cartesian space if two conformations were superimposed onto the same coordinate system. Each protein conformation of identical sequence is represented as a circle, and is separated by some distance vector evaluated as the RMSD100 of two conformations. The maximum RMSD100 describes the greatest pairwise RMSD100 within an ensemble. (B) Illustration of dihedral angle ϕ and ψ variation used to calculate dihedral angle RMSD ($RMSD_{da}$). Orientation of atoms is color-coded and corresponds to the diagram drawn at the bottom of the panel. $RMSD_{da}$ is illustrated as the range of dotted lines, corresponding to the deviation in relative orientation of the third and fourth atoms. (C) Explanation of contact proximity deviation. Two conformations of the same protein are depicted in the left, with two residues, outlined in cyan or orange, shown in their respective positions. These two residues are magnified (top right) in their local side chain environment in Conformation A on the top and Conformation B on the bottom. Contact residues in Conformation A are colored yellow. If the same contacts are maintained in Conformation B, contact residues remain colored yellow in the bottom two boxes. If new contacts are made, contact residues are colored in purple. Even though the cyan residue changes slightly in its relative orientation between conformations, the same contacts are maintained so that the degree of conformational flexibility is relatively low in comparison to the heptad trimer refolding, and would have a low contact proximity deviation score. In contrast, the orange residue completely rearranges its local side chain contacts between conformations as a result of the large conformational rearrangement, and would have a high contact proximity deviation score. (D) Explanation of contact proximity deviation. We assigned a score to each $C_\beta - C_\beta$ distance by applying a soft-bounded, continuously differentiable function that accounts for the proximity of two side chains and approximates the likelihood of two side chains forming a contact, illustrated in the top left of Panel D. We then calculated the deviation of each $C_\beta - C_\beta$ distance across an ensemble as shown in the matrix, with low deviation scores in white and high scores in black. The contact proximity deviation score represents the sum of all $C_\beta - C_\beta$ proximity deviations a single residue undergoes within an ensemble, as shown in the bottom row separated from the matrix.*

RECON MSD and SSD selected models rested in whether or not each of the ten selected models of each conformation where evaluated for design collectively as an ensemble or not. To compare the designed sequence profiles to that of the sequence diversity in natural homologues, the native sequence was used as the PSI-BLAST query sequence to generate PSI-BLAST profiles for each protein. Native sequence recovery was calculated as the mean percentage of conservation of the starting, or native, sequence for all designed positions, and for consistency, we term the percentage of the query sequence used to generate each PSI-BLAST profile as the percent native sequence recovery.

Simultaneously sampling across multiple conformations significantly restricted sequence sampling, or in other words, was more likely to conserve the native sequence, where RECON MSD had total native sequence recovery of 87.8 ± 4.5% versus SSD with 48.9 ± 11.1% native sequence recovery (figure II.3 on the following pageA). In contrast, PSI-BLAST profiles had a native sequence recovery of 82.11 ± 11.2%. Qualitatively, the PSI-BLAST profiles were much more similar to the predicted sequence tolerance of RECON MSD compared to SSD, yet a Mann-Whitney U test[150] indicated a significant difference of mean native sequence recovery of either design protocol compared to PSI-BLAST sequence tolerances, with a significance of $p = 0.0029$ for RECON MSD and $p < 0.00001$ for SSD. Total sequence recovery is a coarse approximation of sequence similarity, and fails to determine if the designed sequence profiles are sampling similar mutation preferences as observed in evolution. Therefore, we calculated an average total deviation score of each observed position-specific mutation profile to the corresponding PSI-BLAST profile of each protein (Panel B in figure II.3 on the next page and S1 Fig, which depicts the root mean square deviation of all aligned positions' profiles between a natural homologues' mutational preferences and those predicted by design. We found that in seven out of eight cases, a RECON MSD mutation profile resembled its corresponding PSI-BLAST profile more closely than the SSD mutation profile, as the root mean square deviation was lower for comparisons between PSI-BLAST profiles and RECON MSD profiles than that of between SSD profiles.

II.2.4. RECON MSD underestimates amino acid exchangeability, but samples a more evolutionarily relevant sequence space than SSD

Although RECON MSD more closely resembled PSI-BLAST sequence profiles on a per-case basis, we wanted to identify trends in sequence sampling in relation to the PSI-BLAST profiles to highlight design-sampling biases. This task was achieved by calculating the frequency an amino acid is conserved or mutated to another residue, or, the mean amino acid substitution frequency. In general, RECON MSD is more likely to conserve a native amino acid compared to a PSI-BLAST profile, whereas SSD is much more likely to replace the native amino acid (Panel A in figure II.4 on page 27 and S2 Fig). We examined amino acid exchangeability as the frequency of exchanging a native for a non-native amino acid. On average, PSI-BLAST profiles exchanged a native for non-native amino acid 1.32 ± 0.03% of the time, versus 0.77 ± 0.02% for RECON MSD and 2.45 ± 0.07% for SSD (Panel B in figure II.4 on page 27). Additionally, we compared the average difference of exchangeability for each residue as observed in the PSI-BLAST profiles versus either RECON MSD or SSD and found that RECON MSD average exchangeability rates of each residue are more similar to PSI-BLAST values than SSD (Panel C in figure II.4 on page 27). With the exception of phenylalanine or tyrosine, the difference between exchangeability rates for residues with larger side chains diminishes for RECON MSD, but becomes more exaggerated for SSD, as compared to observed mutation rates in evolution. This finding suggests that the inclusion of multiple conformations during design encourages better placement of bulky side chains, albeit with conservative placement. However, when comparing the linear regression model of

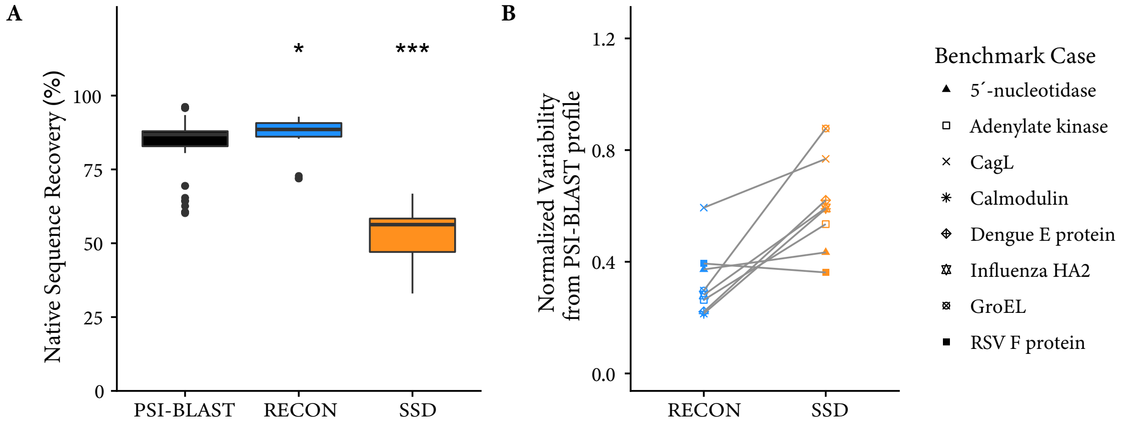


Figure II.3.: Design native sequence recovery and mutation profile variability comparisons to PSI-BLAST profiles.*(A) Comparison of total native sequence recovery of relaxed and unminimized RECON MSD and SSD designs to PSI-BLAST sequence profiles generated using the native sequence. For this figure and all subsequent boxplots, shaded regions of each box plot denote values within the first and third quartiles (interquartile range, or IQR), with the median indicated as a solid line and whiskers representing values ±1.5 × IQR. Outliers are represented as dots. Asterisks indicate the significance of difference of means of each design in comparison to the PSI-BLAST profile, with a z-test $p < 0.01$ represented by one asterisk, and $p < 0.00001$ by three asterisks. The p-value provided in this figure and all subsequent figures represents a two-sided, 95% confidence interval. (B) Mutation frequency root mean square deviations of designs in comparison to a PSI-BLAST profile, normalized by protein length. The y-axis values represent the average variability of mutation profiles for each designed residue in relation to a PSI-BLAST profile, represented as $y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{20}(aa_{PSI-BLAST} - aa_{design})^2}$ where aa_j represents the frequency of an amino acid observed at position i for each of all twenty amino acids (j), and y is the sum of all i differences for all amino acids within a protein of length n residues. A y-value of 0 would indicate that the design profile is identical to the PSI-BLAST profile, and an increase in y-value indicates the average frequency variance of the sequence profile for each residue is more dissimilar to a PSI-BLAST profile.*

individual exchangeability rates of either RECON MSD or SSD to PSI-BLAST rates, both designs were roughly equally dissimilar to PSI-BLAST exchangeability rates, with RECON MSD having a correlation coefficient of $r = 0.35$ and SSD with $r = 0.64$ (S3 Fig). Given that exchangeability rates were not normally distributed, a Kendall τ_β rank correlation coefficient[151] was computed to measure the ordinal association of design and PSI-BLAST amino acid exchangeability rates, where a coefficient of $\tau_\beta = 0$ would indicate that the amino acid exchangeability rates are identical. We found RECON MSD to have a $\tau_\beta = 0.283$, $p \leq 2.22 \times 10^{-16}$ versus $\tau_\beta = 0.372$, $p \leq 2.22 \times 10^{-16}$ for SSD when measured for its association to PSI-BLAST amino acid exchangeability rates. In addition, we compared the difference in exchangeability rates between design and PSI-BLAST by calculating the ratio of transformed exchangeability rates (e) as $\frac{e_{design}+0.00001}{e_{PSI-BLAST}+0.0001}$ to avoid division by zero, where an individual exchangeability rate would be equivalent between design and PSI-BLAST if $e_{ratio} = 1$. We found that for RECON MSD the mean $e_{ratio} = 2.49$ and for SSD the mean $e_{ratio} = 22.7$. A Mann-Whitney U test of matched individual ratios found a significant difference between RECON MSD and SSD exchangeability rate ratios to PSI-BLAST exchangeability rates, with $p < 0.0001$. Taken together, RECON MSD is sampling individual mutation preferences significantly more closely to that observed in evolution than SSD.

II.2.5. RECON MSD prediction of mutation preferences matches mutation profiles of natural homologues

The application of RECON MSD is not only optimized to engineer a stable, flexible protein, which is also readily accomplished by methods such as consensus sequence design.[152] Rather, RECON MSD explores the sequence space consistent with a protein's flexibility. Thus, we hypothesized, that the sequence space sampled by RECON MSD has similarity to the sequence space sampled in evolution. However, RECON MSD might explore additional sequences which have not yet been explored by natural selection. To answer this question, we compared calmodulin sequence profiles predicted by RECON MSD and SSD to naturally selected calmodulin sequence variation within a curated dataset of calmodulin representative of evolution across all eukaryotes.[153] We also compared the mutation profiles of influenza virus HA_2 predicted by either design with mutation profiles obtained from the NCBI Influenza Virus Resource (IVR).[154] Given that the HA_2 structural models used for this benchmark are of the influenza A H3N2 subtype,[155–158] we included all influenza virus type A HA_2 sequences deposited in the IVR to generate an HA_2 profile. Briefly, in each case, we performed a multiple sequence alignment of all sequences within each database to generate sequence profiles. From each profile, we measured the root mean square deviation of mutation frequencies to corresponding mutation frequencies predicted by RECON MSD or SSD. For a complete description of the generation of sequence profiles, please consult section II.5 on page 37.

The calmodulin and influenza virus type A HA_2 mutation frequencies had a root mean standard deviation of 0.473 and 0.580 with respect to RECON MSD profiles and 0.632 and 0.799 to SSD profiles, respectively (figure II.5 on page 29). Although RECON MSD profiles were more likely to match the mutation tolerances observed within either multiple sequence alignment, the improvement was not uniform for all residues. In general, RECON MSD was more likely to predict matching mutation profiles for residues that undergo local conformational rearrangements, For calmodulin, RECON MSD was more likely to improve the prediction of mutation tolerances within the calmodulin EF-hand at conserved motif positions 3, 5, 9, and 12.[159] Even though neither ligands nor water molecules were included during protein design, sequence profiles for positions 9 and 12, which are known to provide a bridged water or direct binding to Ca^{2+}, respectively, were similar. Within HA_2 RECON MSD profiles, residues, charged residues within the B loop, which rearranges into an alpha helix in the post-fusion

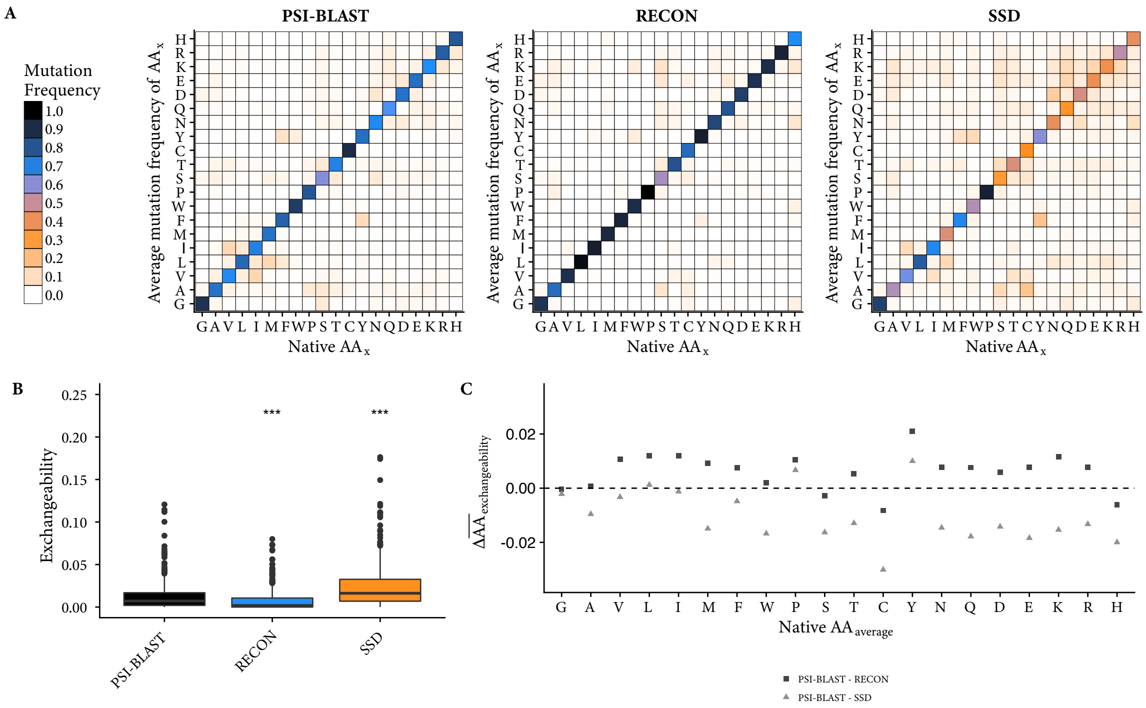


Figure II.4.: Comparison of exchangeability rates. *(A) Average amino acid exchangeability of PSI-BLAST, RECON MSD, and SSD sequence profiles. Single-letter amino acid codes were used for both x and y axes, with the x axis representing the original amino acid and the y axis representing the average mutation frequency the original amino acid to the indicated mutation. (B) Comparison of exchangeability rates between profiles, excluding rates of native sequence conservation rates. The y axis represents the mean frequency a native amino acid is replaced with a specific, non-native amino acid, which we term as amino acid exchangeability. (C) Difference of mean amino acid-specific exchangeability observed in a PSI-BLAST profile compared to a design profile. The x axis represents each type of amino acid present in the native sequence. The y axis represents the difference in average exchangeability frequency of each amino acid type, or rather, the average frequency a native amino acid type is replaced with any other non-native amino acid. A positive value indicates the native amino acid is less likely to be exchanged for a non-native amino acid during design, whereas a negative value indicates the native amino acid is more likely to be exchanged, as compared to a PSI-BLAST profile.*

conformation, and the S5 loop region, which stabilizes the rearranged B loop alpha helix,[89] were more likely to have similar predicted mutational tolerances as observed in influenza type A mutation profiles. However, in some cases, RECON MSD residue profiles either failed to improve or even worsened in predicting mutation profile similarity to profiles obtained from multiple sequence alignments as compared to SSD. This was particularly true for positions with small, non-polar residues that maintained contacts within the loosely packed interior within the HA_2 trimer, including S93, V100, S113, and I149. Additionally, RECON MSD poorly predicted the conservation of charged residues whose side chains faced the protein interior in at least one conformation of either calmodulin or HA_2 trimer, including calmodulin residues Q47, R84, D53, D90, Q132, and D126, and HA_2 residue H106.

Because we used only H3N2 backbones to predict HA_2 mutation profiles, we subdivided the HA_2 profile obtained from all influenza type A multiple sequence alignment into different groups and subtypes including H1 and H2 from group 1 and H3, H4, H7, and H3N2 from group 2 to compare design profile similarity to subtype-specific mutation tolerances (figure II.6 on page 31, S4 Fig). Separation of HA_2 sequences by subtype revealed a divergence in similarity according to related subtypes, with RECON MSD sampling mutation profiles much closer to subtypes within the H3 clade, including H3 and H4. Even so, a Levene's test for equality of variances[160] comparing mutation frequency variance within the H3N2 profile generated by RECON MSD to any influenza subtype A HA_2 IVR profile indicated no significant difference between mutation frequency variances within the RECON MSD H3N2 profile and any HA_2 IVR subtype profile. This suggests that RECON MSD samples similar mutational tolerances found across naturally selected influenza subtype A mutation tolerances, despite the diverging similarity to group 1 profiles. Given the high sequence profile similarity between RECON MSD and IVR H3 clade profiles, we conclude that RECON MSD can be used to predict possible sequence variation of closely related homologues from a single sequence. This might be particular useful in the case of predicting common mutations that arise due to genetic drift or reassortment, given that the HA_2 profile modeled by RECON MSD was not significantly different from multiple HA_2 subtype profiles. In cases where RECON MSD substantially deviates from observed mutation frequencies, particularly from the consensus sequence, such mutations warrant further experimental investigation to examine whether these stabilizing mutations within the H3 clade are artifacts of the Rosetta energy scoring function and/or RECON algorithm sampling, or that they indeed are evolutionary unexplored stabilizing mutations.

II.2.6. Sequence conservation is dependent on its contact map as computed by $C_\beta - C_\beta$ distance deviations

To consider the effect of conformational flexibility on sequence conservation, we examined the dependency of native sequence recovery on different aspects of conformational flexibility using the aforementioned metrics, maximum RMSD100, $RMSD_{da}$, and contact proximity deviation. We performed a Kendall τ_β rank correlation test on each profile to test for the strength of dependency of native sequence recovery on each metric (figure II.7 on page 32).[151] Of the three metrics, the native sequence recovery, or rather percent conservation, observed in PSI-BLAST profiles was only dependent on contact proximity deviation z-score, with $p = 1.79 \times 10^{-6}$, versus $p \leq 0.144$ for all other tests. RECON MSD and SSD native sequence recovery depended on both $RMSD_{da}$ and contact proximity deviation z-score ($p < 0.01$). Native sequence recoveries of both designed profiles depended strongly on $RMSD_{da}$, with $p \leq 2.22 \times 10^{-16}$, and had similar τ_β coefficients, with $\tau_\beta = 0.101$ for RECON MSD and $\tau_\beta = 0.0806$ for SSD. This finding may suggest that the Rosetta scoring function employed by

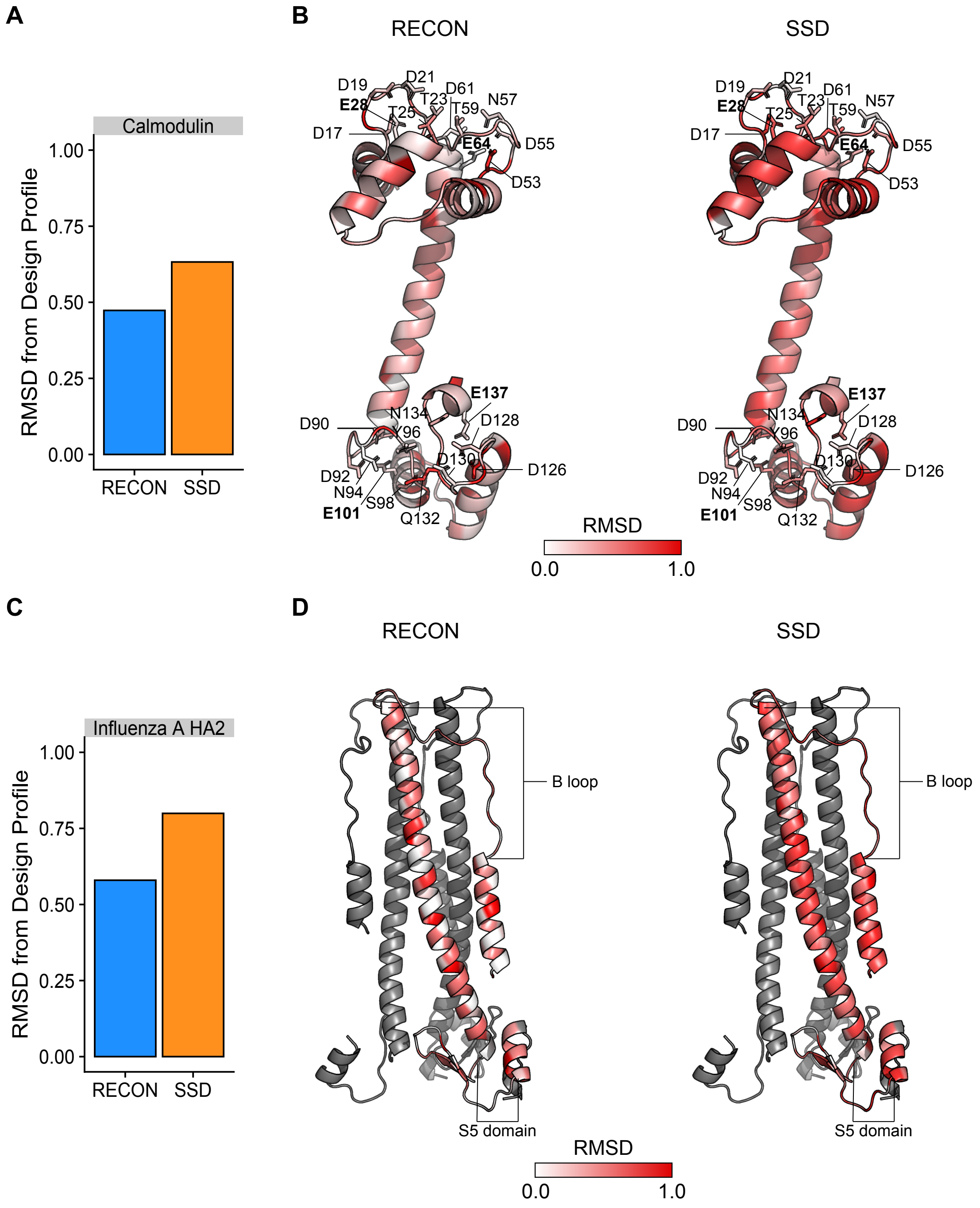


Figure II.5.: *(Continued on the following page.)*

Figure II.5.: Comparison of mutation profiles predicted by protein design to mutation profiles observed within calmodulin and influenza type A HA2 multiple sequence alignments. *(A) Comparison of root mean square deviation of mutation frequencies derived from calmodulin natural homologues to mutation profiles predicted by RECON MSD or SSD. Calmodulin natural homologue mutation preferences were derived from the multiple sequence alignment of calmodulin homologues. The root mean square deviation (RMSD) here represents the mean standard deviation of an individual residue's mutation profile, consisting as the mean sum of squared differences of all twenty amino acid frequencies as determined by the multiple sequence alignment of calmodulin homologue sequences in relation to either RECON MSD or SSD residue profile. (B) Residue profile standard deviations between calmodulin multiple sequence alignment profiles and design profiles mapped onto the unbound conformation of calmodulin (PDB ID 1CLL). Here, RMSD represents the mean sum of squared differences of all twenty amino acid frequencies of each residue between homologue and design profiles. Residues whose sequence profiles were predicted to have identical mutation profiles as that within the corresponding position with the multiple sequence alignment are colored in white. The greater the dissimilarity between the homologue mutation profile and design profile, the greater the saturation in red, with complete saturation indicating an RMSD of 1.0. Residues within all four of the conserved EF-hand motifs are labeled, with the bidentate ligand at position 12 critical for Ca^{2+} binding labeled in boldface. (C) Comparison of root mean square deviation of mutation frequencies derived from influenza type A sequence alignments to mutation profiles predicted by RECON MSD or SSD. RMSD is calculated in a similar fashion as in Panel A. (D) Residue profile standard deviations between HA2 multiple sequence alignment profiles and design profiles mapped onto the pre-fusion conformation of the HA2 trimer (PDB ID 2HMG). RMSD is calculated and labeled the same as in Panel B, but only one HA2 monomer is labeled with RMSD values of the influenza A IVR residue profiles in relation to RECON MSD or SSD profiles. The N- and C-terminal residues of loop regions that undergo large local conformational rearrangements in the post-fusion form are labeled. This includes the B loop that rearranges into an alpha helix and the S5 domain, which stabilizes the alpha helical form of the B loop. Residues within the CR8020 broadly neutralizing epitope (32), including N146 and E150, are also labeled.*

both protein design algorithms is too restrictive in sampling for residues at hinge points, given that the same dependency on $\mathrm{RMSD}_{\mathrm{da}}$ is not observed for PSI-BLAST sequence conservation. In contrast, both PSI-BLAST and RECON MSD had similar τ_β coefficients predicted with the same confidence, with $\tau_\beta = 0.0639$, $p = 1.79 \times 10^{-6}$ and $\tau_\beta = 0.0787$, $p = 1.19 \times 10^{-7}$ respectively, for the dependence of native sequence recovery on contact proximity deviation z-score, whereas SSD dis not exhibit the same dependence. This observation suggests that there is an evolutionary constraint on residues that are required to maintain a re-assortment of their local physicochemical environments necessary for a conformational change, and that RECON MSD closely models this evolutionary constraint by considering the multiple local side-chain environments within a protein ensemble.

II.2.7. Sequences suitable for conformational plasticity are energetically frustrated

The encouraged sequence convergence employed by the RECON MSD algorithm identifies amino acid sequences that have the lowest total energy across all states.[129] To examine the energetic impact of requiring a single amino acid sequence to adopt multiple states, we used a similar energy score term described previously as the sum total energy score normalized by the number of designed positions (see section II.5 on page 37). For RECON MSD designs, this approach would include lowest mean energy score of the designed ensemble, whereas the SSD energy score would include the lowest energy scores for each state. In all eight cases, RECON MSD selects sequences with a significantly higher energy score than SSD with a paired student's t-test,[161] with $p < 1 \times 10^{-4}$ (figure II.8 on page 33). We also compared the design energy scores to the ten lowest-energy relaxed structures, which only included the native sequences, and found that RECON MSD samples lower energy sequences relative the relaxed native

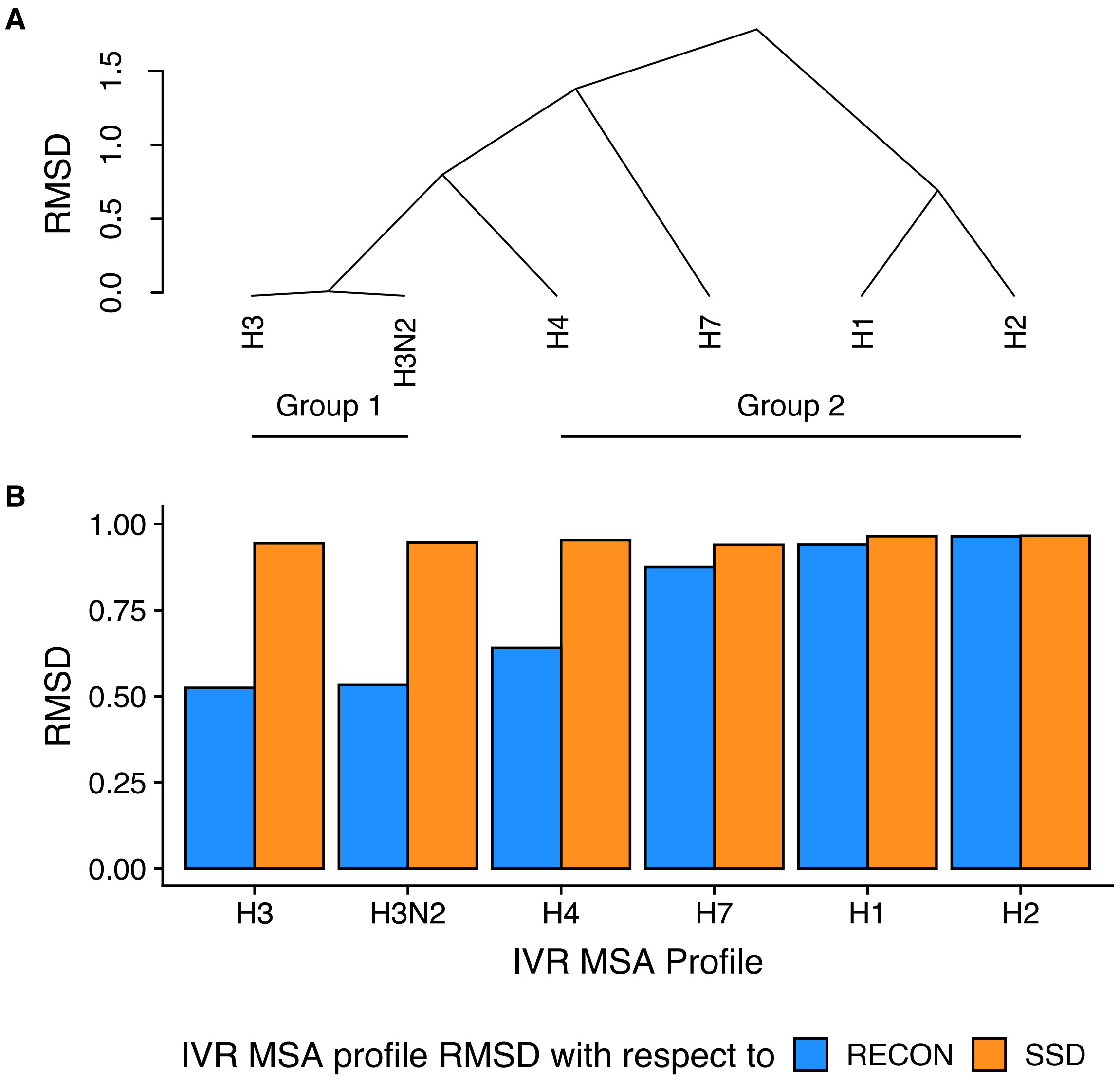


Figure II.6.: Root mean square deviations of residue mutation frequencies of influenza A subtypes and HA2 profiles predicted by RECON MSD and SSD.*(A) Dendrogram of root mean square deviations (RMSD) of influenza A subtype HA2 profiles sorted by pairwise RMSD. The mutation frequencies derived from the multiple sequence alignment profile of each influenza A subtype was compared to all other subtypes by calculating the mean standard deviation of each aligned position's mean sum of squared differences of all twenty amino acid frequencies with respect to each other subtype profile. Pairwise RMSD values were sorted to form clades, with the height along the y axis indicating the pairwise RMSD between each clade. (B) RMSD of each IVR subtype multiple sequence alignment (MSA) profile with respect to RECON MSD and SSD. The x axis represents each IVR subtype profile sorted as in Panel A. The y axis represents the RMSD, calculated in the same fashion as in Panel A, of each subtype profile in relation to either RECON MSD or SSD.*

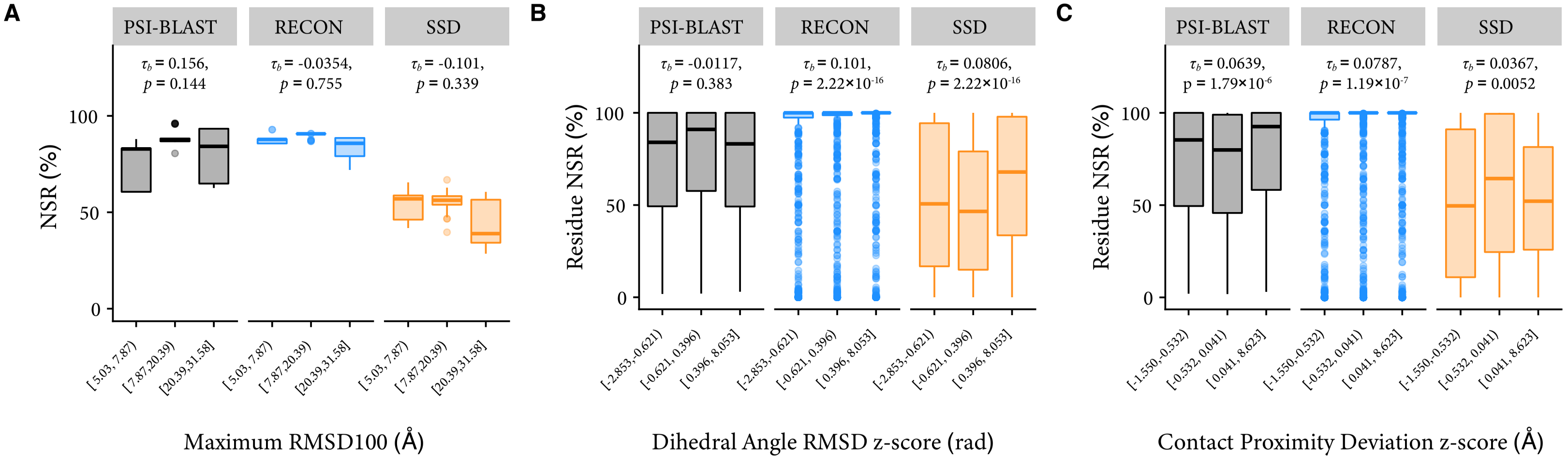


Figure II.7.: Relationship of conformational flexibility and native sequence recovery by sequence profiles. *The x axis is binned into three groups of equal number of data points to show the distribution of native sequence recovery between groups of low, middle, and high values for each metric. A Kendall τ_β rank correlation test was performed on each profile to measure the strength of dependence of native sequence recovery on each metric, indicated in each plot along with its associated p-value. (A) Comparison of native sequence recovery dependence on maximum RMSD100 between sequence profiles. (B) Comparison of native sequence recovery dependence on $RMSD_{da}$ between sequence profiles. $RMSD_{da}$ values of each protein were not equally distributed, nor were of similar range. Therefore, a z-score of was used to normalize $RMSD_{da}$ values of each protein to compare dihedral angle deviation scores, shown along the x axis. A similar approach was implemented to normalize contact map deviation scores. (C) Comparison of native sequence recovery dependence on contact deviation scores.*

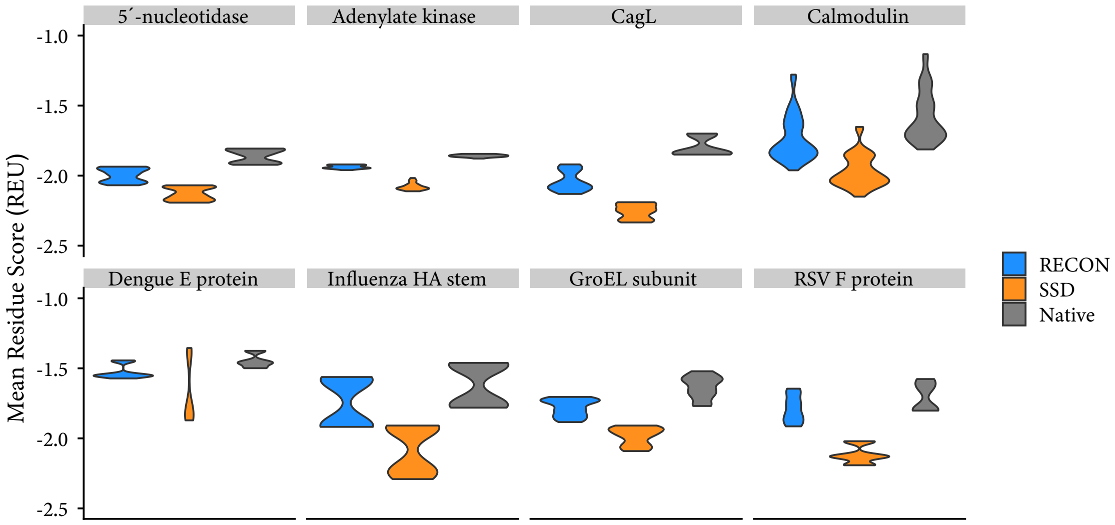


Figure II.8.: Average per-residue total energy score of the lowest ten percent scoring models for RECON MSD, SSD, and starting relaxed (Native) models. *One hundred simulations were performed for each group and the lowest ten total energy scoring models were used for the comparison. The total scores were normalized so that the calculated total score was divided by the number of residues within each model to obtain a mean residue score. For RECON MSD models, the total calculated score also had to be normalized by the number of states within each model. The violin plot width indicates the normalized energy score density of each group.*

structures. Given that RECON MSD conserves, on average, 88% of native sequences, the few mutations RECON MSD introduces to the native sequence are sufficient to sample a lower energy sequence space than the native sequence. In comparison, SSD sequences are the most stable as SSD replaces the native sequence at a much higher frequency, since SSD optimizes the sequence space for each conformation and can identify much lower energy sequences tolerable for each individual conformation. Therefore, given that RECON MSD is constrained in identifying sequences that are suitable to adopt multiple conformations and that the sequence space identified by RECON MSD is, on average, higher in total energy than the sequence space identified by SSD, it can be inferred that the sequence space available for proteins that populate multiple energy minima is more likely to be energetically frustrated, or at least not as energetically stable, than the sequence space available for a protein which populates a single energy minimum.

II.2.8. Stability decreases for residues with larger $C_\beta - C_\beta$ contact map deviations

We used a Kendall τ_β rank correlation test to analyze the dependency of the modeled sequence energy score on global and local conformational changes. For the comparison with global conformational changes, we compared the mean total score of the ten lowest-energy scoring design models, normalized by the number of residues within each protein, to the maximum RMSD100 of an ensemble. We found that there is a negative dependence of mean total score on the maximum RMSD100 for SSD models ($\tau_\beta = -0.143$, $p = 1.16 \times 10^{-5}$), but not so for RECON MSD models ($\tau_\beta = 0.0177$, $p = 0.586$; figure II.9 on the next page). Conversely, there was a small, but significant positive dependence of individual residue scores on contact proximity deviation z-scores for RECON MSD models ($\tau_\beta = 0.0356$,

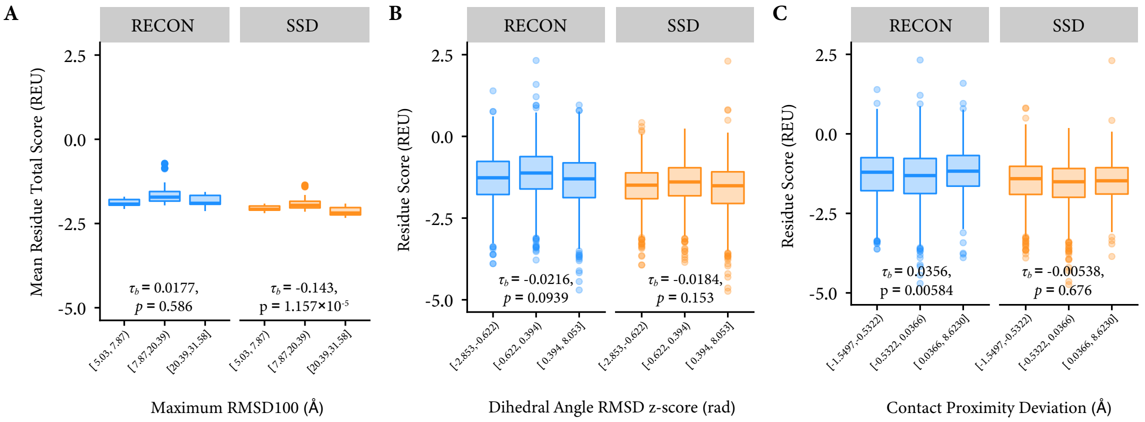


Figure II.9.: Comparison of conformational diversity and per-residue total scores. *All panels are binned into low, medium, and high x values, with equal number of data points for each bin. A Kendall τ_β rank correlation test was performed on each profile to measure the strength of dependence of native sequence recovery on the x axis value, indicated in each plot along with its associated p-value. (A) Comparison of maximum RMSD100 and mean total energy score, normalized by the number of residues. (B) Comparison of normalized RMSD$_{da}$ z-score and mean total energy score of each residue. (C) Comparison of normalized contact proximity deviation z-score and mean total energy score of each residue.*

p = 0.00584), but not so SSD models ($\tau_\beta = -0.00538$, p = 0.677). There was no dependence of individual residue scores on RMSD$_{\text{da}}$ for either design approach (figure II.9), which is surprising given that both native sequence recoveries for RECON MSD and SSD were strongly dependent on RMSD$_{\text{da}}$.

It should be noted that we found the metrices RMSD$_{\text{da}}$ and contact proximity deviation were not independent variables, as we determined that contact proximity deviation is significantly, although not strongly, negatively correlated with RMSD$_{\text{da}}$ (S5 Fig), meaning that residues with contact proximity deviation values close to or at zero were also more likely to have a higher RMSD$_{\text{da}}$ values. In either design case, residues with high RMSD$_{\text{da}}$ values tend to have lower scores, i.e. score favorably, suggesting that the native contacts and/or hydrogen bonding formed at positions restricted in rearranging side chain proximities for one or all states were more likely to score favorably, or at least more favorably than non-native side chains, by the Rosetta scoring function.[162] A favorable reference score would prevent the native residue in being redesigned with non-native residues, hence the correlation of high RMSD$_{\text{da}}$ values and higher native sequence recovery. Even so, the τ_β correlation of RMSD$_{\text{da}}$ and the designed sequences energy was not significant in either design case, such that the degree of backbone flexibility does not directly influence a residue's stability. These data suggest that RMSD$_{\text{da}}$ is restricted in optimizing the stability of residues that must rearrange their local side-chain environments, but not to the same extent in optimizing local backbone flexibility. SSD, on the other hand, is not restricted in optimizing side-chain placement within an ensemble, and therefore can select amino acid sequences that are more stabilizing for individual conformations.

II.3. Discussion

II.3.1. Contact proximity deviation captures local and global conformational rearrangements as a single metric

Methods like Local-Global Alignment and contact area differences are useful in circumventing the overestimation of global structural dissimilarity by either emphasizing local backbone segment structure similarity or by side chain placement similarity, respectively.[163,164] In particular, contact area differences between two homologues have been shown to be as accurate as RMSD, if not more, in comparing the structural similarity of proteins with very high sequence similarity.[165] However, contact area differences are sensitive to errors in side chain atom placement within a structural model, so that accounting for side chain mutations while measuring the relative position of equivalent residues is not feasible.

To overcome this limitation, we introduced the contact proximity deviation metric (see Methods section). Contact proximity quantifies the relative placement of a residue within a structure and is sensitive to a change in conformation without relying on side-chain contacts. To decide whether two residues are in contact, we analyze the $C_\beta - C_\beta$ distance. However, instead of a hard cutoff distance we use a smooth transition function to avoid discontinuities when distances change by small margins. Contact proximity deviation for a residue becomes the sum of changes in contact proximity when comparing two structures. Therefore, contact proximity deviation quantifies the magnitude of local rearrangements around a residue of interest inflicted by a global conformational change, independent of side chain identity.

Thus, combining the contact proximity deviation metric with interaction network analysis[166] could provide a useful tool to investigate how conformational rearrangements alter residue networks. Additionally, in cases such as the design of protein switches or structural analysis of mutations where protein flexibility, but not sequence, needs to be conserved, it is helpful to have a metric that highlights residues that experience rearrangements in their contacts. Lastly, the measurement of local structural variability by NMR RDC has shown that regions of high flexibility within ubiquitin ensembles align with multiple protein-protein interfaces.[167] Thus, contact proximity deviation provides a possible approach to study protein-protein interaction interfaces.

II.3.2. Sampling functional mutation preferences requires evaluation of sequence stability as an ensemble.

MSD approaches use the energetic contributions of multiple conformations to steer sequence selection and cull any sequences that do not improve the energy score of the designed ensemble as a whole.[168] For most approaches, sequences that do not improve all or the majority of conformations within an ensemble are culled, which is appropriate when the goal is to stabilize a protein ensemble within an energy minimum. However, protein function may not select for sequences that are limited to a single energy minimum. To better estimate the mutational preferences critical for function, it is necessary to use an approach that models local side-chain environments within the context of an ensemble.

Within figure II.4 on page 27, we illustrated that, although RECON MSD failed to accurately predict all amino acid exchangeability rates, the consideration of multiple local side chain environments during protein design improved the prediction of sequence conservation and overall accuracy of amino acid substitution frequencies as compared to modeling local side chain environments independently, particularly when modeling bulky side chains. In combination with figure II.8 on page 33 and figure II.9 on the previous page, we demonstrated that the consideration of local side chain stability within the context of an ensemble restricts stability optimization of the ensemble, especially for residues that require side-chain rearrangements during a conformational rearrangement. Taken together, we demonstrated that selection of mutation profiles by RECON MSD is much more similar to mutation rates observed in homologs if each mutation is evaluated across every conformation, or state, within an ensemble, and then culled if the mutation is evaluated to be destabilizing for an individual state within an ensemble. This approach does not necessarily select sequences that improve the stability of every state within an ensemble, but rather places the importance of modeling an ideal conformation-specific, local side-chain environment to prevent local side-chain destabilization within the context of an ensemble.

II.3.3. RECON MSD can be used to predict evolutionary sequence conservation of flexible proteins

Sequence similarity searches, such as PSI-BLAST, are fast and easy to use. Predicting mutation preferences from structure, especially if the sequence is known to form multiple conformations, remains to be a challenge. Advances in structure-based evolution design methods rely on iterative approaches that match sequence and structure similarities to predict sequence entropy.[169] For proteins that undergo conformational rearrangements, using this type of approach to search for structural similarity limits the sequence search space to similar conformations, possibly preventing the identification of sequences capable of adopting multiple conformations. Although other MSD methods have improved the selection of more evolutionarily similar sequences, they are limited in their capacity to simultaneously sample conformation and sequence space so that the relevance of conformational plasticity in evolutionary dynamics have not been fully interrogated.

The caveat to using RECON MSD to predict mutation preferences is accounting for Rosetta sampling biases. First, RECON MSD does not currently allow for the formation or destruction of disulfide bonds, which is critical for conformation stability, and does not accurately model the frequency of cysteine conservation. Consideration of alternate protonation states due changes in pH are also not explicitly modeled, which we see from our amino acid exchangeability comparisons that RECON MSD underrepresents exchangeability of polar residues and frequently mutates histidine to lysine or arginine, which has a pKa much higher than histidine or which is not charged. Additionally, in figure II.7 on page 32, we showed that RECON MSD is likely to overestimate sequence conservation of hinge regions that have large dihedral angle RMSDs. Even though we used a gentle minimization prior to design, minimization significantly increases the estimated stability of the native residue, making the replacement of the native amino acid unfavorable, as shown in S1 Fig. Given that residues located at hinge points within flexible loops are intrinsically disordered and typically contain less than ideal Ramachandran dihedral angles, it is likely that minimization specifically overcorrects these bond angles to fit the energy scoring function, preventing accurate sampling of rotamer placement. With the addition of explicit disulfide bond formation, use of a pKa-dependent rotamer library, and improvement of minimization prior to design, the RECON MSD algorithm could prove to be a valuable tool in predicting accurate mutation profiles.

With that being said, we used RECON MSD to demonstrate that sequence conservation and mutation preferences of a single sequence can be approximated using the evaluation of local residue physicochemical changes, provided that this one sequence folds into select, multiple conformations. In Fig 3, we showed that the estimated sequence conservation of RECON MSD designs differs by roughly 5% from the sequence conservation observed in PSI-BLAST profiles, with RECON MSD being more conservative. More specifically, we demonstrated that RECON MSD samples a very similar sequence space for hemagglutinin (HA_2) compared to what has been observed in H3 clade influenza subtypes (figure II.6 on page 31). Being able to predict the tolerated sequence space for viral antigens such as influenza HA has possible applications for antiviral drug design. It would have been preferred to compare the design profiles to deep sequencing data, as the represented mutation frequencies within the IVR database likely underestimate rare mutations. However, given the correspondence of the RECON MSD predicted HA_2 sequence profiles and IVR subtype-specific sequence profiles, it stands to reason that RECON MSD can serve an in silico approximation for costly deep sequencing, or at least serves as an initial screening for potential, more frequently observed mutations of drug targets, such as pathogens or oncoproteins.

The computational time required for the RECON MSD design simulations within this benchmark ranged from 2 h to 36 h. Compared to experimental approaches that have tested for functionally tolerated mutations in either dengue virus envelope protein or influenza hemagglutinin protein,[132,170,171] RECON MSD is much faster and less costly in identifying biologically relevant mutations. Additionally, RECON MSD is not limited to sampling mutations singly, pairwise, or as limited networks, but rather can sample mutations as an interaction network of each local side-chain environment. Traditional intra-protein co-evolution methods, such as direct coupling analysis,[172] mutual information,[173–175] or McLachlan-based substitution correlation methods,[176,177] are not reliable in detecting co-variation or correlation of mutation frequencies of highly conserved sequences,[178] and so they fail to detect contact dependencies of sequences with low sequence variation. In the case of this benchmark, we see that flexible sequences tend to be more highly conserved, especially when residues need to maintain distinct contacts between conformations. Therefore, current co-evolution methods cannot be used to detect residue contact dependencies of flexible, highly conserved sequences, whereas this benchmark suggests that RECON MSD is well-suited to identifying the evolutionary potential of a flexible sequence.

II.4. Conclusions

We demonstrated that RECON MSD significantly improves the similarity to evolutionary mutation preferences from SSD selected mutation profiles by selecting sequences which are energetically favorable for an ensemble of local side-chain interactions. Specifically, in instances where the goal of protein design is to preserve an ensemble of conformations for functionality, we suggest a greater emphasis on designing local physicochemical environments for each and all conformations within an ensemble, and to place less of an emphasis of finding sequences representing the most thermostabilizing for either each state individually or as an average of all states. Furthermore, the new conformational diversity metric contact proximity deviation we describe in this paper allows for the comparison of protein ensembles, assuming they are of similar length but not sequence, by quantifying position-specific relocation due to one or more conformational changes. Therefore, in conjunction with contact proximity deviation, RECON MSD warrants further use as a bioinformatic tool to estimate mutation preferences of homologous proteins, especially for proteins known to undergo similar domain or fold reorganization between conformations.

II.5. Methods

II.5.1. Selection and preparation of benchmark datasets

Our criteria for benchmark datasets included proteins that had at least two published conformations with greater than 5 Å RMSD and at least one peptide chain greater than 100 residues in length. To identify these proteins, we performed a BLAST search to identify proteins with 100% sequence identity and with gaps of three or less residues in length. Structures with similar backbone conformations of less than 0.5 Å RMSD were excluded from design so that the structure with the longest matching consecutive sequence was kept as the template structure.

Structures were downloaded from the Protein Data Bank (PDB; www.rcsb.org) and processed manually to remove all atoms other than the residue atoms intended for design. Any residues that did not align or positions that were not present in all template structures were not considered for design and were removed from the template. For a detailed description of which residues were included for design, see S1 Table. Native structures were subject to minimization and repacking in Rosetta using FastRelax constrained to start coordinates with the talaris2013 score function placing a backbone movement constraint on all C_{α} atoms of 0.5 Å standard deviations to prevent substantial movement away from the native structure.[162,179] The lowest total energy score model was selected from the 100 relaxed models for design. For comparisons using un-relaxed models, the native structure was used instead of the relaxed model.

II.5.2. RECON multi-state and single-state design

Benchmarking using RECON MSD design was performed using four rounds of fixed backbone design and a convergence step using the greedy selection algorithm, as previously described (12, 13), with the exception that only repacking, and not backbone minimization was allowed following the convergence step to prevent over-optimization of the energy score following design. For parallelized production runs of multistate design, each state within an ensemble was handled on its own processor, requiring up to 32GB of RAM per node for the largest design case, RSV F trimer. Similarly, single-state design was performed using four rounds of fixed backbone rotamer optimization followed by repacking using the identical designable residues as specified for RECON MSD designs. The talaris2013 scoring function was used for both RECON MSD and SSDs.[160] One hundred designs were generated for each benchmark structure using either RECON MSD or SSD.

II.5.3. Generation of sequence profiles

The lowest ten out of a hundred scoring models were used for quantification of sequence tolerance. In the case of SSD, the ten lowest total scoring models were used from the design simulation of each PDB structure and then were grouped by protein to form an ensemble containing $10 \times N$ models, with N being the number of conformations within an ensemble. For RECON MSD, the total score of each model designed within an ensemble design run was averaged with all other conformations modeled during the same RECON MSD run to create a fitness score, which then was sorted to identify the ten designed ensembles with the lowest fitness score, again containing $10 \times N$ models for each ensemble. A Shannon entropy bitscore[58] was calculated using WebLogo for each designed position within an ensemble as $I_i = p_i \times log_2(20 \times p_i)$, with i as the amino acid and p_i as the frequency of that amino acid. Here, the calculated amino acid frequency includes the frequency at the same position within the ten lowest-scoring models of all designed states, whether designed independently by SSD or designed simultaneously by RECON MSD, such that an amino acid represented a 100% of the time at a particular position in all states has a bitscore of 4.32.[180] The frequency of each possible mutation, *i.e.* all twenty amino acid frequencies recovered from design at each position, was calculated from bitscores of each position to generate a $20 \times n$ matrix, with n being the number of designed residues within each ensemble.

PSI-BLAST profiles were obtained by querying a non-redundant protein database using default parameters, increasing the number of iterations to ten iterations, as well as querying the database with e-value thresholds ranging from 1×10^{-5} to 1×10^{2}. We reported only PSI-BLAST profiles generated using default parameters, which includes two iterations and an e-value of 0.005. PSI-BLAST profiles using non-default parameters were qualitatively identical to the PSI-BLAST profiles generated using the default parameters, and were therefore not reported. We omitted any sequences within the queried sequence profiles which were not included for design to generate a $20 \times n$ matrix corresponding to each benchmark case containing the amino acid frequencies obtained from the position specific-scoring matrix (PSSM) as described in the previous paragraph.

II.5.4. Methods for comparison of sequence profiles

We compared the PSSM generated from the PSI-BLAST query to the PSSM generated by either RECON MSD or SSD by calculating the percentage of native sequence recovery, which was determined as the sum of the bitscores of the native amino acids at each position divided by the sum of the information bitscore of all amino acids at all positions.[181] Additionally, we calculated the average total variance in designed position mutation frequencies obtained from the bitscores of each position with the PSSM, as described in figure II.3 on page 25:

$$y = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{20} (aa_{PSI-BLAST} - aa_{design})^2}{n}}$$

where aa_j represents the frequency of an amino acid observed at position i for each of all twenty amino acids (j), and y is the sum of all i differences for all amino acids within a protein with a length of n residues.[182] Average amino acid substitution rates were determined as the mean cumulative substitution frequency of each amino acid i to amino acid j, where

$$\overline{aa_{ij}} = \begin{bmatrix} \frac{1}{aa_i} \sum aa_{11} & \dots & \frac{1}{aa_{1j}} \sum aa_{1j} \\ \vdots & \ddots & \vdots \\ \frac{1}{aa_{i1}} \sum aa_{i1} & \dots & \frac{1}{aa_{ij}} \sum aa_{ij} \end{bmatrix}$$

We define amino acid exchangeability as the subset of average amino acid substation rates that exclude the substitution rates of $i \rightarrow j$, where j is identical to i, or in other words, all substitution rates that include the average conservation frequencies of the native amino acid.[183] The mean amino acid exchangeability rates, as shown in Panel C in figure II.4 on page 27, were calculated as

$$\overline{aa_{ij}} = \frac{1}{19} \sum_{j=1}^{19} \frac{1}{aa_i} \sum aa_{i,j}$$

where the mean exchangeability rate of each amino acid i is the mean of all exchangeability rates of amino acid i to amino acid j, excluding the conservation rate of amino acid i. Kendall τ_β rank comparison tests, linear regression, Wilcox comparison of means, and student t tests were performed in R. Levene's test for equality of variance was performed using Python 3.7 using the scipy.stats package using the median as the center.

II.5.5. Preparation of calmodulin and influenza type A HA2 mutation profiles and consensus sequences

The entire influenza type A HA sequences were obtained from the IVR using the following parameters – type: A, Host: any, Country/Region: any, Protein: HA, Subtype H: any, Subtype N: any. Subdivision of the entire influenza type A HA sequences were obtained by changing the parameter Subtype H to 1, 2, 3, 4, or 7 with Subtype N as any, or for the H3N2 subtype, Subtype H: 3 and Subtype N:2. The number sequences obtained from each query are listed in Table 2. Sequences within the calmodulin dataset containing non-redundant calmodulin sequences across eukaryotes obtained from the Supplementary Material[153] or influenza type A full-length HA sequences obtained from the Influenza Virus Resource (25) were aligned using a locally installed Clustal Omega version 1.2.4.[184,185] Given the limited number of sequence gaps and high sequence conservation, we believe the multiple sequence alignments performed on either dataset were accurate. The consensus sequence was determined from the multiple sequence alignment using a locally installed EMBOSS 6.4.0.0 package cons using default parameters. The frequency of each amino acid type at each aligned position was determined using WebLogo[180] using the following parameters – sequence-type: protein, format: logodata, composition: none.

II.5.6. Description of conformational metrics used in this benchmark

We use the maximum RMSD within an ensemble to represent the largest amplitude of dissimilarity within an ensemble, defined as:

$$RMSD_{max} = max_s \sqrt{\tfrac{1}{n} \textstyle\sum_{i=1}^{n} \|v_i - w_i\|^2}$$

where n represents the number of residues, s represents the number of structures within an ensemble, and $\sqrt{\frac{1}{n} \sum_{i=1}^{n} \|v_i - w_i\|^2}$ represents each pairwise RMSD within an ensemble (Panel A in figure II.2 on page 22).[148] For the local backbone dissimilarity metric, we use dihedral angle RMSD to describe the deviation of each equivalent dihedral angle, or pair ϕ and ψ angles, within an ensemble containing s structures, as

$$RMSD_{dihedral} = \sqrt{\tfrac{1}{s} \textstyle\sum_{i=1}^{s} \|\phi_s - \bar{\phi}\|^2 + \tfrac{1}{s} \sum_{i=1}^{s} \|\psi_s - \bar{\psi}\|^2}$$

where $\bar{\phi}$ and $\bar{\psi}$ represent the mean ϕ and ψ angle of each equivalent residue. The contact map dissimilarity metric we introduce here is based off the neighbor count weight metric,[186] which scores the likelihood of a neighboring contact by assigning each $C_\beta - C_\beta$ distance a score, which we term contact proximity as

$$cp = \begin{cases} 1 & \text{if } d \leq bound_{low} \\ \frac{1}{2} \cos \frac{d - bound_{low}}{bound_{high} - bound_{low}} & \text{if } bound_{low} < d < bound_{high} \\ 0 & \text{if } d \geq bound_{high} \end{cases}$$

For glycines, a pseudo-C_β atom is defined using the amide N, C_α, and carboxyl C atom coordinates in the PDBtools package (github.com/harmslab/pdbtools) before calculating $C_\beta - C_\beta$ distances. The lower and upper bounds represent thresholds where a $C_\beta - C_\beta$ distance certainly does and does not contain any side-chain atoms that are in contact with another residue's side-chain atom. We define the lower bound, $bound_{low}$, as 4.0 Å and the upper bound, $bound_{high}$ 12.8 Å, where the lower bound was determined to be a reliable threshold to define solvent-inaccessible side-chains due to side-chain contacts, or in other words, a $C_\beta - C_\beta$ distance less than the lower bound is very likely to form at least one side chain interaction.[186] The upper bound was determined by the maximum $C_\beta - C_\beta$ distance where at least one atom from each side chain formed an interaction.[187] Finally, the contact proximity deviation for each residue was calculated as the sum of all $C_\beta - C_\beta$ contact proximity score deviations for that residue (Panel D in figure II.2 on page 22). With this metric, we can quantify the changes in side chain local environments that are not due to local hinge-points, but instead, show local side chain environment changes that are due to larger conformational rearrangements.

II.6. Acknowledgements

The authors would like to thank Amanda Duran, Axel Fischer, and Diego del Alamo for their assistance and helpful discussions during the development of this project. Calculations for the contact map deviations were made possible by members of the Michael Harms Laboratory as contributors to the PDBtools opensource software (github.com/harmslab/pdbtools). We also thank the Vanderbilt University Advanced Computing Center for Research and Education for the computational resources necessary to complete this project. All work was funded by the U19 AI117905 grant for Structure Based Design of Antibodies.

CHAPTER III

PREDICTING SITES OF VULNERABILITY WITH LOCAL SITES OF CONFORMATIONAL CHANGE

This chapter is based on unpublished work. Marion F. Sauer conducted the initial experiments, analysis, and writing the article.

Conformational B-cell epitope prediction methods traditionally use amino acid properties, such as hydrophobicity or solvent-accessible surface area, to describe propensities of residues to be located within an epitope. For meta-stable viral fusion glycoproteins, sites of vulnerability often overlap with sites of conformational rearrangements. Provided that viral fusion is entropy-driven and requires changes in stability to propogate the large-scale conformational changes necessary for membrane fusion, one possible mechanism to block viral fusion protein function by targeting residues that must undergo local changes in stability and conformation to overcome the entropy barriers blocking these large-scale conformational rearrangements. This chapter provides preliminary work using the residue descriptors contact proximity deviation, introduced in chapter II on page 16, and total Rosetta *energy score as predictors of conformational B-cell epitopes. Moreover, traditional conformational epitope prediction methods do not attempt to classify the predicted epitope residues into distinct clusters, or rather specific minimal epitopes, whereas the latter part of this chapter discusses one potential method to achieve this goal and the potential shortcomings of defining minimal epitopes.*

III.1. Introduction

When a specific protein surface of an invading pathogen, or antigenic determinant, is recognized via secreted antibodies or B-cell receptors to elicit a humoral immune response, this antigenic determinant is termed a B-cell epitope. Although B cell epitopes can consist of either a linear peptide sequence or a three-dimensional protein surface, recognition of 90% of B-cell epitopes are thought to be conformation-specific.[52] Therefore, B-cell epitopes are typically defined by a discrete cluster of amino acids with a surface area of 600 Å to 1000 Å that form a binding interface with the paratope (Sela-Culang2013). Correct identification of conformational B-cell epitopes within an antigenic protein is paramount for the design of molecules, such as subunit vaccines, that imitate the binding surface of antigenic determinants to raise specific antibodies and can be used as prophylatic or therapeutic vaccines. Existing epitope mapping methods, such as X-ray crystallography, cryo-EM, HDX, and competition assays have provided detailed biophysical descriptions of commonly targeted epitopes, but are time-intensive, laborious, and expensive. Given the intrinsic mutability of re-emerging infectious diseases due to sequence mutations and possible structural variation, experimental determination of B-cell epitopes are slow in responding to changes in mechanisms of epitope recognition.

Computational methods for the prediction of what defines a B-cell epitope could provide a useful starting point for further experimental validation to accelerate the discovery of either commonly conserved or novel epitope targets. The earliest efforts to predict conformational epitopes relied on using amino acid sequence properties, especially hydrophobicity/hydrophilicity and solvent-accessible surface area (SASA), to describe epitope propensities, but reached a success rate of less than 75% accurate predicition of epitopes.[188–191] Even the most recent conformational epitope prediction algorithms, which employ Kmer or random forest classification to distinguish epitopes from non-epitopes, rely on similar amino acid physicochemical features used thirty years ago as part of their epitope predictors.[192,193] As such, the maximum success rate in conformational B-cell prediction was demonstrated by the EPITOPIA server, which reached a success rate of 80.6%.[194]

One limitation of B-cell epitope prediction is distinguishing between false from true positives and negatives. Experimental epitope mapping techniques can yield contradicting boundaries or inexact locations of conformational epitopes, which decreases the sensitivity of classifier predictions.[193,195] Although collections of structural epitope and non-epitope data has increased significantly in the past several years with the creation of structural epitope databases such as IEDB-3D, a component of the Immune Epitope Databank (IEDB),[196,197] structural epitope data typically include short continuous peptide motifs. However, of the 20 – 25 amino acids within a structural epitope, hotspots of only 2 – 5 discontiniguous residues contribute to the majority of relative free energy necessary for binding to a paratope.[84,86,198] Therefore, the resolution of which residues are critical for binding cannot be distinguised from nearby residues unless experimentally validated.

Although an epitope must be exposed on the surface of an antigen to form a paratope-epitope interaction, not all residues that form an epitope are surface-accessible within all conformations of a highly flexible protein. For instance, many viral fusion glycoproteins, such as influenza HA, HIV Env, or DV E proteins, are known to undergo substantial conformational rearrangments that facilitate the attachment and fusion of a virion to the host cell. These viral glycoproteins also are common targets of bnAbs, but not all epitopes persist as antigenic determinants when these glycoproteins assume distinct conformations during the fusion process. RSV F Site Ø is surface-accessible when RSV F protein remains in its pre-fusion conformation, but vanishes during the reorganization of the RSV F protein into its post-fusion state.[68] Similarly, recent epitope mapping discoveries have identified that the transient breathing of the metastable, pre-fusion state of RSV F, HIV Env, and influenza HA allows for the temporary exposure of occluded epitopes.[98,116–120,199] Even if the residues that constitute an epitope are surface-accessible for multiple conformations, antibody recognition, *i.e.* binding affinity, has been shown to be specific to the available conformational epitope on the RSV F protein.[6]

Viral glycoprotein sites of vulnerability, or epitopes known to interact with neutralizing or protective antibodies, tend to overlap with regions known to undergo local conformational rearrangments. Influenza H3 and H1 HA bnAbs CR8020 and CR6261, respectively, target epitopes that either undergo local backbone torsion angle deviations or relative transformation in relation to the viral membrane.[88,89] HIV bnAbs such as 2F5, PGT145, and Fab17b target common sites of vulnerability within the Env trimer, including the MPER, CD4 receptor, and the trimer apex containing the V1/V2 and V3 regions.[116,199,200] Each site of vulnerability is known to exhibit conformational heterogeneity upon either priming, induced by CD4 binding, or triggering to the open, post-fusion conformation as induced by co-receptor binding. Although viral glycoproteins follow a diverse set of mechanisms to achieve fusion, each viral glycoprotein's inherent flexibility is necessary to facilitate the entropy-driven process of viral fusion. Therefore, as with the aforementioned bnAbs, it appears that one common mechanism of neutralization and/or protection is by stabilizing local conformational change and preventing the conformational rearrangments necessary for fusion to occur.

By extension, prediction of epitopes could be improved by the consideration of conformational changes exhibited by an antigenic determinant. So far, only single amino acid properties such as sequence conservation, SASA, electrostatic potential, and in some cases amino acid side chain propensity for rotameric flexibility have been used as predictors of antigenic determinants of viral glycoproteins. Such properties are agnostic to the global conformational changes required for function and provide little insight into mechanisms of protection and/or neutralization. However, given that goal of B-cell epitope prediction is to distinguish which residues are or are not within an epitope, any predictor variable that quantifies conformational change must be descriptive of an individual residue. Residue flexibility can be quantified a number of ways — either by B-factor, RMSD_{da}, or changes in residue contacts. A residue's B-factor describes the temperature-dependent static disorder of a crystal lattice,[201] whereas RMSD_{da} represents the mean torsional displacement an individual peptide bond undergoes within a protein ensemble. Both of these metrics, however, do not relate local residue flexibility to larger-scale residue displacements, whereas quantifying residue contact neighbor changes have shown to be a good indicator of local peptide backbone dynamics and relative displacement due to larger conformational rearrangments.[130,202]

For each conformation an antigenic determinant assumes, the specific protein fold is associated with a relative Gibbs free energy, or ΔG value, approximated as its sum of all hydrophobic, hydrogen bond, van der Waals, and ionic bond interactions that all of its substituents, that is, the amino acid side chains create.[203] Conformational rearrangments require a reconfiguration of the protein scaffold and/or side chains, so that the ΔG contributions of each amino acid are likely distinct between each conformation. As mentioned previously, one possible protection mechanism is through stabilization of locally flexible residues. Therefore, the conformation-dependent estimated ΔG, in addition to the quantification of contact neighbor changes, of each residue is necessary to identify unstable, flexible residues that would be prime paratope-binding targets. The main motivation of this paper is to determine if a residue's stability and flexibility can distinguish between not only residues that are or are not known epitopes, but also the conformations in which a set of residues may form potential hotspots within a conformational B-cell epitope.

III.2. Results

Validation of B-cell epitope prediction by quantifying residue contact neighbor changes and residue stability raises three questions: 1) Do residues within epitopes have distinct patterns of contact neighbor changes and/or relative free energy? 2) Are these patterns conformation specific? 3) How are epitope boundaries defined? The first two questions are readily addressed by comparing residues' contact proximity deviation and Rosetta total energy scores calculated for each determined conformation to known epitope locations. The last question can be addressed by considering the surface area of an epitope and making the assumption that an epitope is one that includes more than one hotspot residues, and ideally includes a constellation of hotspot residues within close proximity. In other words, clustering by not only residue hotspot location, but also the likelihood that a residue is a hotspot, may be sufficient to determine which residues form a minimal conformational epitope. This likelihood is determined by its predictor attributes, *e.g.* contact proximity deviation or Rosetta total score.

The results are presented in correspondence to the above questions. For the purpose of this dissertation, the results introduce only preliminary data with the aim that this chapter will be expanded upon and submitted for publication. Future experiments and analysis necessary to determine what predictors best describe epitope residue attributes and minimal epitopes sufficient for antibody binding are discussed in section III.3 on page 47.

III.2.1. Conformation-dependent epitopes are more likely to include flexible, unstable residues

The contact proximity deviation metric introduced in chapter II on page 16 was shown to approximate a residue's conformational flexibility in relation to the global conformational rearrangements of the protein ensemble, provided that there are at least two determined structural models of differing conformations. Therefore, in conjunction with evaluating the relative energy of each residue using the Rosetta energy scoring junction,[204] the contact proximity deviation and Rosetta total score were used to approximate each residue's flexibility and stability within the DV E and RSV F proteins. Given that the motivation of this chapter is to describe a generalized method to discern epitope from non-epitope residues, the determined scores of all residues evaluated either by contact proximity deviation or total score were normalized by z-score for further analysis. Next, each residue was identified as an epitope or not based on experimental determination of known DV E and RSV F epitopes from linear peptide data within the IEDB[205] and from structurally discontinuous epitope maps.[206,207] A Pearson correlation coefficient[208] was calculated for residues that were or were not found to be within an epitope to test for a linear correlation of a residue's normalized contact proximity deviation and total energy score. The normalized total score used was specific to either the pre-fusion or post-fusion conformation, so that two Pearson correlation coefficients were computed specific to each conformation. Comparison of the Pearson coefficients showed that only residues known to be located within epitopes were significantly more likely to have high normalized contact proximity deviation scores and high normalized Rosetta total energy scores, but only for the pre-fusion conformation total scores, as shown in figure III.1 on the next page.

Although the correlation coefficients did not indicate a strong correlation ($-0.75 < R < 0.75$) even though significant for either RSV F or DV E pre-fusion total score and contact proximity deviations, the quotient of the Fisher's z-transformed R values[209] indicated that RSV F residues classified as epitopes were 3.36 times more likely to be an epitope as contact proximity deviation and total score increased. DV E residues classified as epitopes were 16.0 more likely to be an epitope with increased contact proximity deviation and increased pre-fusion-specific residue total score. However, multiple linear regression analysis indicated that the coefficient of determination using normalized contact proximity deviation and total Rosetta pre-fusion energy scores was 0.221 when evaluated for only DV E and RSV F.

III.2.2. Epitopes cannot be identified by local unstable and flexible hotspots alone

As mentioned in the introduction, the "resolution" of epitope mapping does not always possess the capability to distinguish between individual residues that are critical for antibody binding from surrounding local residues, which limits the predictive power computational methods to identify B-cell epitopes. Furthermore, epitope prediction methods are practical only when they have the power to distinguish not only residues that might or might not be epitopes, but also which set of residues will contribute to a broadly-neutralizing antibody response. As shown in section **??** on page ??, the strength of prediction using regression analysis alone was low. In part this is due to the false positive classification of residues as epitopes, which likely includes those resides with low normalized contact proximity deviation and/or Rosetta total scores. Another factor, which is critical for the formation of an antibody-antigen interaction, is the spatial orientation of each classified epitope residue in relation to each other. Distinct epitopes, such as RSV F Site IIa and Site IIb, have been shown to be immediately adjacent to each other[108,207] so that the definition of an epitope cannot be easily defined by hotspots within some minimum distance of each other.

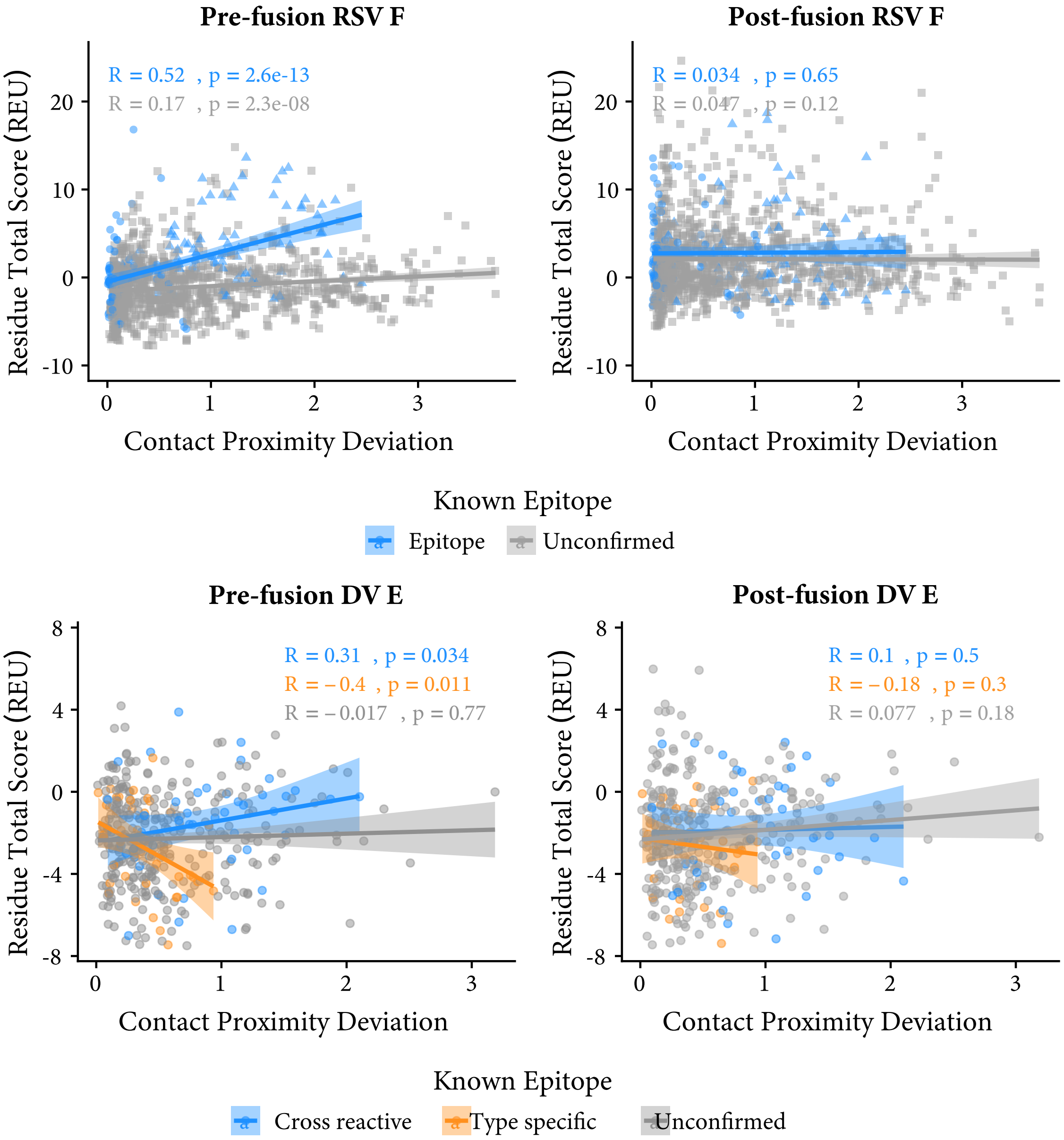


Figure III.1.: Correlation comparisons of epitope residues versus non-epitope residues using normalized contact proximity deviations and ROSETTA total score. *(A) Correlation comparisons of RSV F residues known to form a binding interface with neutralizing antibodies (shown in blue) and residues not yet determined to be within an epitope (shown in grey) within the pre-fusion and post-fusion conformations. The structures used to evaluate contact proximity deviation included two structurally determined pre-fusion conformations (Protein Databank (PDB) ID 4MMS and 4ZYP) and two post-fusion conformations (PDB ID 3RKI and 3RRR). Total ROSETTA energy scores were evaluated using unminimized models of 3RKI and 4MMS. (B) Correlation comparisons of DV E residues known to form a binding interface with bnAb (which target multiple DV serotypes (shown in blue) or serotype-specific antibodies (shown in orange) versus non-epitope residues (grey) for the DV E pre-fusion conformation at neutral pH and the post-fusion conformation. The structures used to evaluate contact proximity deviation scores include PDB models immature, uncleaved 1OAN; immature, cleaved 3C5X determined at pH 2.0, immature, cleaved 3C6E determined at pH 7.0, and post-fusion 1OK8. Total ROSETTA energy scores were evaluated using unminimized models of 3C6E and 1OK8.*

Therefore, in addition to identifying which residues may be epitopes, it also necessary to identify which groups of residues form a minimal epitope, which is likely best achieved through the use of unsupervised clustering methods to avoid overfitting. The spatial distance of each residue needs to be preserved in the clustering approach so that creating a tensor including a residue's cartesian coordinates and other epitope predictor values will distort the spatial information necessary to identify epitopes. In other words, a set of predictors must first be used to cull any residues which are unlikely to be found within an epitope. Next, the remaining residues should be clustered to identify constellations of residues which form a minimal epitope. Commonly used unsupervised clustering methods, such as k-means clustering[210] or DBSCAN,[211] however, make assumptions about either the number of points (in this case residues) or the density of residues necessary to form an epitope. The Girvan-Newman algorithm,[212] on the other hand, makes no assumptions about the relative size of a cluster, but instead clusters by hierarchy of connectiveness within the network of points. Given that a epitope typically has a surface area of 800 Å^2, the connectivity of each epitope residue is defined by its possibility of being within the same epitope interface as another residue, which requires the distance to any other epitope residue be equal to or less than the square root of a typical epitope surface area, or approximately 16 Å. The advantage of the Girvan-Newman algorithm is that it has been shown to appropriately cluster community networks that are spatially close, but distinct in their interaction networks.[212] Like RSV F Site IIa and IIb, it is critical that spatially close, but antigenically distinct, be appropriately clustered.

The Girvan-Newman algorithm was used to cluster residues that were identified as epitope residues. This was achieved by 1) Calculating the sum of normalized contact proximity deviation and normalized total Rosetta energy score; 2) Any residues that scored above a certain percentage of all score sums were defined as a epitope residue; 3) Connectivity of the epitope network was defined by defining an edge as any $C_\alpha - C_\alpha$ distance equal or less than 16 Å between any two residues defined as an epitope; 4) The Girvan-Newman algorithm was applied to identify any residues within the same community cluster; 5) The sum of all sums of the normalized contact proximity and energy score for each residue within a community were used to test for any difference in relative flexibility and stability between identified epitopes, even when compared across conformation-specific z-scores (figure III.2 on the next page). The enrichment for properly assigned epitopes within each cluster was slightly higher the multiple linear regression coefficient of determination 0.221 for the pre-fusion specific community clusters of either DV E residues as shown in figure III.2 on the following page, suggesting that the elimination of low contact proximity deviation and total energy scoring residues enriched for positive identification of epitope residues.

III.3. Discussion

The approaches discussed in section III.2 on page 44 need to be expanded to other viral glycoproteins before reaching more conclusive results. Additionally, the parameters of contact proximity deviation and total Rosetta energy scores are not sufficient to predict B-cell epitopes alone. Based on the location of clusters and the lack of any clearly defined community clusters as shown in figure C.1 on page 115 and figure C.2 on page 116, one immediately obvious descriptor that was excluded was surface accessibility, which can be more readily approximated using the contact neighbor vector metric[186] that precise SASA calculations. Additionally, the approach suggested in section III.2.2 on page 45 would need include previously defined descriptors to compare the success rate of epitope residue enrichment. The contributions of each descriptor should likely be optimized by multiple linear regression, although this

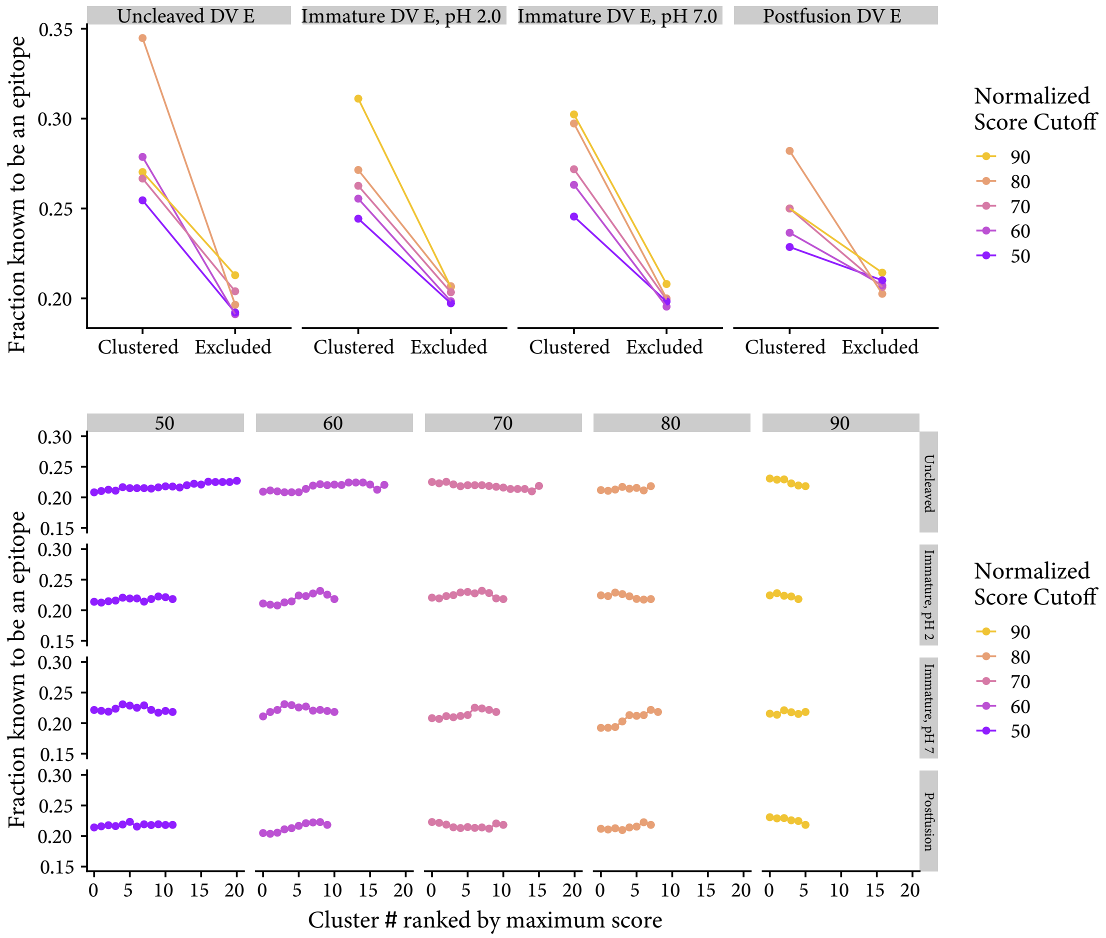


Figure III.2.: Clustering enrichment and ranking of predicted epitopes. *The clustering of predicted epitope residues was performed by taking a certain threshold, or percentage, or residues with a sum of normalized contact proximity deviation and total* Rosetta *energy scores (with energy scores determined using each conformation indicated in the top grey panels), and are indicated by the purple to gold coloring. The top panel depicts the number of residues identified (Clustered) or not identified (Excluded) as epitopes by the clustering approach described in the last paragraph of section III.2.2 on page 45. The height of each point indicates the percentage of residues within either the Clustered or Excluded residues that have been experimentally determined to form an antibody-antigen interface, which also included linear peptide epitopes. The bottom panel depicts the normalized sum of contact proximity deviation and total, conformation-specific* Rosetta *energy score for each identified community cluster and the number of clusters identified using the range of thresholds to exclude non-epitope residues from community detection.*

may overfit the descriptor weights given the limited number of glycoproteins determined in their pre- and post-fusion conformations.

The definition of a minimal epitope described here assumes that the spatial distribution of hotspots alone is sufficient to detect unique clusters, which is achieved through the use of the Girvan-Newman community detection hierarchal clustering. This particular clustering method may not be ideal, and the particular clusters defined by the Girvan-Newman clustering algorithm should be compared to clusters defined by other means, including DBSCAN. Moreover, the estimation of relative flexibility and stability of hotspot residues within each epitope is insufficient to estimate the relative flexibility and stability of a site of vulnerability, as it does not account for the estimated flexibility and stability of surrounding residues which are necessary to form a minimal epitope. One possible way to account for contact proximity deviation and energy scores of any surface-accessible residues within some distance-bound region between the collection of hotspot residues which define a minimal epitope. This approach is very crude, however, and the goal of community detection is to identify critical residues within a single binding interface that, when bound by an antibody, is sufficient to confer either neutralization or protection by blocking the conformational rearrangements necessary for fusion.

III.4. Acknowledgements

I would like to thank Diego del Alamo for suggesting to use the Girvan-Newman algorithm and Dr. Axel Fischer for his help with multiple linear regression analysis. All work was funded by the U19 AI117905 grant for Structure Based Design of Antibodies.

CHAPTER IV

CONCLUSIONS

The body of this dissertation primarily focuses on computational methods that can be applied to determine the dependency of mutation preferences and the formation of sites of vulnerability due to changes in viral fusion glycoprotein conformation during attachment and fusion. Given that not all chapters specifically focus on viral fusion proteins, this chapter presents the results pertinent to viral fusion glycoproteins and, when relevent, to the general understanding of structural biology. Additionally, this chapter discusses the limits of these conclusions and what measures might be taken to overcome, or at least address, some of these limitations.

IV.1. Simulation of mutational tolerance is improved by replicating the physicochemical constraints within a protein ensemble

Given the dogma of structural biology — sequence determines structure, it seems almost a tautology to say that mutation preferences are determined by the protein backbone ensemble a single sequence must assume. That said, validating this principle with computational protein design methods was previously intractable for larger proteins, given that the combined sequence and conformation search space required for exhausive sampling would require the combined sampling of all rotamer conformations on all possible protein backbone conformations. The RECON multi-state design (MSD) method as described in chapter II on page 16 eliminates part of the sampling complexity by assuming that each conformation has its own low-energy sequence. From the set of lowest-energy sequences of each conformation, each designed "mutation" is evaluated using the same Rosetta energy score function on all other backbone conformations to select the mutation that is lower in evaluated energy score than the rest of the selected mutations at each position.[129] With the inclusion of energy score weights that encourage the selection of the starting, or native, sequence, the RECON MSD algorithm greatly reduces the sequence and conformational search space. In doing so, the RECON MSD algorith is one of the first protein design algorithms used to identify the mutational tolerances of large, flexible proteins.

The results in Sections II.2.3 - II.2.5 indicate that this approach improves the prediction of mutational tolerance of highly conserved proteins, which were conserved 82.11 ± 11.2% in all cases, as opposed to the more traditional single-state design (SSD) approach. The accuracy of sequence mutation profiles however was not uniform; the frequency of bulkier side chain mutations was much more likely to match between the predicted RECON MSD sequence profiles and the sequence profiles obtained from multiple sequence alignments of natural homologues, suggesting that local changes in side chain contacts due to conformational rearrangments limits the substitution of bulky side chains. However, the overall estimation of mutational tolerance is more conservative than observed in natural homologues' sequence diversity, for reasons discussed in Section II.3.2, which primarily discusses the limitations of the Rosetta scoring function and design method limitations. There are additional limitaions not discussed in Section II.3.2 that are discussed in the next sub-sections that should be considered for the further use of RECON MSD as a bioinformatic tool to predict sequence profiles.

IV.1.1. Limited representation of a protein ensemble biases predicted mutational tolerance

The results presented in Section II.2.6 indicate that local changes in $C_\beta - C_\beta$ distances are correlated with the degree of sequence conservation, whereas either changes in the maximum amplitude or local backbone torsion angle conformational changes do not significantly affect sequence conservation. However, in earlier work that was not included in chapter II on page 16, the percent native sequence recovery was evaluated using different combinations, including number of templates, of designed ensembles of RSV F protein and calmodulin protein backbones. The predicted native sequence recovery deviated by more than 5% between designed ensembles of varying number of backbone templates and maximum RMSD100 values. The estimated sequence conservation within homologues of either RSV F protein or calmodulin are greater than 90%, which was most closely simulated by the inclusion of more template backbones. The differences in percent native sequence recovery given the number of templates suggests that incomplete sampling of conformational space most likely does not accurately assess the constraints a functional ensemble imposes on the fitness selection of tolerated sequences.

Therefore, in cases where very few conformations have been determined as high-resolution structures, the use of RECON MSD to predict mutation preferences of an incompletely represented ensemble will most likely yield biased approximations of sequence tolerance. Additionally, the native backbones used within the benchmark described in chapter II on page 16 were minimized using the Rosetta FastRelax, so that each backbone approximated a local energy minimum. However, a local energy minimum is defined by an equilibrium of conformations, *i.e.*, the conformation used to represent each local energy minimum may or may not be representative of the full conformational space available at each local energy minimum. The incorporation of additional conformations, such as metastable intermediates, and the broader sampling of conformational space available at each loc al energy minimum, which could be achieved through the use of methods like Rosetta Backrub, could yield a more accurate representation of the conformational space within a functional ensemble.

IV.1.2. RECON multi-state design does not consider kinetic barriers and their contributions to mutational tolerance

The RECON MSD assumes that the thermodynamic stability required to assume each conformation is the primary selection pressure on mutational tolerance. This assumption does not account for the kinetic barriers that must be overcome to transition from one conformation to another, in particular the activation energy required to overcome the energy barrier between two local energy minima. In cases where the changes in the entropic contributions to ΔG are sufficient to overcome the kinetic barrier, as in the case of HA_2, RSV F, or DV E proteins, the thermodynamic equilibrium of the protein system is more closely approximated by the energetic contributions of each conformation within an ensemble as compared when the entropic contributions are insufficient to overcome the activation barrier. For instance, the conformational change of calmodulin and the subsequent favorable change in Ca^{2+}-binding affinity — which includes a change from a K_d of approximately 10 μM to that of a much lower K_d depending on the target enzyme — requires binding to an ATP-activated calmodulin-dependent kinase.[213] The activation energy provided by ATP hydrolysis is necessary to overcome the kinetic barrier for conformational change, and is highly regulated by the concentration of both calmodulin and ATPase, such that the conformational change and Ca^{2+} binding occurs with 20 μs.[214,215] Therefore, the mutational tolerance of calmodulin is not only dependent on the stability of all local energy minima, but also the kinetic requirements of Ca^{2+}-binding for proper Ca^{2+} signaling. As discussed in Section II.2.5, the root mean square differences of mutational tolerance predicted by RECON MSD with respect to the mutation frequencies within the multiple sequence alignment of functional homologues was less than the mutation frequencies predicted by SSD for either calmodulin or influenza H3N2 HA_2. However these differences were less pronounced for calmodulin than influenza H3 HA_2, with the quotient of SSD to RECON MSD root mean square differences ($RMSD_{SSD}/RMSD_{RECON}$ with respect to calmodulin functional homologue mutation frequencies, which was equal to $\frac{0.473}{0.632} = 0.748$, as compared to influenza H3N2 HA_2, $\frac{0.534}{0.895} = 0.597$. The larger disparity in mutation frequencies predicted by SSD in relation to influenza H3 HA_2 multiple sequence alignment frequencies as compared to sequences predicted by RECON MSD suggests that the mutation tolerance of HA_2 is more restricted by the stability of its pre- and post-fusion conformations. The lesser disparity in SSD to RECON MSD in relation to the mutation frequencies within calmodulin homologues suggests the sequences which are optimal for each conformation are also closer to the optimal sequence for assuming multiple binding states. In other words, the stability of each calmodulin conformation is not as influential on its mutation tolerance as compared to influenza H3N2 HA_2. However, in both cases, the root mean square difference of designed sequence preferences with respect to homologue mutation preferences was far from zero, indicating that the estimation of mutation preferences based on thermodynamic stability of the designed ensemble was insufficient to accurately predict sequence fitness alone. Ideally, the incorporation of higher-energy conformations, especially transition states, would improve the estimation of the free energy landscape and associated sequence tolerance of each residue.

IV.1.3. The incorporation of co-evolution constraints would likely improve the accuracy of mutational tolerance predictions

Even though the protein backbones included within each of the eight protein ensembles did not represent a complete protein ensemble, the majority of predicted mutation profiles were similar to the mutation profiles as observed within natural homologues. However, depending on the protein ensemble, more than 10% of the predicted sequence profiles contained no similarity to the homologues' sequence profiles. This could be due to the Rosetta energy function identifying low-energy sequences not yet sampled by natural selection, but more likely is due to inaccuracies in sampling. The energy score term `fa_dun`, which calculates the internal energy of a rotamer, is highly dependent on the ϕ and ψ angle, such that dihedral angle deviations as small as 0.1 Å will result in the replacement of the native rotamer and/or native side chain.[216] A study on alternate location of amino acid side chains indicated that less than 50% of Arg, Glu, Gln, Lys and Met dihedral angles will retain the same dihedral angle conformation when determined at resolution of 1.0 Å and 3.0 Å, particularly when these long side-chain residue types are solvent exposed.[217] Therefore, the ambigous determination of long side chain coordinates is prone to inaccurate rotamer internal energy scoring, which is evident in the dissimilarity of calculated amino acid exchangeability rates of these side chain types within RECON MSD in relation to PSI-BLAST profiles as presented in Section II.2.4. The RECON MSD algorithm uses the `favor_native_residue` constraint which provides an added "bonus" to the native residue during scoring to discourage selection of non-native residue side chains.[129] However, based on the over-estimation of total percent native sequence recovery and the dissimilarity of amino acid exchangeability rates of either very short or long side chains, the `favor_native_residue` constraint is over-restrictive for the exchangeability of shorter amino acid side chains while not accounting for disfavorable dihedral angle conformations of long amino acid side chains. One possible alternative for the `favor_native_residue` constraint would be the use of mutation covariation constraints which would encourage sampling of amino acid exchangeability rates and mutation tolerances that are closer to those within natural homologues.

IV.2. Mutation tolerance prediction of viral glycoproteins is clade-specific

In Section II.2.5, the comparison of sequence profiles obtained from the RECON MSD of influenza H3N2 HA_2 with respect to influenza A subtype-specific HA_2 multiple sequence alignment profiles indicated that the predicted mutation profiles were more similar to multiple sequence alignment profiles of subtypes within the H3 clade. H3 and H4 HA_2 sequences contained no gaps in the aligned sequence, unlike the Group 2 H7 or Group 1 H1 and H2 subtype HA_2 sequences in relation to the H3N2 native sequence used for design. As mentioned in the introduction, deep mutational scanning of influenza H1 and H3 HA as well as HIV BG505 and BF520 Env strains indicate that between strains, mutations that tend to be conserved within flexible regions also tend to be unique between strains.[57,59] Simulation of sequence stability using the Eris algorithm[218] of H1N1 and H3N2 HA sequence lineages obtained from 2009 - 2016 indicated that lineage fitness preferentially selected for higher HA stability in later years, but the convergence of which mutations were tolerated within the H1N1 and H3N2 sequences were unique.[219] The H3 clade also can be described as structurally distinct in terms of its HA structure from the H1, H7, and H9 clades.[220] Although these past finding apply mostly to HA sequence fitness selection, accurate prediction of mutation tolerance of viral proteins in general is most likely limited within structurally similar clades, such that predicted mutations that lower the total stability of the starting sequence are most likely sequences that will persist in later lineages. Future efforts to predict persistant viral lineages may benefit from structure-based, especially ensemble-based, predictons of mutation tolerances, such as in the use of RECON MSD. However, either additional improvements will need to be added to account for sequence insertions and/or deletions as well as classification of clade-specific conformations.

IV.3. Sites of vulnerability contain unstable, flexible residues

The preliminary work presented in chapter III on page 42 indicated that residues within RSV F and DV E protein that rearrange their local side chain contacts and have a higher Rosetta total energy score relative to other residues are very likely located within a B-cell epitope. The rearrangement from the determined pre-fusion to post-fusion conformations, however, also changes the relative stability of each residue. Although broadly neutralizing RSV F or DV E epitopes are more likely to include flexible, unstable residues, these residues have a higher relative total score only in the pre-fusion conformation. It is likely that the favorable change in stability helps overcome the entropic barriers necessary to complete fusion. The spontaneous formation of an antibody-antigen interaction requires a negative ΔG of all interactions, and has been shown to be both exothermic and enthalpy-driven.[221–223] The negative enthalpy changes ($-\Delta H$) upon binding far outweighs the loss of entropy ($-\Delta S$) for the unstable residue(s) such that $\Delta G_{antibody-antigen} << \Delta G_{antigen}$, and increases the entropic barriers within the antigen needed to be overcome to assume the post-fusion conformation. Therefore, by forming an antibody-antigen interaction with residues that a relatively higher ΔG value in relation to other residues, the favorable change in relative free energy is higher and is more likely to prevent either attachment or fusion from taking place.

Even the most successful conformational B-cell epitope prediction algorithm, with an accurate epitope prediction rate of 80.6%, does not account for the relative stability, let alone the flexibility of a residue within its predictor set.[194] The Rosetta total energy score and the contact proximity deviation metric, introduced in chapter III on page 42, of each residue was used to quantify each residue's relative stability and flexibility, respectively. Although the fold enrichment for unstable, flexible residues within known epitopes reaches a relatively low enrichment of 0.224 or 0.251 when calculated for the pre-fusion conformation (with an enrichment of 1.00 indicating perfect distinction between epitope and non-epitope residues), there was no enrichment for the prediction of epitopes within the DV E post-fusion conformation, and a negative enrichment within the RSV F post-fusion conformation. This result suggests that the majority of currently known epitopes, particularly those that are targeted by broadly neutralizing antibodies, are conformation-selective for the pre-fusion conformation.

There a few possible explanations for this observation. As mentioned in the previous paragraph, the formation of the antibody-antigen complex that targets the pre-fusion conformation likely blocks fusion through the stabilization of the binding interface. Additionally, as reported in chapter II on page 16, residues with large contact proximity deviations are also more likely to be conserved — of those that are surface accessible, the conservation of an unstable binding interface provides a common site of vulnerability and potential target for broadly neutralizing antibodies. However, epitope mapping efforts initially screen for antibody binding and neutralization using a stabilized protein construct, which is usually limited to a single conformation of the viral glycoprotein. The creation of nanoparticle vaccines of RSV F stabilized in multiple conformations showed preferential binding affinity for the meta-stable pre-fusion conformation by known broadly neutralizing antibodies against RSV F protein.[6] This suggests that earlier screening for RSV F neutralizing antibodies, which may or may not have been screened against the pre-fusion conformation, did not select against pre-fusion specific conformation B-cell epitopes. However, it is possible that the lack of screening against alternative conformations may miss the identification of broadly neutralizing antibodies.

The suggested approach as discussed in Section III.3 aims to identify hotspot residues as candidates for potential epitopes. Even with improvements in computational approaches to predict conformational B-cell epitopes, the likelihood of false negative predictions is high, given that the complete conformational transitions of viral glycoproteins remains yet to be determined. With improvements in the identification of alternative conformations through more refined uses of cryo-EM, such as the trimeric breathing that exposes "cryptic" epitopes within the influenza HA,[119,120] the discovery of additional surface-accessible epitopes will decrease the false negative rate. Together with the suggested improvements to conformational B-cell epitope prediction in Section III.3, these candidate "hotspots" could be used to screen for novel epitopes within antigens.

IV.4. Accounting for conformational selection and/or flexibility of an antigen should improve the effectiveness of reverse vaccinology methods

The aim of this dissertation was not to attribute a viral fusion glycoprotein's ability to undergo large conformational changes as the sole driver of fusion glycoprotein fitness and antigenicity. Indeed, the breadth of factors which determines viral evolution and antigenicity exceeds the scope of this dissertation. Rather, the aim of this dissertation is to focus on how the thermostability changes necessary to assume multiple conformations limits the sequence conservation of select residues, which in turn, provides a conserved binding interface for broadly neutralizing antibody interactions. With RECON MSD it is possible to approximate the mutation preferences of viral glycoproteins. In conjuction with the suggested conformational B-cell epitope prediction method as described in chapter III on page 42, it may be possible to test how these mutations affect the relative stability and flexibility of epitopes by threading the RECON MSD-predicted mutations onto determined conformational ensembles. This particular approach may allow for the identification of a minimal epitope that is conserved in terms of its sequence, relative stability, and flexibility across one or more viral clades. Moreover, this approach may be useful in identifying which conformation, rather than sequence, elicits the greatest neutralizing antibody response via binding.

Subunit vaccine design, such as the design of the FFL_001 epitope-focused immunogen scaffold,[224] has been used to boost subdominant, broadly neutralizing antibody recognition of conformation-specific epitopes.[225] In many cases, recognition of flexible regions is often occluded by glycan shields or the conformation of the glycoprotein itself, such that antigen recognition is prevented by the limited time period in which flexible region is surface-accessible. Therefore, the degree of affinity maturation and number of germline antibodies that target these time-limited epitopes is less than, or subdominant to, other antibodies that target more easilty accessible epitopes. One goal of subunit vaccines is to provide a stabilized structural scaffold corresponding to these kinetically-limited epitopes to encourage affinity maturation and prevelance of antibodies, particulary bnAbs, which can then more readily detect the same structural epitope during infection. The beauty behind subunit vaccines is that they can be engineered to be both sequence- and conformation-specific to boost a very specific antibody response.[225] With the identification of potential antigenic mutations and the specific conformations that elicit a bnAb response, which can be accelerated through the application of RECON MSD as described in chapter II on page 16 and the conformational B-cell epitope prediction method presented in chapter III on page 42, subunit vaccines could provide a platform to elicit not only a single bnAb response, but a panel of bnAbs responses. Thus, the effectiveness of a single vaccination could potentially last for a much greater time period than current vaccination time tables.

There are several potential caveats to using a subunit vaccine cocktail. The most immediately obvious caveat is one that relates to Fab-mediated neutralization and/or protection. As discussed in appendix D on page 117, antibody neutralization is not only conferred through the binding affinity of its paratope-epitope binding interface, but also through binding angle, or rather conformation of the entire antibody-antigen complex. Therefore, insufficient representation of the epitope by a subunit vaccine could potentially boost a non-neutralizing antibody response, as the limited structural representation of the epitope could promote the formation of an incorrect binding angle or other fc-mediated effect. Therefore, it is paramount that the reverse vaccinology method account for not only the precise conformation of the epitope binding interface, but also any secondary interfaces which direct the correct conformation of antibody binding pose.

[1] Stephen C. Harrison. *Viral membrane fusion*. 2015. DOI: 10.1016/j.virol.2015.03.043.

[2] Judith M. White, Sue E. Delos, Matthew Brecher, and Kathryn Schornberg. *Structures and mechanisms of viral membrane fusion proteins: Multiple variations on a common theme*. 2008. DOI: 10.1080/10409230802058320.

[3] John J. Skehel and Don C. Wiley. “Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin”. In: *Annual Review of Biochemistry* 69.1 (June 2000), pp. 531–569. DOI: 10.1146/annurev.biochem.69.1.531.

[4] W. Weissenhorn, A. Dessen, S. C. Harrison, J. J. Skehel, and D. C. Wiley. “Atomic structure of the ectodomain from HIV-1 gp41”. In: *Nature* (1997). DOI: 10.1038/387426a0.

[5] David C. Chan, Deborah Fass, James M. Berger, and Peter S. Kim. “Core structure of gp41 from the HIV envelope glycoprotein”. In: *Cell* (1997). DOI: 10.1016/S0092-8674(00)80205-6.

[6] Nita Patel, Mike J. Massare, Jing Hui Tian, Mimi Guebre-Xabier, Hanxin Lu, Haixia Zhou, Ernest Maynard, Daniel Scott, Larry Ellingsworth, Gregory Glenn, and Gale Smith. “Respiratory syncytial virus prefusogenic fusion (F) protein nanoparticle vaccine: Structure, antigenic profile, immunogenicity, and protection”. In: *Vaccine* (2019). DOI: 10.1016/j.vaccine.2019.07.089.

[7] M. Gordon Joyce, Baoshan Zhang, Li Ou, Man Chen, Gwo Yu Chuang, Aliaksandr Druz, Wing Pui Kong, Yen Ting Lai, Emily J. Rundlet, Yaroslav Tsybovsky, Yongping Yang, Ivelin S. Georgiev, Miklos Guttman, Christopher R. Lees, Marie Pancera, Mallika Sastry, Cinque Soto, Guillaume B.E. Stewart-Jones, Paul V. Thomas, Joseph G. Van Galen, Ulrich Baxa, Kelly K. Lee, John R. Mascola, Barney S. Graham, and Peter D. Kwong. “Iterative structure-based improvement of a fusion-glycoprotein vaccine against RSV”. In: *Nature Structural and Molecular Biology* (2016). DOI: 10.1038/nsmb.3267.

[8] Gert Bolt, Lars Østergaard Pedersen, and Helle Harder Birkeslund. “Cleavage of the respiratory syncytial virus fusion protein is required for its surface expression: Role of furin”. In: *Virus Research* (2000). DOI: 10.1016/S0168-1702(00)00149-0.

[9] R. J. Sugrue, C. Brown, G. Brown, J. Aitken, and H. W. McL. Rixon. “Furin cleavage of the respiratory syncytial virus fusion protein is not a requirement for its transport to the surface of virus-infected cells”. In: *Journal of General Virology* (2001). DOI: 10.1099/0022-1317-82-6-1375.

[10] M. Begoña Ruiz-Argüello, Diana Martín, Steve A. Wharton, Lesley J. Calder, Steve R. Martín, Olga Cano, Miguel Calero, Blanca García-Barreno, John J. Skehel, and José A. Melero. “Thermostability of the human respiratory syncytial virus fusion protein before and after activation: Implications for the membrane-fusion mechanism”. In: *Journal of General Virology* (2004). DOI: 10.1099/vir.0.80318-0.

[11] Vladimir N. Malashkevich, Mona Singh, and Peter S. Kim. “The trimer-of-hairpins motif in membrane fusion: Visna virus”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2001). DOI: 10.1073/pnas.151254798.

[12] Daniel L. Floyd, Justin R. Ragains, John J. Skehel, Stephen C. Harrison, and Antoine M. Van Oijen. “Single-particle kinetics of influenza virus membrane fusion”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2008). DOI: 10.1073/pnas.0807771105.

[13] Tijana Ivanovic, Jason L. Choi, Sean P. Whelan, Antoine M. van Oijen, and Stephen C. Harrison. “Influenza-virus membrane fusion by cooperative fold-back of stochastically induced hemagglutinin intermediates”. In: *eLife* (2013). DOI: 10.7554/eLife.00333.

[14] Jason S. McLellan, William C. Ray, and Mark E. Peeples. “Structure and function of respiratory syncytial virus surface glycoproteins”. In: *Current Topics in Microbiology and Immunology* (2013). DOI: 10.1007/978-3-642-38919-1-4.

[15] Victor Buzon, Ganesh Natrajan, David Schibli, Felix Campelo, Michael M. Kozlov, and Winfried Weissenhorn. “Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions”. In: *PLoS Pathogens* (2010). DOI: 10.1371/journal.ppat.1000880.

[16] Beatriz Apellániz, Edurne Rujas, Soraya Serrano, Koldo Morante, Kouhei Tsumoto, Jose M.M. Caaveiro, M. Ángeles Jiménez, and José L. Nieva. “The atomic structure of the HIV-1 gp41 transmembrane domain and its connection to the immunogenic membrane-proximal external region”. In: *Journal of Biological Chemistry* (2015). DOI: 10.1074/jbc.M115.644351.

[17] Y. Li, X. Han, A. L. Lai, J. H. Bushweller, D. S. Cafiso, and L. K. Tamm. “Membrane Structures of the Hemifusion-Inducing Fusion Peptide Mutant G1S and the Fusion-Blocking Mutant G1V of Influenza Virus Hemagglutinin Suggest a Mechanism for Pore Opening in Membrane Fusion”. In: *Journal of Virology* (2005). DOI: 10.1128/jvi.79.18.12065-12076.2005.

[18] D. E. Klein, J. L. Choi, and S. C. Harrison. “Structure of a Dengue Virus Envelope Protein Late-Stage Fusion Intermediate”. In: *Journal of Virology* (2013). DOI: 10.1128/jvi.02957-12.

[19] Yorgo Modis, Steven Ogata, David Clements, and Stephen C. Harrison. “Structure of the dengue virus envelope protein after membrane fusion”. In: *Nature* (2004). DOI: 10.1038/nature02165.

[20] Ying Zhang, Wei Zhang, Steven Ogata, David Clements, James H. Strauss, Timothy S. Baker, Richard J. Kuhn, and Michael G. Rossmann. “Conformational changes of the flavivirus E glycoprotein”. In: *Structure* (2004). DOI: 10.1016/j.str.2004.06.019.

[21] Long Li, Joyce Jose, Ye Xiang, Richard J. Kuhn, and Michael G. Rossmann. “Structural changes of envelope proteins during alphavirus fusion”. In: *Nature* (2010). DOI: 10.1038/nature09546.

[22] Long Li, Shee Mei Lok, I. Mei Yu, Ying Zhang, Richard J. Kuhn, Jue Chen, and Michael G. Rossmann. “The flavivirus precursor membrane-envelope protein complex: Structure and maturation”. In: *Science* (2008). DOI: 10.1126/science.1153263.

[23] Julien Lescar, Alain Roussel, Michelle W. Wien, Jorge Navaza, Stephen D. Fuller, Gisela Wengler, Gerd Wengler, and Félix A. Rey. “The fusion glycoprotein shell of Semliki Forest virus: An icosahedral assembly primed for fusogenic activation at endosomal pH”. In: *Cell* (2001). DOI: 10.1016/S0092-8674(01)00303-8.

[24] Don L. Gibbons, Marie Christine Vaney, Alain Roussel, Armelle Vigouroux, Brigid Reilly, Jean Lepault, Margaret Kielian, and Félix A. Rey. “Conformational change and protein-protein interactions of the fusion protein of Semliki Forest virus”. In: *Nature* (2004). DOI: 10.1038/nature02239.

[25] Aurélie A. Albertini, Cécile Mérigoux, Sonia Libersou, Karine Madiona, Stéphane Bressanelli, Stéphane Roche, Jean Lepault, Ronald Melki, Patrice Vachette, and Yves Gaudin. "Characterization of monomeric intermediates during VSV glycoprotein structural transition". In: *PLoS Pathogens* (2012). DOI: 10.1371/journal.ppat.1002556.

[26] Eduard Baquero, Aurélie A. Albertini, Patrice Vachette, Jean Lepault, Stéphane Bressanelli, and Yves Gaudin. *Intermediate conformations during viral fusion glycoprotein structural transition*. 2013. DOI: 10.1016/j.coviro.2013.03.006.

[27] Stéphane Roche, Stéphane Bressanelli, Félix A. Rey, and Yves Gaudin. "Crystal structure of the low-pH form of the vesicular stomatitis virus glycoprotein G". In: *Science* (2006). DOI: 10.1126/science.1127683.

[28] Stéphane Roche, Félix A. Rey, Yves Gaudin, and Stéphane Bressanelli. "Structure of the prefusion form of the vesicular stomatitis virus glycoprotein G". In: *Science* (2007). DOI: 10.1126/science.1135710.

[29] Eduard Baquero, Aurélie A. Albertini, Patrice Vachette, Jean Lepault, Stéphane Bressanelli, and Yves Gaudin. *Intermediate conformations during viral fusion glycoprotein structural transition*. 2013. DOI: 10.1016/j.coviro.2013.03.006.

[30] A. Ferlin, H. Raux, E. Baquero, J. Lepault, and Y. Gaudin. "Characterization of pH-Sensitive Molecular Switches That Trigger the Structural Transition of Vesicular Stomatitis Virus Glycoprotein from the Postfusion State toward the Prefusion State". In: *Journal of Virology* (2014). DOI: 10.1128/jvi.01962-14.

[31] Eduard Baquero, Aurélie A Albertini, Hélène Raux, Abbas Abou-Hamdan, Elisabetta Boeri-Erba, Malika Ouldali, Linda Buonocore, John K Rose, Jean Lepault, Stéphane Bressanelli, and Yves Gaudin. "Structural intermediates in the fusion-associated transition of vesiculovirus glycoprotein". In: *The EMBO Journal* (2017). DOI: 10.15252/embj.201694565.

[32] Ken A. Dill, S. Banu Ozkan, Thomas R. Weikl, John D. Chodera, and Vincent A. Voelz. *The protein folding problem: when will it be solved?* 2007. DOI: 10.1016/j.sbi.2007.06.001.

[33] David Baker and David A. Agard. "Kinetics versus Thermodynamics in Protein Folding". In: *Biochemistry* (1994). DOI: 10.1021/bi00190a002.

[34] Christian B. Anfinsen. *Principles that govern the folding of protein chains*. 1973. DOI: 10.1126/science.181.4096.223.

[35] Chavela M. Carr, Charu Chaudhry, and Peter S. Kim. "Influenza hemagglutinin is spring-loaded by a metastable native conformation". In: *Proceedings of the National Academy of Sciences of the United States of America* (1997). DOI: 10.1073/pnas.94.26.14306.

[36] Xun Zhao, Mona Singh, Vladimir N. Malashkevich, and Peter S. Kim. "Structural characterization of the human respiratory syncytial virus fusion protein core". In: *Proceedings of the National Academy of Sciences of the United States of America* (2000). DOI: 10.1073/pnas.260499197.

[37] Meher K. Prakash, Alessandro Barducci, and Michele Parrinello. "Probing the mechanism of pH-induced large-scale conformational changes in dengue virus envelope protein using atomistic simulations". In: *Biophysical Journal* (2010). DOI: 10.1016/j.bpj.2010.04.024.

[38] H Zhang. "Investigating the stability of dengue virus envelope protein dimer using well-tempered metadynamics simulations". In: *Proteins: Structure, Function and Bioinformatics* (2019). DOI: https://doi.org/10.1002/prot.25844.

[39] James C. Whisstock and Stephen P. Bottomley. *Molecular gymnastics: serpin structure, folding and misfolding*. 2006. DOI: 10.1016/j.sbi.2006.10.005.

[40] Julie L. Sohl, Sheila S. Jaswal, and David A. Agard. "Unfolded conformations of α-lytic protease are more stable than its native state". In: *Nature* (1998). DOI: 10.1038/27470.

[41] V. V. Hemanth Giri Rao and Shachi Gosavi. "On the folding of a structurally complex protein to its metastable active state". In: *Proceedings of the National Academy of Sciences of the United States of America* (2018). DOI: 10.1073/pnas.1708173115.

[42] Aaron R. Dinner and Martin Karplus. "A metastable state in folding simulations of a protein model". In: *Nature Structural Biology* (1998). DOI: 10.1038/nsb0398-236.

[43] Cheolju Lee, Soon Ho Park, Min Youn Lee, and Myeong Hee Yu. "Regulation of protein function by native metastability". In: *Proceedings of the National Academy of Sciences of the United States of America* (2000). DOI: 10.1073/pnas.97.14.7727.

[44] Ineke Braakman, Helana Hoover-Litty, Krystn R. Wagner, and Ari Helenius. "Folding of influenza hemagglutinin in the endoplasmic reticulum". In: *Journal of Cell Biology* (1991). DOI: 10.1083/jcb.114.3.401.

[45] J. L. Slon Campos, S. Marchese, J. Rana, M. Mossenta, M. Poggianella, M. Bestagno, and O. R. Burrone. "Temperature-dependent folding allows stable dimerization of secretory and virus-Associated e proteins of Dengue and Zika viruses in mammalian cells". In: *Scientific Reports* (2017). DOI: 10.1038/s41598-017-01097-5.

[46] Sven O. Dahms, Marcelino Arciniega, Torsten Steinmetzer, Robert Huber, and Manuel E. Than. "Structure of the unliganded form of the proprotein convertase furin suggests activation by a substrate-induced mechanism". In: *Proceedings of the National Academy of Sciences of the United States of America* (2016). DOI: 10.1073/pnas.1613630113.

[47] Danielle M. Williamson, Johannes Elferich, Parvathy Ramakrishnan, Gary Thomas, and Ujwal Shinde. "The mechanism by which a propeptide-encoded pH sensor regulates spatiotemporal activation of furin". In: *Journal of Biological Chemistry* (2013). DOI: 10.1074/jbc.M112.442681.

[48] R. P. Rand and V. A. Parsegian. "Physical force considerations in model and biological membranes". In: *Canadian Journal of Biochemistry and Cell Biology* (1984). DOI: 10.1139/o84-097.

[49] Esteban Domingo, Donna Sabo, Tadatsugu Taniguchi, and Charles Weissmann. "Nucleotide sequence heterogeneity of an RNA phage population". In: *Cell* (1978). DOI: 10.1016/0092-8674(78)90223-4.

[50] Kenneth Murphy. *Immunobiology, 8th Edition*. 2012.

[51] Siobain Duffy, Laura A. Shackelton, and Edward C. Holmes. *Rates of evolutionary change in viruses: Patterns and determinants*. 2008. DOI: 10.1038/nrg2323.

[52] Rafael Sanjuán, Andrés Moya, and Santiago F. Elena. "The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus". In: *Proceedings of the National Academy of Sciences of the United States of America* (2004). DOI: 10.1073/pnas.0400146101.

[53] P. Carrasco, F. de la Iglesia, and S. F. Elena. "Distribution of Fitness and Virulence Effects Caused by Single-Nucleotide Substitutions in Tobacco Etch Virus". In: *Journal of Virology* (2007). DOI: 10.1128/jvi.00524-07.

[54] Elisa Visher, Shawn E. Whitefield, John T. McCrone, William Fitzsimmons, and Adam S. Lauring. "The Mutational Robustness of Influenza A Virus". In: *PLoS Pathogens* (2016). DOI: 10.1371/journal.ppat.1005856.

[55] Ashley Acevedo, Leonid Brodsky, and Raul Andino. "Mutational and fitness landscapes of an RNA virus revealed through population sequencing". In: *Nature* (2014). DOI: 10.1038/nature12861.

[56] Renata C. Fleith, Francisco P. Lobo, Paula F. Dos Santos, Mariana M. Rocha, Juliano Bordignon, Daisy M. Strottmann, Daniel O. Patricio, Wander R. Pavanelli, Maria Lo Sarzi, Claudia N.D. Santos, Brian J. Ferguson, and Daniel S. Mansur. "Genome-wide analyses reveal a highly conserved Dengue virus envelope peptide which is critical for virus viability and antigenic in humans". In: *Scientific Reports* (2016). DOI: 10.1038/srep36339.

[57] Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. "Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants". In: *Proceedings of the National Academy of Sciences of the United States of America* (2018). DOI: 10.1073/pnas.1806133115.

[58] C E Shannon. "The mathematical theory of communication. 1963". In: *MD Comput* 14.4 (1997), pp. 306–317.

[59] Hugh K. Haddox, Adam S. Dingens, Sarah K. Hilton, Julie Overbaugh, and Jesse D. Bloom. "Mapping mutational effects along the evolutionary landscape of HIV envelope". In: *eLife* (2018). DOI: 10.7554/eLife.34420.

[60] William W. Thompson, David K. Shay, Eric Weintraub, Nancy Cox, Larry J. Anderson, and Keiji Fukuda. "Mortality associated with influenza and respiratory syncytial virus in the United States". In: *Journal of the American Medical Association* (2003). DOI: 10.1001/jama.289.2.179.

[61] Davide Corti, Elisabetta Cameroni, Barbara Guarino, Nicole L. Kallewaard, Qing Zhu, and Antonio Lanzavecchia. *Tackling influenza with broadly neutralizing antibodies*. 2017. DOI: 10.1016/j.coviro.2017.03.002.

[62] Cyrille Dreyfus, Nick S. Laursen, Ted Kwaks, David Zuijdgeest, Reza Khayat, Damian C. Ekiert, Jeong Hyun Lee, Zoltan Metlagel, Miriam V. Bujny, Mandy Jongeneelen, Remko Van Der Vlugt, Mohammed Lamrani, Hans J.W.M. Korse, Eric Geelen, Özcan Sahin, Martijn Sieuwerts, Just P.J. Brakenhoff, Ronald Vogels, Olive T.W. Li, Leo L.M. Poon, Malik Peiris, Wouter Koudstaal, Andrew B. Ward, Ian A. Wilson, Jaap Goudsmit, and Robert H.E. Friesen. "Highly conserved protective epitopes on influenza B viruses". In: *Science* (2012). DOI: 10.1126/science.1222908.

[63] Joshua A. Horwitz, Yotam Bar-On, Ching Lan Lu, Daniela Fera, Ainsley A.K. Lockhart, Julio C.C. Lorenzi, Lilian Nogueira, Jovana Golijanin, Johannes F. Scheid, Michael S. Seaman, Anna Gazumyan, Susan Zolla-Pazner, and Michel C. Nussenzweig. "Non-neutralizing Antibodies Alter the Course of HIV-1 Infection In Vivo". In: *Cell* (2017). DOI: 10.1016/j.cell.2017.06.048.

[64] N. A. Doria-Rose, R. M. Klein, M. M. Manion, S. O'Dell, A. Phogat, B. Chakrabarti, C. W. Hallahan, S. A. Migueles, J. Wrammert, R. Ahmed, M. Nason, R. T. Wyatt, J. R. Mascola, and M. Connors. "Frequency and Phenotype of Human Immunodeficiency Virus Envelope-Specific B Cells from Patients with Broadly Cross-Neutralizing Antibodies". In: *Journal of Virology* (2009). DOI: 10.1128/jvi.01583-08.

[65] Iliyana Mikell, D. Noah Sather, Spyros A. Kalams, Marcus Altfeld, Galit Alter, and Leonidas Stamatatos. “Characteristics of the earliest cross-neutralizing antibody response to HIV-1”. In: *PLoS Pathogens* (2011). DOI: 10.1371/journal.ppat.1001251.

[66] E. S. Gray, M. C. Madiga, T. Hermanus, P. L. Moore, C. K. Wibmer, N. L. Tumba, L. Werner, K. Mlisana, S. Sibeko, C. Williamson, S. S. Abdool Karim, and L. Morris. “The Neutralization Breadth of HIV-1 Develops Incrementally over Four Years and Is Associated with CD4+ T Cell Decline and High Viral Load during Acute Infection”. In: *Journal of Virology* (2011). DOI: 10.1128/jvi.00198-11.

[67] D. N. Sather, J. Armann, L. K. Ching, A. Mavrantoni, G. Sellhorn, Z. Caldwell, X. Yu, B. Wood, S. Self, S. Kalams, and L. Stamatatos. “Factors Associated with the Development of Cross-Reactive Neutralizing Antibodies during Human Immunodeficiency Virus Type 1 Infection”. In: *Journal of Virology* (2009). DOI: 10.1128/jvi.02036-08.

[68] Jason S. McLellan, Man Chen, Sherman Leung, Kevin W. Graepel, Xiulian Du, Yongping Yang, Tongqing Zhou, Ulrich Baxa, Etsuko Yasuda, Tim Beaumont, Azad Kumar, Kayvon Modjarrad, Zizheng Zheng, Min Zhao, Ningshao Xia, Peter D. Kwong, and Barney S. Graham. “Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody”. In: *Science* (2013). DOI: 10.1126/science.1234914.

[69] Mark J. Kwakkenbos, Sean A. Diehl, Etsuko Yasuda, Arjen Q. Bakker, Caroline M.M. Van Geelen, Michaël V. Lukens, Grada M. Van Bleek, Myra N. Widjojoatmodjo, Willy M.J.M. Bogers, Henrik Mei, Andreas Radbruch, Ferenc A. Scheeren, Hergen Spits, and Tim Beaumont. “Generation of stable monoclonal antibody-producing B cell receptor-positive human memory B cells by genetic programming”. In: *Nature Medicine* (2010). DOI: 10.1038/nm.2071.

[70] Min Zhao, Zi-Zheng Zheng, Man Chen, Kayvon Modjarrad, Wei Zhang, Lu-Ting Zhan, Jian-Li Cao, Yong-Peng Sun, Jason S. McLellan, Barney S. Graham, and Ning-Shao Xia. “ Discovery of a Prefusion Respiratory Syncytial Virus F-Specific Monoclonal Antibody That Provides Greater In Vivo Protection than the Murine Precursor of Palivizumab ”. In: *Journal of Virology* (2017). DOI: 10.1128/jvi.00176-17.

[71] David B Roth. “V(D)J Recombination: Mechanism, Errors, and Fidelity Generation of antigen receptor diversity: a double-edged sword HHS Public Access”. In: *Microbiol Spectr.* (2014). DOI: 10.1128/microbiolspec.MDNA3-0041-2014.

[72] Fumihiko Matsuda, Kazuo Ishii, Patrice Bourvagnet, Kei Ichi Kuma, Hidenori Hayashida, Takashi Miyata, and Tasuku Honjo. “The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus”. In: *Journal of Experimental Medicine* (1998). DOI: 10.1084/jem.188.11.2151.

[73] Aihong Li, Montse Rue, Jianbiao Zhou, Hongjun Wang, Meredith A. Goldwasser, Donna Neuberg, Virginia Dalton, David Zuckerman, Cheryl Lyons, Lewis B. Silverman, Stephen E. Sallan, and John G. Gribben. “Utilization of Ig heavy chain variable, diversity, and joining gene segments in children with B-lineage acute lymphoblastic leukemia: Implications for the mechanisms of VDJ recombination and for pathogenesis”. In: *Blood* (2004). DOI: 10.1182/blood-2003-11-3857.

[74] Andrew M. Collins and Corey T. Watson. *Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire*. 2018. DOI: 10.3389/fimmu.2018.02249.

[75] Kazuhiko Kawasaki, Shinsei Minoshima, Eriko Nakato, Kazunori Shibuya, Ai Shintani, James L. Schmeits, Jun Wang, and Nobuyoshi Shimizu. “One-megabase sequence analysis of the human immunoglobulin λ gene locus”. In: *Genome Research* (1997). DOI: 10.1101/gr.7.3.250.

[76] Grace Teng and F. Nina Papavasiliou. “Immunoglobulin Somatic Hypermutation”. In: *Annual Review of Genetics* (2007). DOI: 10.1146/annurev.genet.41.110306.130340.

[77] Christopher J. Jolly, Simon D. Wagner, Cristina Rada, Norman Klix, César Milstein, and Michael S. Neuberger. “The targeting of somatic hypermutation”. In: *Seminars in Immunology* (1996). DOI: 10.1006/smim.1996.0020.

[78] Gabriel D. Victora and Michel C. Nussenzweig. “Germinal Centers”. In: *Annual Review of Immunology* (2012). DOI: 10.1146/annurev-immunol-020711-075032.

[79] Shane Crotty, Phil Felgner, Huw Davies, John Glidewell, Luis Villarreal, and Rafi Ahmed. “Cutting Edge: Long-Term B Cell Memory in Humans after Smallpox Vaccination”. In: *The Journal of Immunology* (2003). DOI: 10.4049/jimmunol.171.10.4969.

[80] Ian J. Amanna, Nichole E. Carlson, and Mark K. Slifka. “Duration of humoral immunity to common viral and vaccine antigens”. In: *New England Journal of Medicine* (2007). DOI: 10.1056/NEJMoa066092.

[81] Xiaocong Yu, Tshidi Tsibane, Patricia A. McGraw, Frances S. House, Christopher J. Keefer, Mark D. Hicar, Terrence M. Tumpey, Claudia Pappas, Lucy A. Perrone, Osvaldo Martinez, James Stevens, Ian A. Wilson, Patricia V. Aguilar, Eric L. Altschuler, Christopher F. Basler, and James E. Crowe. “Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors”. In: *Nature* (2008). DOI: 10.1038/nature07231.

[82] Scott A. Smith and James E. Crowe, Jr. “Use of Human Hybridoma Technology To Isolate Human Monoclonal Antibodies”. In: *Microbiology Spectrum* (2015). DOI: 10.1128/microbiolspec.aid-0027-2014.

[83] Marc H.V. Van Regenmortel. “What is a b-cell epitope?” In: *Methods in Molecular Biology* (2009). DOI: 10.1007/978-1-59745-450-6_1.

[84] Andrew A. Bogan and Kurt S. Thorn. “Anatomy of hot spots in protein interfaces”. In: *Journal of Molecular Biology* (1998). DOI: 10.1006/jmbi.1998.1843.

[85] David C. Benjamin and Samuel S. Perdue. “Site-directed mutagenesis in epitope mapping”. In: *Methods: A Companion to Methods in Enzymology* (1996). DOI: 10.1006/meth.1996.0058.

[86] Nimrod D. Rubinstein, Itay Mayrose, Dan Halperin, Daniel Yekutieli, Jonathan M. Gershoni, and Tal Pupko. “Computational characterization of B-cell epitopes”. In: *Molecular Immunology* (2008). DOI: 10.1016/j.molimm.2007.10.016.

[87] Davide Corti, Jarrod Voss, Steven J. Gamblin, Giosiana Codoni, Annalisa Macagno, David Jarrossay, Sebastien G. Vachieri, Debora Pinna, Andrea Minola, Fabrizia Vanzetta, Chiara Silacci, Blanca M. Fernandez-Rodriguez, Gloria Agatic, Siro Bianchi, Isabella Giacchetto-Sasselli, Lesley Calder, Federica Sallusto, Patrick Collins, Lesley F. Haire, Nigel Temperton, Johannes P.M. Langedijk, John J. Skehel, and Antonio Lanzavecchia. “A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins”. In: *Science* (2011). DOI: 10.1126/science.1205669.

[88] Damian C. Ekiert, Gira Bhabha, Marc Andre Elsliger, Robert H.E. Friesen, Mandy Jongeneelen, Mark Throsby, Jaap Goudsmit, and Ian A. Wilson. “Antibody recognition of a highly conserved influenza virus epitope”. In: *Science* (2009). DOI: 10.1126/science.1171491.

[89] Damian C. Ekiert, Robert H.E. Friesen, Gira Bhabha, Ted Kwaks, Mandy Jongeneelen, Wenli Yu, Carla Ophorst, Freek Cox, Hans J.W.M. Korse, Boerries Brandenburg, Ronald Vogels, Just P.J. Brakenhoff, Ronald Kompier, Martin H. Koldijk, Lisette A.H.M. Cornelissen, Leo L.M. Poon, Malik Peiris, Wouter Koudstaal, Ian A. Wilson, and Jaap Goudsmit. “A highly conserved neutralizing epitope on group 2 influenza A viruses”. In: *Science* (2011). DOI: 10.1126/science.1204839.

[90] C. Dreyfus, D. C. Ekiert, and I. A. Wilson. “Structure of a Classical Broadly Neutralizing Stem Antibody in Complex with a Pandemic H2 Influenza Virus Hemagglutinin”. In: *Journal of Virology* (2013). DOI: 10.1128/jvi.02975-12.

[91] Peter S. Lee, Nobuko Ohshima, Robyn L. Stanfield, Wenli Yu, Yoshitaka Iba, Yoshinobu Okuno, Yoshikazu Kurosawa, and Ian A. Wilson. “Receptor mimicry by antibody F045-092 facilitates universal binding to the H3 subtype of influenza virus”. In: *Nature Communications* (2014). DOI: 10.1038/ncomms4614.

[92] Rui Xu, Jens C. Krause, Ryan Mcbride, James C. Paulson, James E. Crowe, and Ian A. Wilson. “A recurring motif for antibody recognition of the receptor-binding site of influenza hemagglutinin”. In: *Nature Structural and Molecular Biology* (2013). DOI: 10.1038/nsmb.2500.

[93] Brett D. Welch, Yuanyuan Liu, Christopher A. Kors, George P. Leser, Theodore S. Jardetzky, and Robert A. Lamb. “Structure of the cleavage-activated prefusion form of the parainfluenza virus 5 fusion protein”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2012). DOI: 10.1073/pnas.1213802109.

[94] Hsien Sheng Yin, Xiaolin Wen, Reay G. Paterson, Robert A. Lamb, and Theodore S. Jardetzky. “Structure of the parainfluenza virus 5 F protein in its metastable, prefusion conformation”. In: *Nature* (2006). DOI: 10.1038/nature04322.

[95] Joyce J.W. Wong, Reay G. Paterson, Robert A. Lamb, and Theodore S. Jardetzky. “Structure and stabilization of the Hendra virus F glycoprotein in its prefusion form”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2016). DOI: 10.1073/pnas.1523303113.

[96] João M. Dias, Ana I. Kuehne, Dafna M. Abelson, Shridhar Bale, Anthony C. Wong, Peter Halfmann, Majidat A. Muhammad, Marnie L. Fusco, Samantha E. Zak, Eugene Kang, Yoshihiro Kawaoka, Kartik Chandran, John M. Dye, and Erica Ollmann Saphire. “A shared structural solution for neutralizing ebolaviruses”. In: *Nature Structural and Molecular Biology* (2011). DOI: 10.1038/nsmb.2150.

[97] Jeffrey E. Lee, Marnie L. Fusco, Ann J. Hessell, Wendelien B. Oswald, Dennis R. Burton, and Erica Ollmann Saphire. “Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor”. In: *Nature* (2008). DOI: 10.1038/nature07082.

[98] Morgan S.A. Gilman, Polina Furmanova-Hollenstein, Gabriel Pascual, Angélique B. van ‘t Wout, Johannes P.M. Langedijk, and Jason S. McLellan. “Transient opening of trimeric prefusion RSV F proteins”. In: *Nature Communications* (2019). DOI: 10.1038/s41467-019-09807-5.

[99] Dominique P. Frueh, Andrew C. Goodrich, Subrata H. Mishra, and Scott R. Nichols. *NMR methods for structural studies of large monomeric and multimeric proteins*. 2013. DOI: 10.1016/j.sbi.2013.06.016.

[100] Dmitry Lyumkis, Jean Philippe Julien, Natalia De Val, Albert Cupo, Clinton S. Potter, Per Johan Klasse, Dennis R. Burton, Rogier W. Sanders, John P. Moore, Bridget Carragher, Ian A. Wilson, and Andrew B. Ward. "Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer". In: *Science* (2013). DOI: 10.1126/science.1245627.

[101] Andrew B. Ward and Ian A. Wilson. *The HIV-1 envelope glycoprotein structure: nailing down a moving target*. 2017. DOI: 10.1111/imr.12507.

[102] Andrew I. Flyak, Philipp A. Ilinykh, Charles D. Murin, Tania Garron, Xiaoli Shen, Marnie L. Fusco, Takao Hashiguchi, Zachary A. Bornholdt, James C. Slaughter, Gopal Sapparapu, Curtis Klages, Thomas G. Ksiazek, Andrew B. Ward, Erica Ollmann Saphire, Alexander Bukreyev, and James E. Crowe. "Mechanism of Human Antibody-Mediated Neutralization of Marburg Virus". In: *Cell* (2015). DOI: 10.1016/j.cell.2015.01.031.

[103] Marie Pancera, Tongqing Zhou, Aliaksandr Druz, Ivelin S. Georgiev, Cinque Soto, Jason Gorman, Jinghe Huang, Priyamvada Acharya, Gwo Yu Chuang, Gilad Ofek, Guillaume B.E. Stewart-Jones, Jonathan Stuckey, Robert T. Bailer, M. Gordon Joyce, Mark K. Louder, Nancy Tumba, Yongping Yang, Baoshan Zhang, Myron S. Cohen, Barton F. Haynes, John R. Mascola, Lynn Morris, James B. Munro, Scott C. Blanchard, Walther Mothes, Mark Connors, and Peter D. Kwong. "Structure and immune recognition of trimeric pre-fusion HIV-1 Env". In: *Nature* (2014). DOI: 10.1038/nature13808.

[104] Jeong Hyun Lee, Natalia De Val, Dmitry Lyumkis, and Andrew B. Ward. "Model building and refinement of a natively glycosylated HIV-1 Env protein by high-resolution cryoelectron microscopy". In: *Structure* (2015). DOI: 10.1016/j.str.2015.07.020.

[105] Matteo Bianchi, Hannah L. Turner, Bartek Nogal, Christopher A. Cottrell, David Oyen, Matthias Pauthner, Raiza Bastidas, Rebecca Nedellec, Laura E. McCoy, Ian A. Wilson, Dennis R. Burton, Andrew B. Ward, and Lars Hangartner. "Electron-Microscopy-Based Epitope Mapping Defines Specificities of Polyclonal Antibodies Elicited during HIV-1 BG505 Envelope Trimer Immunization". In: *Immunity* (2018). DOI: 10.1016/j.immuni.2018.07.009.

[106] Benoît Arragain, Juan Reguera, Ambroise Desfosses, Irina Gutsche, Guy Schoehn, and Hélène Malet. "High resolution cryo-EM structure of the helical RNA-bound Hantaan virus nucleocapsid reveals its assembly mechanisms". In: *eLife* (2019). DOI: 10.7554/eLife.43075.

[107] Charlotte A. Scarff, Martin J.G. Fuller, Rebecca F. Thompson, and Matthew G. Iadaza. "Variations on negative stain electron microscopy methods: Tools for tackling challenging systems". In: *Journal of Visualized Experiments* (2018). DOI: 10.3791/57199.

[108] Jarrod J. Mousa, Marion F. Sauer, Alexander M. Sevy, Jessica A. Finn, John T. Bates, Gabriela Alvarado, Hannah G. King, Leah B. Loerinc, Rachel H. Fong, Benjamin J. Doranz, Bruno E. Correia, Oleksandr Kalyuzhniy, Xiaolin Wen, Theodore S. Jardetzky, William R. Schief, Melanie D. Ohi, Jens Meiler, and James E. Crowe. "Structural basis for nonneutralizing antibody competition at antigenic site II of the respiratory syncytial virus fusion protein". In: *Proceedings of the National Academy of Sciences of the United States of America* (2016). DOI: 10.1073/pnas.1609449113.

[109] Elizabeth A. Christian, Kristen M. Kahle, Kimberly Mattia, Bridget A. Puffer, Jennifer M. Pfaff, Adam Miller, Cheryl Paes, Edgar Davidson, and Benjamin J. Doranz. “Atomic-level functional model of dengue virus Envelope protein infectivity”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2013). DOI: 10.1073/pnas.1310962110.

[110] Edgar Davidson and Benjamin J. Doranz. “A high-throughput shotgun mutagenesis approach to mapping B-cell antibody epitopes”. In: *Immunology* (2014). DOI: 10.1111/imm.12323.

[111] Romain Rouet, David Lowe, and Daniel Christ. *Stability engineering of the human antibody repertoire*. 2014. DOI: 10.1016/j.febslet.2013.11.029.

[112] Richard Y.C. Huang, Stanley R. Krystek, Nathan Felix, Robert F. Graziano, Mohan Srinivasan, Achal Pashine, and Guodong Chen. “Hydrogen/deuterium exchange mass spectrometry and computational modeling reveal a discontinuous epitope of an antibody/TL1A Interaction”. In: *mAbs* (2018). DOI: 10.1080/19420862.2017.1393595.

[113] Lars G. Fägerstam, Åsa Frostell, Robert Karlsson, Mari Kullman, Anita Larsson, Magnus Malmqvist, and Helena Butt. “Detection of antigen—antibody interactions by surface plasmon resonance. Application to Epitope Mapping”. In: *Journal of Molecular Recognition* (1990). DOI: 10.1002/jmr.300030507.

[114] Pamela Holzlöhner and Katja Hanack. “Generation of murine monoclonal antibodies by hybridoma technology”. In: *Journal of Visualized Experiments* (2017). DOI: 10.3791/54832.

[115] Natsue Omi, Yuichi Tokuda, Yoko Ikeda, Morio Ueno, Kazuhiko Mori, Chie Sotozono, Shigeru Kinoshita, Masakazu Nakano, and Kei Tashiro. “Efficient and reliable establishment of lymphoblastoid cell lines by Epstein-Barr virus transformation from a limited amount of peripheral blood”. In: *Scientific Reports* (2017). DOI: 10.1038/srep43833.

[116] Sonu Kumar, Anita Sarkar, Pavel Pugach, Rogier W. Sanders, John P. Moore, Andrew B. Ward, and Ian A. Wilson. “Capturing the inherent structural dynamics of the HIV-1 envelope glycoprotein fusion peptide”. In: *Nature Communications* (2019). DOI: 10.1038/s41467-019-08738-5.

[117] Goran Bajic, Max J. Maron, Yu Adachi, Taishi Onodera, Kevin R. McCarthy, Charles E. McGee, Gregory D. Sempowski, Yoshimasa Takahashi, Garnett Kelsoe, Masayuki Kuraoka, and Aaron G. Schmidt. “Influenza Antigen Engineering Focuses Immune Responses to a Subdominant but Broadly Protective Viral Epitope”. In: *Cell Host and Microbe* (2019). DOI: 10.1016/j.chom.2019.04.003.

[118] Akiko Watanabe, Kevin R. McCarthy, Masayuki Kuraoka, Aaron G. Schmidt, Yu Adachi, Taishi Onodera, Keisuke Tonouchi, Timothy M. Caradonna, Goran Bajic, Shengli Song, Charles E. McGee, Gregory D. Sempowski, Feng Feng, Patricia Urick, Thomas B. Kepler, Yoshimasa Takahashi, Stephen C. Harrison, and Garnett Kelsoe. “Antibodies to a Conserved Influenza Head Interface Epitope Protect by an IgG Subtype-Dependent Mechanism”. In: *Cell* (2019). DOI: 10.1016/j.cell.2019.03.048.

[119] Sandhya Bangaru, Shanshan Lang, Michael Schotsaert, Hillary A. Vanderven, Xueyong Zhu, Nurgun Kose, Robin Bombardi, Jessica A. Finn, Stephen J. Kent, Pavlo Gilchuk, Iuliia Gilchuk, Hannah L. Turner, Adolfo García-Sastre, Sheng Li, Andrew B. Ward, Ian A. Wilson, and James E. Crowe. “A Site of Vulnerability on the Influenza Virus Hemagglutinin Head Domain Trimer Interface”. In: *Cell* (2019). DOI: 10.1016/j.cell.2019.04.011.

[120] Hannah L. Turner, Jesper Pallesen, Shanshan Lang, Sandhya Bangaru, Sarah Urata, Sheng Li, Christopher A. Cottrell, Charles A. Bowman, James E. Crowe, Ian A. Wilson, and Andrew B. Ward. "Potent anti-influenza H7 human monoclonal antibody induces separation of hemagglutinin receptor-binding head domains". In: *PLoS Biology* (2019). DOI: 10.1371/journal.pbio.3000139.

[121] Yan Wu and George F. Gao. *"Breathing" Hemagglutinin Reveals Cryptic Epitopes for Universal Influenza Vaccine Design*. 2019. DOI: 10.1016/j.cell.2019.04.034.

[122] Cyrus Levinthal. "How to Fold Graciously. In Mossbauer Spectroscopy in Biological Systems". In: *University of Illinois Press* (1969). DOI: citeulike-article-id:380320.

[123] Ken A. Dill. "Dominant Forces in Protein Folding". In: *Biochemistry* (1990). DOI: 10.1021/bi00483a001.

[124] Ken A. Dill and Justin L. MacCallum. *The protein-folding problem, 50 years on*. 2012. DOI: 10.1126/science.1219021.

[125] Hahnbeom Park, Philip Bradley, Per Greisen, Yuan Liu, Vikram Khipple Mulligan, David E. Kim, David Baker, and Frank Dimaio. "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules". In: *Journal of Chemical Theory and Computation* (2016). DOI: 10.1021/acs.jctc.6b00819.

[126] Brian Kuhlman and Philip Bradley. *Advances in protein structure prediction and design*. 2019. DOI: 10.1038/s41580-019-0163-x.

[127] James U. Bowie, Roland Lüthy, and David Eisenberg. "A method to identify protein sequences that fold into a known three-dimensional stucture". In: *Science* (1991). DOI: 10.1126/science.1853201.

[128] Janusz M. Bujnicki. *Protein-structure prediction by recombination of fragments*. 2006. DOI: 10.1002/cbic.200500235.

[129] Alexander M. Sevy, Tim M. Jacobs, James E. Crowe, and Jens Meiler. "Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences". In: *PLoS Computational Biology* 11.7 (2015). DOI: 10.1371/journal.pcbi.1004300.

[130] Marion F. Sauer, Alexander M. Sevy, James E. Crowe, and Jens Meiler. "Multi-state design of flexible proteins predicts sequences optimal for conformational change". In: *PLoS Computational Biology* (2020). DOI: 10.1371/journal.pcbi.1007339.

[131] E Humphris-Narayanan, E Akiva, R Varela, O Conchuir S, and T Kortemme. "Prediction of mutational tolerance in HIV-1 protease and reverse transcriptase using flexible backbone protein design". In: *PLoS Comput Biol* 8.8 (2012), e1002639. DOI: 10.1371/journal.pcbi.1002639.

[132] Elizabeth a Christian, Kristen M Kahle, Kimberly Mattia, Bridget a Puffer, Jennifer M Pfaff, Adam Miller, Cheryl Paes, Edgar Davidson, and Benjamin J Doranz. "Atomic-level functional model of dengue virus Envelope protein infectivity". In: *Proceedings of the National Academy of Sciences* 110.46 (2013), pp. 1–6. DOI: 10.1073/pnas.1310962110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1310962110.

[133] C A Smith and T Kortemme. "Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction". In: *J Mol Biol* 380.4 (2008), pp. 742–756. DOI: 10.1016/j.jmb.2008.05.023.

[134] Colin A. Smith and Tanja Kortemme. “Predicting the tolerated sequences for proteins and protein interfaces using rosettabackrub flexible backbone design”. In: *PLoS ONE* 6.7 (2011). DOI: 10.1371/journal.pone.0020451.

[135] S Y Rhee, M J Gonzales, R Kantor, B J Betts, J Ravela, and R W Shafer. “Human immunodeficiency virus reverse transcriptase and protease sequence database”. In: *Nucleic Acids Res* 31.1 (2003), pp. 298–303. DOI: 10.1093/nar/gkg100.

[136] G D Friedland, N A Lakomek, C Griesinger, J Meiler, and T Kortemme. “A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family”. In: *PLoS Comput Biol* 5.5 (2009), e1000393. DOI: 10.1371/journal.pcbi.1000393.

[137] James A. Davey and Roberto A. Chica. *Multistate approaches in computational protein design*. 2012. DOI: 10.1002/pro.2128.

[138] M Fromer and J M Shifman. “Tradeoff between stability and multispecificity in the design of promiscuous proteins”. In: *PLoS Comput Biol* 5.12 (2009), e1000627. DOI: 10.1371/journal.pcbi.1000627.

[139] Benjamin D. Allen and Stephen L. Mayo. “An efficient algorithm for multistate protein design based on faster”. In: *Journal of Computational Chemistry* 31.5 (2010), pp. 904–916. DOI: 10.1002/jcc.21375.

[140] Andrew Leaver-Fay, Ron Jacak, P. Benjamin Stranges, and Brian Kuhlman. “A generic program for multistate protein design”. In: *PLoS ONE* 6.7 (2011). DOI: 10.1371/journal.pone.0020937.

[141] A M Sevy, N C Wu, I M Gilchuk, E H Parrish, S Burger, D Yousif, M B M Nagel, K L Schey, I A Wilson, J E Crowe Jr., and J Meiler. “Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses”. In: *Proc Natl Acad Sci U S A* 116.5 (2019), pp. 1597–1602. DOI: 10.1073/pnas.1806004116.

[142] A. T. Heiny, Olivo Miotto, Kellathur N. Srinivasan, Asif M. Khan, G. L. Zhang, Vladimir Brusic, Tin Wee Tan, and J. Thomas August. “Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets”. In: *PLoS ONE* (2007). DOI: 10.1371/journal.pone.0001190.

[143] F Friedberg and A R Rhoads. “Evolutionary aspects of calmodulin”. In: *IUBMB Life* 51.4 (2001), pp. 215–221. DOI: 10.1080/152165401753311753.

[144] M Kobayashi, J Buck, and L R Levin. “Conservation of functional domain structure in bicarbonate-regulated ”soluble” adenylyl cyclases in bacteria and eukaryotes”. In: *Dev Genes Evol* 214.10 (2004), pp. 503–509. DOI: 10.1007/s00427-004-0432-2.

[145] R T Shealy, A D Murphy, R Ramarathnam, E Jakobsson, and S Subramaniam. “Sequence-function analysis of the K+-selective family of ion channels using a comprehensive alignment and the KcsA channel structure”. In: *Biophys J* 84.5 (2003), pp. 2929–2942. DOI: 10.1016/S0006-3495(03)70020-4.

[146] T Hrabe, Z Li, M Sedova, P Rotkiewicz, L Jaroszewski, and A Godzik. “PDBFlex: exploring flexibility in protein structures”. In: *Nucleic Acids Res* 44.D1 (2016), pp. D423–8. DOI: 10.1093/nar/gkv1316.

[147] A Stein and T Kortemme. "Improvements to robotics-inspired conformational sampling in rosetta". In: *PLoS One* 8.5 (2013), e63090. DOI: 10.1371/journal.pone.0063090.

[148] I Kufareva and R Abagyan. "Methods of protein structure comparison". In: *Methods Mol Biol* 857 (2012), pp. 231–257. DOI: 10.1007/978-1-61779-588-6_10.

[149] O Carugo and S Pongor. "A normalized root-mean-square distance for comparing protein three-dimensional structures". In: *Protein Sci* 10.7 (2001), pp. 1470–1473. DOI: 10.1110/ps.690101.

[150] H B Mann D.R. and Whitney. "On a test whether one or two random variables is stochastically larger than the other". In: *The Annals of Mathematical Statistics* 18.1 (), pp. 50–60.

[151] M G Kendall. *Rank Correlation Methods*. Ed. by C Griffin. 2nd ed. The University of California, 1948, p. 160.

[152] M Sternke, K W Tripp, and D Barrick. "Consensus sequence design as a general strategy to create hyperstable, biologically active proteins". In: *Proc Natl Acad Sci U S A* 116.23 (2019), pp. 11275–11284. DOI: 10.1073/pnas.1816707116.

[153] D B Halling, B J Liebeskind, A W Hall, and R W Aldrich. "Conserved properties of individual Ca2+-binding sites in calmodulin". In: *Proc Natl Acad Sci U S A* 113.9 (2016), E1216–25. DOI: 10.1073/pnas.1600385113.

[154] Y Bao, P Bolotov, D Dernovoy, B Kiryutin, L Zaslavsky, T Tatusova, J Ostell, and D Lipman. "The influenza virus resource at the National Center for Biotechnology Information". In: *J Virol* 82.2 (2008), pp. 596–601. DOI: 10.1128/JVI.02005-07.

[155] W I Weis, A T Brunger, J J Skehel, and D C Wiley. "Refinement of the influenza virus hemagglutinin by simulated annealing". In: *J Mol Biol* 212.4 (1990), pp. 737–761. DOI: 10.1016/0022-2836(90)90234-D.

[156] R J Russell, P S Kerry, D J Stevens, D A Steinhauer, S R Martin, S J Gamblin, and J J Skehel. "Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion". In: *Proc Natl Acad Sci U S A* 105.46 (2008), pp. 17736–17741. DOI: 10.1073/pnas.0807142105.

[157] J Chen, J J Skehel, and D C Wiley. "N- and C-terminal residues combine in the fusion-pH influenza hemagglutinin HA(2) subunit to form an N cap that terminates the triple-stranded coiled coil." In: *Proceedings of the National Academy of Sciences of the United States of America* 96.16 (Aug. 1999), pp. 8967–72.

[158] Per A. Bullough, Frederick M. Hughson, John J. Skehel, and Don C. Wiley. "Structure of influenza haemagglutinin at the pH of membrane fusion". In: *Nature* 371.6492 (Sept. 1994), pp. 37–43. DOI: 10.1038/371037a0.

[159] Yubin Zhou, Teryl K. Frey, and Jenny J. Yang. *Viral calciomics: Interplays between Ca2+ and virus*. 2009. DOI: 10.1016/j.ceca.2009.05.005.

[160] H Levene. *Robust tests for equality of variances*. Stanford University Press, p. 517.

[161] B. Student. "The probable error of a mean". In: *Biometrika* (1908). DOI: 10.1093/biomet/6.1.1.

[162] A Leaver-Fay, M Tyka, S M Lewis, O F Lange, J Thompson, R Jacak, K Kaufman, P D Renfrew, C A Smith, W Sheffler, I W Davis, S Cooper, A Treuille, D J Mandell, F Richter, Y E Ban, S J Fleishman, J E Corn, D E Kim, S Lyskov, M Berrondo, S Mentzer, Z Popovic, J J Havranek, J Karanicolas, R Das, J Meiler, T Kortemme, J J Gray, B Kuhlman, D Baker, and P Bradley. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules". In: *Methods Enzymol* 487 (2011), pp. 545–574. DOI: 10.1016/B978-0-12-381270-4.00019-6.

[163] A Zemla. "LGA: A method for finding 3D similarities in protein structures". In: *Nucleic Acids Res* 31.13 (2003), pp. 3370–3374. DOI: 10.1093/nar/gkg571.

[164] R A Abagyan and M M Totrov. "Contact area difference (CAD): a robust measure to evaluate accuracy of protein models". In: *J Mol Biol* 268.3 (1997), pp. 678–685. DOI: 10.1006/jmbi.1997.0994.

[165] B Marsden and R Abagyan. "SAD–a normalized structural alignment database: improving sequence-structure alignments". In: *Bioinformatics* 20.15 (2004), pp. 2333–2344. DOI: 10.1093/bioinformatics/bth244.

[166] Kathleen F. O'Rourke, Scott D. Gorman, and David D. Boehr. *Biophysical and computational methods to analyze amino acid interaction networks in proteins*. 2016. DOI: 10.1016/j.csbj.2016.06.002.

[167] O F Lange, N A Lakomek, C Fares, G F Schroder, K F Walter, S Becker, J Meiler, H Grubmuller, C Griesinger, and B L de Groot. "Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution". In: *Science* 320.5882 (2008), pp. 1471–1475. DOI: 10.1126/science.1157092.

[168] James A. Davey and Roberto A. Chica. "Multistate Computational Protein Design with Backbone Ensembles". In: 2017, pp. 161–179. DOI: 10.1007/978-1-4939-6637-0_7.

[169] J R Brender, D Shultis, N A Khattak, and Y Zhang. "An Evolution-Based Approach to De Novo Protein Design". In: *Methods Mol Biol* 1529 (2017), pp. 243–264. DOI: 10.1007/978-1-4939-6637-0_12.

[170] F Tsai, P J Homan, H Agrawal, A V Misharin, H Abdala-Valencia, G K Haines 3rd, S Dominguez, C L Bloomfield, R Saber, A Chang, C Mohan, J Hutcheson, A Davidson, G R S Budinger, P Bouillet, A Dorfleutner, C Stehlik, D R Winter, C M Cuda, and H Perlman. "Bim suppresses the development of SLE by limiting myeloid inflammatory responses". In: *J Exp Med* 214.12 (2017), pp. 3753–3773. DOI: 10.1084/jem.20170479.

[171] Bargavi Thyagarajan and Jesse D. Bloom. "The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin". In: *eLife* 3 (2014). DOI: 10.7554/eLife.03300.

[172] F Morcos, A Pagnani, B Lunt, A Bertolino, D S Marks, C Sander, R Zecchina, J N Onuchic, T Hwa, and M Weigt. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". In: *Proc Natl Acad Sci U S A* 108.49 (2011), E1293–301. DOI: 10.1073/pnas.1111471108.

[173] G B Gloor, L C Martin, L M Wahl, and S D Dunn. "Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions". In: *Biochemistry* 44.19 (2005), pp. 7156–7165. DOI: 10.1021/bi050293e.

[174] E R Tillier and T W Lui. “Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments”. In: *Bioinformatics* 19.6 (2003), pp. 750–755. DOI: 10.1093/bioinformatics/btg072.

[175] L C Martin, G B Gloor, S D Dunn, and L M Wahl. “Using information theory to search for co-evolving residues in proteins”. In: *Bioinformatics* 21.22 (2005), pp. 4116–4124. DOI: 10.1093/bioinformatics/bti671.

[176] U Gobel, C Sander, R Schneider, and A Valencia. “Correlated mutations and residue contacts in proteins”. In: *Proteins* 18.4 (1994), pp. 309–317. DOI: 10.1002/prot.340180402.

[177] O Olmea, B Rost, and A Valencia. “Effective use of sequence correlation and conservation in fold recognition”. In: *J Mol Biol* 293.5 (1999), pp. 1221–1239. DOI: 10.1006/jmbi.1999.3208.

[178] D de Juan, F Pazos, and A Valencia. “Emerging methods in protein co-evolution”. In: *Nat Rev Genet* 14.4 (2013), pp. 249–261. DOI: 10.1038/nrg3414.

[179] S A Combs, S L Deluca, S H Deluca, G H Lemmon, D P Nannemann, E D Nguyen, J R Willis, J H Sheehan, and J Meiler. “Small-molecule ligand docking into comparative models with Rosetta”. In: *Nat Protoc* 8.7 (2013), pp. 1277–1298. DOI: 10.1038/nprot.2013.074.

[180] G E Crooks, G Hon, J M Chandonia, and S E Brenner. “WebLogo: a sequence logo generator”. In: *Genome Res* 14.6 (2004), pp. 1188–1190. DOI: 10.1101/gr.849004.

[181] Jordan R. Willis, Bryan S. Briney, Samuel L. DeLuca, James E. Crowe, and Jens Meiler. “Human Germline Antibody Gene Segments Encode Polyspecific Antibodies”. In: *PLoS Computational Biology* 9.4 (2013). DOI: 10.1371/journal.pcbi.1003045.

[182] N A Weiss. *A Course in Probability*. Addison-Wesley.

[183] L Y Yampolsky and A Stoltzfus. “The exchangeability of amino acids in proteins”. In: *Genetics* 170.4 (2005), pp. 1459–1472. DOI: 10.1534/genetics.104.039107.

[184] F Sievers, A Wilm, D Dineen, T J Gibson, K Karplus, W Li, R Lopez, H McWilliam, M Remmert, J Soding, J D Thompson, and D G Higgins. “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega”. In: *Mol Syst Biol* 7 (2011), p. 539. DOI: 10.1038/msb.2011.75.

[185] F Sievers and D G Higgins. “Clustal Omega for making accurate alignments of many protein sequences”. In: *Protein Sci* 27.1 (2018), pp. 135–145. DOI: 10.1002/pro.3290.

[186] E Durham, B Dorr, N Woetzel, R Staritzbichler, and J Meiler. “Solvent accessible surface area approximations for rapid and accurate protein structure prediction”. In: *J Mol Model* 15.9 (2009), pp. 1093–1108. DOI: 10.1007/s00894-009-0454-9.

[187] S Ovchinnikov, H Kamisetty, and D Baker. “Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information”. In: *Elife* 3 (2014), e02030. DOI: 10.7554/eLife.02030.

[188] T. P. Hopp and K. R. Woods. “Prediction of protein antigenic determinants from amino acid sequences”. In: *Proceedings of the National Academy of Sciences of the United States of America* (1981). DOI: 10.1073/pnas.78.6.3824.

[189] A. S. Kolaskar and Prasad C. Tongaonkar. “A semi-empirical method for prediction of antigenic determinants on protein antigens”. In: *FEBS Letters* (1990). DOI: 10.1016/0014-5793(90)80535-Q.

[190] Urmila Kulkarni-Kale, Shriram Bhosle, and A. S. Kolaskar. “CEP: A conformational epitope prediction server”. In: *Nucleic Acids Research* (2005). DOI: 10.1093/nar/gki460.

[191] Martin J. Blythe and Darren R. Flower. “Benchmarking B cell epitope prediction: Underperformance of existing methods”. In: *Protein Science* (2009). DOI: 10.1110/ps.041059505.

[192] Edgar Ernesto Gonzalez Kozlova, Loïc Cerf, Francisco Santos Schneider, Benjamin Thomas Viart, Christophe NGuyen, Bethina Trevisol Steiner, Sabrina de Almeida Lima, Franck Molina, Clara Guerra Duarte, Liza Felicori, Carlos Chávez-Olórtegui, and Ricardo Andrez Machado-de-Ávila. “Computational B-cell epitope identification and production of neutralizing murine antibodies against Atroxlysin-I”. In: *Scientific Reports* (2018). DOI: 10.1038/s41598-018-33298-x.

[193] Wen Zhang, Yi Xiong, Meng Zhao, Hua Zou, Xinghuo Ye, and Juan Liu. “Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature”. In: *BMC Bioinformatics* (2011). DOI: 10.1186/1471-2105-12-341.

[194] Nimrod D. Rubinstein, Itay Mayrose, Eric Martz, and Tal Pupko. “Epitopia: A web-server for predicting B-cell epitopes”. In: *BMC Bioinformatics* (2009). DOI: 10.1186/1471-2105-10-287.

[195] Yao Lian, Meng Ge, and Xian Ming Pan. “EPMLR: Sequence-based linear B-cell epitope prediction method using multiple linear regression”. In: *BMC Bioinformatics* (2014). DOI: 10.1186/s12859-014-0414-y.

[196] Julia Ponomarenko, Nikitas Papangelopoulos, Dirk M. Zajonc, Bjoern Peters, Alessandro Sette, and Philip E. Bourne. “IEDB-3D: Structural data within the immune epitope database”. In: *Nucleic Acids Research* (2011). DOI: 10.1093/nar/gkq888.

[197] Jason A. Greenbaum, Pernille Haste Andersen, Martin Blythe, Huynh Hoa Bui, Raul E. Cachau, James Crowe, Matthew Davies, A. S. Kolaskar, Ole Lund, Sherrie Morrison, Brendan Mumey, Yanay Ofran, Jean Luc Pellequer, Clemencia Pinilla, Julia V. Ponomarenko, G. P.S. Raghava, Marc H.V. Van Regenmortel, Erwin L. Roggen, Alessandro Sette, Avner Schlessinger, Johannes Sollner, Martin Zand, and Bjoern Peters. *Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools*. 2007. DOI: 10.1002/jmr.815.

[198] Inbal Sela-Culang, Vered Kunik, and Yanay Ofran. “The structural basis of antibody-antigen recognition”. In: *Frontiers in Immunology* (2013). DOI: 10.3389/fimmu.2013.00302.

[199] Jeong Hyun Lee, Raiees Andrabi, Ching Yao Su, Anila Yasmeen, Jean Philippe Julien, Leopold Kong, Nicholas C. Wu, Ryan McBride, Devin Sok, Matthias Pauthner, Christopher A. Cottrell, Travis Nieusma, Claudia Blattner, James C. Paulson, Per Johan Klasse, Ian A. Wilson, Dennis R. Burton, and Andrew B. Ward. “A Broadly Neutralizing Antibody Targets the Dynamic HIV Envelope Trimer Apex via a Long, Rigidified, and Anionic β-Hairpin Structure”. In: *Immunity* (2017). DOI: 10.1016/j.immuni.2017.03.017.

[200] Peter D Kwong and Ian a Wilson. “HIV-1 and influenza antibodies: seeing antigens in new ways.” In: *Nature immunology* 10.6 (2009), pp. 573–578. DOI: 10.1038/ni.1746.

[201] K. N. Trueblood, H. B. Bürgi, H. Burzlaff, J. D. Dunitz, C. M. Gramaccioli, H. H. Schulz, U. Shmueli, and S. C. Abrahams. “Atomic displacement parameter nomenclature report of a subcommittee on atomic displacement parameter nomenclature”. In: *Acta Crystallographica Section A: Foundations of Crystallography* (1996). DOI: 10.1107/S0108767396005697.

[202] Guillaume Brysbaert, Théo Mauri, Jérôme de Ruyck, and Marc F. Lensink. “Identification of Key Residues in Proteins Through Centrality Analysis and Flexibility Prediction with RINspector”. In: *Current Protocols in Bioinformatics* (2019). DOI: 10.1002/cpbi.66.

[203] Aaron R. Dinner, Andrej Šalib, Lorna J. Smitha, Christopher M. Dobsona, and Martin Karplus. *Understanding protein folding via free-energy surfaces from theory and experiment*. 2000. DOI: 10.1016/S0968-0004(00)01610-8.

[204] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O’Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design”. In: *Journal of Chemical Theory and Computation* (2017). DOI: 10.1021/acs.jctc.7b00125.

[205] Randi Vita, Swapnil Mahajan, James A. Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R. Cantrell, Daniel K. Wheeler, Alessandro Sette, and Bjoern Peters. “The Immune Epitope Database (IEDB): 2018 update”. In: *Nucleic Acids Research* (2019). DOI: 10.1093/nar/gky1006.

[206] William B. Messer, Ruklanthi De Alwis, Boyd L. Yount, Scott R. Royal, Jeremy P. Huynh, Scott A. Smith, James E. Crowe, Benjamin J. Doranz, Kristen M. Kahle, Jennifer M. Pfaff, Laura J. White, Carlos A. Sariol, Aravinda M. De Silva, and Ralph S. Baric. “Dengue virus envelope protein domain I/II hinge determines long-lived serotype-specific dengue immunity”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2014). DOI: 10.1073/pnas.1317350111.

[207] Vicente Mas, Harish Nair, Harry Campbell, Jose A. Melero, and Thomas C. Williams. “Antigenic and sequence variability of the human respiratory syncytial virus F glycoprotein compared to related viruses in a comprehensive dataset”. In: *Vaccine* (2018). DOI: 10.1016/j.vaccine.2018.09.056.

[208] K. Pearson. “VII. Note on regression and inheritance in the case of two parents”. In: *Proceedings of the Royal Society of London* (1895). DOI: 10.1098/rspl.1895.0041.

[209] R. A. Fisher. “Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population”. In: *Biometrika* (1915). DOI: 10.2307/2331838.

[210] E. Forgy. “Cluster analysis of multivariate data : efficiency versus interpretability of classifications”. In: *Biometrics* (1965).

[211] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. 1996.

[212] M. Girvan and M. E.J. Newman. “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2002). DOI: 10.1073/pnas.122653799.

[213] M. T. Swulius and M. N. Waxham. *Ca2+/calmodulin-dependent protein kinases*. 2008. DOI: 10.1007/s00018-008-8086-2.

[214] D. H. Ross and H. L. Cardenas. “Calmodulin Stimulation of Ca2+-Dependent ATP Hydrolysis and ATP-Dependent Ca2+ Transport in Synaptic Membranes”. In: *Journal of Neurochemistry* (1983). DOI: 10.1111/j.1471-4159.1983.tb11828.x.

[215] Yoon Park Hye, Sally A. Kim, Jonas Korlach, Elizabeth Rhoades, Lisa W. Kwok, Warren R. Zipfel, M. Neal Waxham, Watt W. Webb, and Lois Pollack. “Conformational changes of calmodulin upon Ca2+ binding studied with a microfluidic mixer”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2008). DOI: 10.1073/pnas.0710810105.

[216] Maxim V. Shapovalov and Roland L. Dunbrack. “A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions”. In: *Structure* (2011). DOI: 10.1016/j.str.2011.03.019.

[217] Zhichao Miao and Yang Cao. “Quantifying side-chain conformational variations in protein structure”. In: *Scientific Reports* (2016). DOI: 10.1038/srep37024.

[218] Shuangye Yin, Feng Ding, and Nikolay V. Dokholyan. *Eris: An automated estimator of protein stability [2]*. 2007. DOI: 10.1038/nmeth0607-466.

[219] Eili Y. Klein, Deena Blumenkrantz, Adrian Serohijos, Eugene Shakhnovich, Jeong-Mo Choi, João V. Rodrigues, Brendan D. Smith, Andrew P. Lane, Andrew Feldman, and Andrew Pekosz. “Stability of the Influenza Virus Hemagglutinin Protein Correlates with Evolutionary Dynamics”. In: *mSphere* (2018). DOI: 10.1128/mspheredirect.00554-17.

[220] R. J. Russell, S. J. Gamblin, L. F. Haire, D. J. Stevens, B. Xiao, Y. Ha, and J. J. Skehel. “H1 and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes”. In: *Virology* (2004). DOI: 10.1016/j.virol.2004.04.040.

[221] Frederick P. Schwarz, Diana Tello, Fernando A. Goldbaum, Roy A. Mariuzza, and Roberto J. Poljak. “Thermodynamics of Antigen-antibody Binding using Specific Anti-lysozyme Antibodies”. In: *European Journal of Biochemistry* (1995). DOI: 10.1111/j.1432-1033.1995.00388.x.

[222] Roberto Reverberi and Lorenzo Reverberi. “Factors affecting the antigen-antibody reaction”. In: *Blood Transfusion* (2007). DOI: 10.2450/2007.0047-07.

[223] Hiroki Akiba and Kouhei Tsumoto. *Thermodynamics of antibody - Antigen interaction revealed by mutation analysis of antibody variable regions*. 2015. DOI: 10.1093/jb/mvv049.

[224] Bruno E. Correia, John T. Bates, Rebecca J. Loomis, Gretchen Baneyx, Chris Carrico, Joseph G. Jardine, Peter Rupert, Colin Correnti, Oleksandr Kalyuzhniy, Vinayak Vittal, Mary J. Connell, Eric Stevens, Alexandria Schroeter, Man Chen, Skye MacPherson, Andreia M. Serra, Yumiko Adachi, Margaret A. Holmes, Yuxing Li, Rachel E. Klevit, Barney S. Graham, Richard T. Wyatt, David Baker, Roland K. Strong, James E. Crowe, Philip R. Johnson, and William R. Schief. “Proof of principle for epitope-focused vaccine design”. In: *Nature* (2014). DOI: 10.1038/nature12966.

[225] Fabian Sesterhenn, Marie Galloux, Sabrina S. Vollers, Lucia Csepregi, Che Yang, Delphyne Descamps, Jaume Bonet, Simon Friedensohn, Pablo Gainza, Patricia Corthésy, Man Chen, Stéphane Rosset, Marie Anne Rameix-Welti, Jean François Éléouët, Sai T. Reddy, Barney S. Graham, Sabine Riffault, and Bruno E. Correia. “Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen”. In: *PLoS Biology* (2019). DOI: 10.1371/journal.pbio.3000164.

[226] Caroline Breese Hall, Geoffrey A. Weinberg, Marika K. Iwane, Aaron K. Blumkin, Kathryn M. Edwards, Mary A. Staat, Peggy Auinger, Marie R. Griffin, Katherine A. Poehling, Dean Erdman, Carlos G. Grijalva, Yuwei Zhu, and Peter Szilagyi. “The Burden of respiratory syncytial virus infection in young children”. In: *New England Journal of Medicine* (2009). DOI: 10.1056/NEJMoa0804877.

[227] Deena Shefali-Patel, Mireia Alcazar Paris, Fran Watson, Janet L. Peacock, Morag Campbell, and Anne Greenough. “RSV hospitalisation and healthcare utilisation in moderately prematurely born infants”. In: *European Journal of Pediatrics* (2012). DOI: 10.1007/s00431-012-1673-0.

[228] Harish Nair, D. James Nokes, Bradford D. Gessner, Mukesh Dherani, Shabir A. Madhi, Rosalyn J. Singleton, Katherine L. O’Brien, Anna Roca, Peter F. Wright, Nigel Bruce, Aruna Chandran, Evropi Theodoratou, Agustinus Sutanto, Endang R. Sedyaningsih, Mwanajuma Ngama, Patrick K. Munywoki, Cissy Kartasasmita, Eric AF Simões, Igor Rudan, Martin W. Weber, and Harry Campbell. “Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis”. In: *The Lancet* (2010). DOI: 10.1016/S0140-6736(10)60206-1.

[229] Jason S. McLellan. *Neutralizing epitopes on the respiratory syncytial virus fusion glycoprotein*. 2015. DOI: 10.1016/j.coviro.2015.03.002.

[230] Edward M. Connor. “Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants”. In: *Pediatrics* (1998). DOI: 10.1542/peds.102.3.531.

[231] Gregory M. Glenn, Louis F. Fries, D. Nigel Thomas, Gale Smith, Eloi Kpamegan, Hanxin Lu, David Flyer, Dewal Jani, Somia P. Hickman, and Pedro A. Piedra. “A randomized, blinded, controlled, dose-ranging study of a respiratory syncytial virus recombinant Fusion (F) nanoparticle vaccine in healthy women of childbearing age”. In: *Journal of Infectious Diseases* (2016). DOI: 10.1093/infdis/jiv406.

[232] Gale Smith, Rama Raghunandan, Yingyun Wu, Ye Liu, Michael Massare, Margret Nathan, Bin Zhou, Hanxin Lu, Sarathi Boddapati, Jingning Li, David Flyer, and Gregory Glenn. “Respiratory Syncytial Virus Fusion Glycoprotein Expressed in Insect Cells Form Protein Nanoparticles That Induce Protective Immunity in Cotton Rats”. In: *PLoS ONE* (2012). DOI: 10.1371/journal.pone.0050852.

[233] Rama Raghunandan, Hanxin Lu, Bin Zhou, Mimi Guebre Xabier, Michael J. Massare, David C. Flyer, Louis F. Fries, Gale E. Smith, and Gregory M. Glenn. “An insect cell derived respiratory syncytial virus (RSV) F nanoparticle vaccine induces antigenic site II antibodies and protects against RSV challenge in cotton rats by active and passive immunization”. In: *Vaccine* (2014). DOI: 10.1016/j.vaccine.2014.09.030.

[234] Sheng Jiun Wu, Albert Albert Schmidt, Eric J. Beil, Nicole D. Day, Patrick J. Branigan, Changbao Liu, Lester L. Gutshall, Concepción Palomo, Julie Furze, Geraldine Taylor, José A. Melero, Ping Tsui, Alfred M. Del Vecchio, and Marian Kruszynski. “Characterization of the epitope for anti-human respiratory syncytial virus F protein monoclonal antibody 101F using synthetic peptides and genetic approaches”. In: *Journal of General Virology* (2007). DOI: 10.1099/vir.0.82753-0.

[235] James E. Crowe, Cai Yen Firestone, Roberta Crim, Judy A. Beeler, Kathleen L. Coelingh, Carlos F. Barbas, Dennis R. Burton, Robert M. Chanock, and Brian R. Murphy. “Monoclonal antibody-resistant mutants selected with a respiratory syncytial virus-neutralizing human antibody Fab fragment (Fab 19) define a unique epitope on the fusion (F) glycoprotein”. In: *Virology* (1998). DOI: 10.1006/viro.1998.9462.

[236] James E. Crowe, Jr., Page S. Gilmour, Brian R. Murphy, Robert M. Chanock, Lingxun Duan, Roger J. Pomerantz, and Glenn R. Pilkington. “Isolation of a Second Recombinant Human Respiratory Syncytial Virus Monoclonal Antibody Fragment (Fab RSVF2–5) that Exhibits Therapeutic Efficacy In Vivo”. In: *The Journal of Infectious Diseases* (1998). DOI: 10.1086/517397.

[237] Herren Wu, David S. Pfarr, Syd Johnson, Yambasu A. Brewah, Robert M. Woods, Nita K. Patel, Wendy I. White, James F. Young, and Peter A. Kiener. “Development of Motavizumab, an Ultra-potent Antibody for the Prevention of Respiratory Syncytial Virus Infection in the Upper and Lower Respiratory Tract”. In: *Journal of Molecular Biology* (2007). DOI: 10.1016/j.jmb.2007.02.024.

[238] J A Beeler and K van Wyke Coelingh. “Neutralization epitopes of the F glycoprotein of respiratory syncytial virus: effect of mutation upon fusion function.” In: *Journal of virology* (1989).

[239] Anders Krarup, Daphné Truan, Polina Furmanova-Hollenstein, Lies Bogaert, Pascale Bouchier, Ilona J.M. Bisschop, Myra N. Widjojoatmodjo, Roland Zahn, Hanneke Schuitemaker, Jason S. McLellan, and Johannes P.M. Langedijk. “A highly stable prefusion RSV F vaccine derived from structural analysis of the fusion mechanism”. In: *Nature Communications* (2015). DOI: 10.1038/ncomms9143.

[240] Larry J. Anderson, John C. Hierholzer, Cecilia Tsou, R. Michael Hendry, Bruce F. Fernie, Yvonne Stone, and Kenneth McIntosh. “Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies”. In: *Journal of Infectious Diseases* (1985). DOI: 10.1093/infdis/151.4.626.

[241] Xiaodong Zhao, Fu Ping Chen, and Wayne M. Sullender. “Respiratory syncytial virus escape mutant derived in vitro resists palivizumab prophylaxis in cotton rats”. In: *Virology* (2004). DOI: 10.1016/j.virol.2003.10.018.

[242] X. Zhao and W. M. Sullender. “In Vivo Selection of Respiratory Syncytial Viruses Resistant to Palivizumab”. In: *Journal of Virology* (2005). DOI: 10.1128/jvi.79.7.3962-3968.2005.

[243] Xiao Dong Zhao, W. Stillender, and Xi Qiang Yang. “Variable resistance to Palivizumab in cotton rats by respiratory syncytial virus escape mutants”. In: *Chinese Journal of Microbiology and Immunology* (2005). DOI: 10.1086/425515.

[244] Morgan S.A. Gilman, Syed M. Moin, Vicente Mas, Man Chen, Nita K. Patel, Kari Kramer, Qing Zhu, Stephanie C. Kabeche, Azad Kumar, Concepción Palomo, Tim Beaumont, Ulrich Baxa, Nancy D. Ulbrandt, José A. Melero, Barney S. Graham, and Jason S. McLellan. “Characterization of a Prefusion-Specific Antibody That Recognizes a Quaternary, Cleavage-Dependent Epitope on the RSV Fusion Glycoprotein”. In: *PLoS Pathogens* (2015). DOI: 10.1371/journal.ppat.1005035.

[245] Xiaozhou Luo, Tao Liu, Ying Wang, Haiqun Jia, Yuhan Zhang, Dawna Caballero, Juanjuan Du, Rongsheng E. Wang, Danling Wang, Peter G. Schultz, and Feng Wang. “An Epitope-Specific Respiratory Syncytial Virus Vaccine Based on an Antibody Scaffold”. In: *Angewandte Chemie - International Edition* (2015). DOI: 10.1002/anie.201507928.

[246] Xiaolin Wen, Jennifer Pickens, Jarrod J. Mousa, George P. Leser, Robert A. Lamb, James E. Crowe, and Theodore S. Jardetzky. “A Chimeric Pneumovirus Fusion Protein Carrying Neutralizing Epitopes of Both MPV and RSV”. In: *PLoS ONE* (2016). DOI: 10.1371/journal.pone.0155917.

[247] Xiaocong Yu, Patricia A. McGraw, Frances S. House, and James E. Crowe. “An optimized electrofusion-based protocol for generating virus-specific human monoclonal antibodies”. In: *Journal of Immunological Methods* (2008). DOI: 10.1016/j.jim.2008.04.008.

[248] George M. Sheldrick. *A short history of SHELX*. 2008. DOI: 10.1107/S0108767307043930.

[249] Paul D. Adams, Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, Li Wei Hung, Gary J. Kapral, Ralf W. Grosse-Kunstleve, Airlie J. McCoy, Nigel W. Moriarty, Robert Oeffner, Randy J. Read, David C. Richardson, Jane S. Richardson, Thomas C. Terwilliger, and Peter H. Zwart. “PHENIX: A comprehensive Python-based system for macromolecular structure solution”. In: *Acta Crystallographica Section D: Biological Crystallography* (2010). DOI: 10.1107/S0907444909052925.

[250] Paul Emsley and Kevin Cowtan. “Coot: Model-building tools for molecular graphics”. In: *Acta Crystallographica Section D: Biological Crystallography* (2004). DOI: 10.1107/S0907444904019158.

[251] Wolfgang Kabsch, Brünger A. T., Diederichs K., Karplus P. A., Diederichs K., McSweeney S., Ravelli R. B. G., Evans P., French S., Wilson K., Kabsch W., Kabsch W., Kabsch W., Kabsch W., Kursula P., and Weiss M. S. “<i>XDS</i>”. In: *Acta Crystallographica Section D Biological Crystallography* (2010). DOI: 10.1107/S0907444909047337.

[252] Guang Tang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. “EMAN2: An extensible image processing suite for electron microscopy”. In: *Journal of Structural Biology* (2007). DOI: 10.1016/j.jsb.2006.05.009.

[253] Tanvir R. Shaikh, Haixiao Gao, William T. Baxter, Francisco J. Asturias, Nicolas Boisset, Ardean Leith, and Joachim Frank. “Spider image processing for single-particle reconstruction of biological macromolecules from electron micrographs”. In: *Nature Protocols* (2008). DOI: 10.1038/nprot.2008.156.

[254] A. M. Sevy, J. F. Healey, W. Deng, P. C. Spiegel, S. L. Meeks, and R. Li. “Epitope mapping of inhibitory antibodies targeting the C2 domain of coagulation factor VIII by hydrogen-deuterium exchange mass spectrometry”. In: *Journal of Thrombosis and Haemostasis* (2013). DOI: 10.1111/jth.12433.

APPENDICES

APPENDIX A

RECON PROTOCOL CAPTURE

Introduction

The following protocol capture describes how to run RECON multi-state design and single-state design and the analyses performed that were discussed within the manuscript. For simplicity, we use only one of the eight protein ensembles included in the benchmark, dengue virus envelope (DV E) protein, as a case example to run all scripts and analyses.

All native structures, relaxed structures, Rosetta scripts, and other analysis scripts used in this benchmark can be downloaded from https://github.com/mfsauer/RECON_flexible_sequences.

Dependencies

All Rosetta commands for this publication were run with version 6b77f113505c4687d084d54890b1027ff308330d, from March 2016. Note that all analysis scripts will only function properly if they are in the correct directory as provided.

Several scripts used in this protocol require Python, either Python 2.7 or Python 3.7, and the required version is noted for each script. Additionally, several scripts require the Biopython package (https://github.com/biopython/biopython.github.io/). It is recommended that the user have both versions of Python and the Biopython package installed prior to using this protocol.

To generate sequences profiles, the user can install WebLogo locally either using `pip` or downloaded manually from https://github.com/WebLogo/weblogo. Multiple sequence alignments require Clustal Omega —for generating alignments of input structures or designs, the user may use the online version found at https://www.ebi.ac.uk/Tools/msa/clustalo/ or download a local version found at http://www.clustal.org/omega/#Download.

For plotting, it is recommended that the user have R with the following packages installed into their library: ggplot2, cowplot, ape, broom, dendextend, dendsort, ggdendro, ggpubr, ggrepel, ggridges, ggsignif, Hmisc, Kendall, latex2exp, plotly, reshape2, stats, and treeio.

Structure Preparation

All structures (1OAN, 1OK8, 3C5X, 3C6E, 3J27, and 3J2P) were downloaded from the Protein DataBank (PDB; www.rcsb.org) and manually processed in PyMol to remove all waters and non-protein atoms. The FASTA sequence of each chain was generated using

```
python2.7 get_fasta_from_pdb.py 1oan A > 1oan_A.fa
```

DV E protein is a single chain, but for multi-chain proteins the individual `.fa` files were concatenated to form a single `.fa` file of the whole protein. Using the aligned sequences, any residues not aligned at either the N- or C-termini were removed from the original PDB file using PyMol and saved as the native PDB for relaxation and design. For all other protein ensembles, any gaps in sequence alignment were excluded from design. However, for DV E structures 1OK8, 3C5X, and 3C6E, missing densities were replaced (so that the entire E monomer could be designed) using the Rosetta Partial Thread application as follows:

A grishin file was generated for each of the three DV E structures to define where the missing densities were located. Again, for clarity, we describe only the partial thread procedure for 3C6E, but the same protocol was applied in all cases. Below is the Rosetta partial thread application script along with the needed files. The sequence from 1OAN was used as the threaded sequence to fill any gaps in sequence for Any text denoted between two —– indicates a separate file containing the contents between the two lines with the file name indicated on the top line. Any \\ notation indicates that the command line continues with no return.

```
/path/to/rosetta/main/source/bin/partial_thread.default. \\
linuxgccrelease \\
-in:file:fasta 1oan_A.fasta -in:file:alignment 3c6e.grishin \\
-in:file:template 3c6e.pdb

---------------------------3c6e.grishin---------------------------
## 1oanA 3c6eA
# hhsearch
scores_from_program 0 1.00
0 MRCIGISNRDFVEGVSGGSWVDIVLEHGSCVTTMAKNKPTLDFELIKTEAKQPATLRKYCIEAK
LTNTTTESRCPTQGEPTLNEEQDKRFVCKHSMVDRGWGNGCGLFGKGGIVTCAMFTCKKNMEGKIV
QPENLEYTVVITPHSGEEHAVGNDTGKHGKEVKITPQSSITEAELTGYGTVTMECSPRTGLDFNEM
VLLQMKDKAWLVTHRQWFLDLPLPWLPGADQGSNWIQKETLVTFKNPHAKKQDVVVLGSQEGAMHT
ALTGATEIQMSSGNLLFTGHLKCRLRMDKLQLKGMSYSMCTGKFKVVKEIAETQHGTIVIRVQYEG
DGSPCKIPFEIMDLEKRHVLGRLITVNPIVTEKDSPVNIEAEPPFGDSYIIIGVEPGQLKLNWFKK
0 MRCIGMSNRDFVEGVSGGSWVDIVLEHGSCVTTMAKNKPTLDFELIKTEAKQPATLRKYCIEAK
LTNTTTESRCPTQGEPTLNEEQDKRFVCKHSMVDRGWGNGCGLFGKGGIVTCAMFTCKKNMEGKIV
QPENLEYTIVITPHSGEEHA-----GKHGKEIKITPQSSITEAELTGYGTVTMECSPRT-LDFNEM
VLLQMENKAWLVHRQWFLDLPLPWLPGADTQGSNWIQKETLVTFKNPHAKKQDVVVLGSQEGAMHT
ALTGATEIQMSSGNLLFTGHLKCRLRMDKLQLKGMSYSMCTGKFKVVKEIAETQHGTIVIRVQY-G
DGSPCKIPFEIMDLEKRHVLGRLITVNPIVTEKDSPVNIEAEPPFGDSYIIIGVEPGQLKLNWFKK
--
-----------------------------------------------------------------

# After running rename output model
mv 3c6e.pdb.pdb 3c6e_threaded.pdb
```

The partial thread application maps the missing sequence to the template structure, `3c6e.pdb`, containing the missing densities. From the threaded model, the missing backbone and side chains are rebuilt using the Rosetta application RosettaCM hybridize

```
/path/to/rosetta/main/source/bin/rosetta_scripts.default. \\
linuxgccrelease \\
@rosetta_cm.options -s 3c6e_threaded.pdb
```

```
--------------------rosetta_cm.options-----------------------
-database /path/to/rosetta/main/database
-parser:protocol hybridize.xml
-default_max_cycles 200
-dualspace
-------------------------------------------------------------

----------------------hybridize.xml--------------------------
<ROSETTASCRIPTS>
<SCOREFXNS>
    <stage1 weights=score3 symmetric=0>
        <Reweight scoretype=atom_pair_constraint weight=0.5/>
    </stage1>
    <stage2 weights=score4_smooth_cart symmetric=0>
        <Reweight scoretype=atom_pair_constraint weight=0.5/>
    </stage2>
    <fullatom weights=talaris2013_cart symmetric=0>
        <Reweight scoretype=atom_pair_constraint weight=0.5/>
    </fullatom>
</SCOREFXNS>
<MOVERS>
<Hybridize name=hybridize stage1_scorefxn=stage1 \\
stage2_scorefxn=stage2 fa_scorefxn=fullatom batch=1 \\
stage1_increase_cycles=1.0 stage2_increase_cycles=1.0>
<Template pdb=”3c6e_threadedl.pdb” cst_file=”AUTO” \\
weight=1.000 />
</Hybridize>
</MOVERS>
<PROTOCOLS>
    <Add mover=hybridize/>
</PROTOCOLS>
<OUTPUT scorefxn=talaris2013 />
</ROSETTASCRIPTS>
-----------------------------------------------------------
```

Refinement of input structures

All native structures were subject to a constrained FastRelax prior to design:

```
/path/to/rosetta/main/source/bin/rosetta_scripts.default. \\
linuxgccrelease \\
@relax.flags -s 3c6e_rebuilt.pdb -scorefile 3c6e_relaxed.fasc

--------------relax.flags--------------
-database /path/to/rosetta/main/database/
```

```
-linmem_ig 10
-in:file:fullatom
-in:detect_disulf false
-relax:fast
-relax:constrain_relax_to_start_coords
-out:file:fullatom
-out:suffix _relax
-use_input_sc
-nstruct 100
----------------------------------------
```

The lowest scoring model ranked by total score was chosen as the relaxed model, labeled `*_relaxed.pdb`.

Design of ensembles

For each ensemble, a resfile was created to specifiy which residues were to be considered for design. In either RECON multi-specificity or single-state design, the same resfile was used. For design of the DV E ensemble design, residues 1-394 were considered for design, and all six PDB files contained Chain A of the E monomer, which were used for design. Supplementary Table 1 lists the number, chain, and native residue considered for design for each PDB file —the matched number and chain correspond to the first and second column of each resfile. If the chain or residue numbering differed between PDB files of the same protein ensemble, separate resfiles were created for each protein. Although not relevant in the DV E example, it is paramount that the same number of positions are listed in each resfile and that each position listed in the same order for all conformations/PDB files, should there be more than one resfile needed per ensemble. This is because each matching/equivalent position between resfiles will be modeled as a matching side chain state within an ensemble. Below is an example of the start and end of a resfile in this format:

```
---------------denvE.resfile---------------
NATRO
start
1 A ALLAA
2 A ALLAA
3 A ALLAA
...            # Continue for residues 4-393
394 A ALLAA
-------------------------------------------
```

RECON multi-state design

As mentioned in the manuscript, all aligned positions within each conformation were considered for design in protein ensembles as long as the the protein ensemble contained at least two conformations/PDB files that had a root mean square distance of 5 Å and at least 120 aligned positions of the same sequence. For RECON multi-state design of the DV E ensemble, the following scripts were used:

```
mkdir designs/  # Create directory for output models
```

```
mpiexec -n 6 \\
/path/to/rosetta/main/source/bin/rosetta_scripts.mpi. \\
linuxgccrelease \\
@msd.options -l models.list -scorefile denvE-msd.fasc

--------------------models.list---------------------
1oan_diff.pdb
1ok8_diff.pdb
3c5x_diff.pdb
3c6e_diff.pdb
3j27.pdb
3j2p_diff.pdb
----------------------------------------------------



--------------------msd.options---------------------
-database /path/to/rosetta/main/database
-in:file:fullatom
-in:detect_disulf false
-mute protocols.simple_moves.GenericMonteCarloMover
-parser:protocol msd.xml
-run:msd_job_dist
-use_input_sc
-linmem_ig 50
-out:file:fullatom
-out:pdb_gz
-out:suffix _msd_
-out:path:pdb designs/
-nstruct 1
----------------------------------------------------



----------------------msd.xml-----------------------
<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=denvE.resfile debug=1 />
        <MSDMover name=msd2 design_mover=design \\
        constraint_weight=1.0 \\
        resfiles=denvE.resfile debug=1 />
        <MSDMover name=msd3 design_mover=design \\
        constraint_weight=1.5 \\
        resfiles=denvE.resfile debug=1 />
        <MSDMover name=msd4 design_mover=design \\
        constraint_weight=2.0 \\
        resfiles=denvE.resfile debug=1 />
        <FindConsensusSequence name=finish \\
        scorefxn=tal resfiles=denvE.resfile debug=1 \\
        task_operations=ifcl repack_one_res=1 />
        <FastRelax name=relax scorefxn=talaris2013 \\
        task_operations=ifcl,rtr repeats=1 />
    </MOVERS>
    <FILTERS>
        <FitnessFilter name=fitness \\
        output_to_scorefile=1 />
    </FILTERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
        <Add mover=msd1 />
        <Add mover=msd2 />
        <Add mover=msd3 />
        <Add mover=msd4 />
        <Add mover=finish />
        <Add filter=fitness />
        <Add mover=relax />
      </PROTOCOLS>
</ROSETTASCRIPTS>
-----------------------------------------------------
```

This protocol will run in parallel over 6 processors, one for each state. For the benchmark, this protocol was run 100 times to generate 100 designed ensembles of DV E. The protocol will run four rounds of multi-state design followed by repacking only, not a subsequent minimization as described previously, to avoid over-optimization to the Rosetta energy score function during design and to conserve the peptide backbone geometry of the original relaxed state. A step was added to calculate the fitness, or energy of an ensemble, defined as the sum of total energy over all input states divided by the number of all input states. The ten ensembles with the lowest fitness were used for benchmark analysis.

Single-state design

The protocol for single-state design for each state or PDB file within an ensemble uses the same Rosetta talaris2013 scoring funtion, but each state is designed independently following this protocol:

```
/path/to/rosetta/main/source/bin/rosetta_scripts. \\
linuxgccrelease \\
-s 3c6e_relaxed.pdb -scorefile 3c6e_ssd.fasc

------------------ssd.options------------------
-database /path/to/rosetta/main/database
-in:file:fullatom
-in:detect_disulf false
-mute protocols.simple_moves.GenericMonteCarloMover
-parser:protocol ssd.xml
-parser:script_vars resfile=denvE.resfile
-use_input_sc
-linmem_ig 50
-out:file:fullatom
-out:pdb_gz
-out:suffix _ssd_
-nstruct 5
-----------------------------------------------




--------------------ssd.xml--------------------
<ROSETTASCRIPTS>
    <SCOREFXNS>
    </SCOREFXNS>
    <TASKOPERATIONS>
        <InitializeFromCommandline name=ifcl />
        <RestrictToRepacking name=rtr />
        <ReadResfile name=rrf filename=%%resfile%% />
    </TASKOPERATIONS>
    <MOVERS>
        Design movers
        <PackRotamersMover name=design \\
        scorefxn=talaris2013 \\
        task_operations=ifcl,rrf />
        <FastRelax name=relax scorefxn=talaris2013 \\
        task_operations=ifcl,rtr repeats=1 />
    </MOVERS>
    <FILTERS>
    </FILTERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
        <Add mover=design />
        <Add mover=design />
        <Add mover=design />
        <Add mover=design />
        <Add mover=relax />
    </PROTOCOLS>
</ROSETTASCRIPTS>
----------------------------------------------
```

Generation of sequence profiles using sequences of natural homologues

The following describes the procurement of position-specific scoring matrices (PSSMs), or profiles, of mutation frequencies.

Design profiles

We used the WebLogo tool to generate a fasta file with the sequences of all designs, a sequence logo summarizing the bitscores of all twenty amino acids at each position, and a tab file summarizing the percentage of each amino acid type that populated each designed position. Given the length of the proteins used in the benchmark the sequence logo was not used, whereas the tab file was used for analysis (after first converting the percentages to frequencies).

```
# Generate fasta alignment of each state
# designed by RECON MSD
cat pdb.list | awk ’{system(‘‘design_analysis.py \\
  --native ’’\$1’’.pdb \\
  --format eps --resfile denvE.resfile --multiproc \\
  --units probability designs/’’\$1’’_msd*pdb’’)}’

# Generate fasta alignment of each state
# designed by SSD
cat pdb.list | awk ’{system(‘‘design_analysis.py \\
  --native ’’\$1’’.pdb \\
  --format eps --resfile denvE.resfile --multiproc \\
  --units probability designs/’’\$1’’_ssd*pdb’’)}’
```

```
-------pdb.list-------
1oan_relaxed
1ok8_relaxed
3c5x_relaxed
3c6e_relaxed
```

```
3j27_relaxed
3j2p_relaxed
———————————————————————
```

PSI-BLAST profiles

PSI-BLAST profiles were generated using a search query of non-redundant sequences using a databased downloaded from the NCBI BLAST server (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) on 2 May, 2014, and and run locally using `psiblast` version 2.2.29 as

```
/path/to/blast/2.2.29/bin/psiblast \\
−query \$NATIVE.fasta −db nr \\
−num_iterations 2 −out \$NATIVE.txt \\
−out_ascii_pssm \$NATIVE.ascii
```

The `$NATIVE.ascii` here represents the generated PSSM of the native sequence derived from the non-redunant database query, which was generated for each PDB native sequence used in the benchmark. The mutation profiles consist of columns 23-42 of the PSSM. Any sequences and the corresponding 20 amino acid type mutation frequencies that did not align to the sequences used for design were manually removed from the generated mutation profile.

Influenza Virus Resource database hemagglutinin stem profiles

Sequences were downloaded on October 14-15, 2019 from the Influenza Virus Resource Database using the search criteria listed in methods for influenza A HA and subtypes H1, H2, H3, H3N2, H4, and H7, as a single file of unaligned FASTA sequences. A multiple sequence alignment was performed using a locally installed Clustal Omega version 1.2.4 as:

```
clustalo −i H3N2.fa −o H3N2_aligned.fa
```

From the multiple sequence alignment, the frequencies of all twenty amino acids present at each aligned position were calculated using WebLogo 3, which was installed locally with Python 3.7,

```
weblogo −−sequence−type “protein \\−
−format logodata −−composition “none \\
<H3N2_aligned.fa> H3N2.logodata
```

The consensus sequence was determined from the multiple sequence alignment using `EMBOSS v.6.6.0.0` with the cons package (ftp://emboss.open-bio.org/pub/EMBOSS/), although was not reported in the manuscript.

```
cons −sequence H3N2_aligned.fa \\
−outseq H3N2.cons
```

The above three command line procedures were applied to each influenza A sequence group. It should be noted the sequences included within the entire influenza A FASTA sequences contained approximately 40 amino acids that were designated with amino acid type 'J', indicating an ambiguous designation between leucine or isoleucine, with only one sequence containing one amino acid type 'J' and each J did not occur at the same aligned position. Each 'J' designation was converted to leucine, since a 'J' designation was not allowed in the multiple sequence alignment using Clustal Omega.

The .logodata file generated using WebLogo contained the mutation frequencies of each aligned position for the entire HA protomer. HA1 and unaligned HA2 C-terminus residues were removed from each profile such that only positions that aligned with the native sequence of PDB ID 2HMG, chain F, residues 40-153 were included in each profile for analysis.

Calmodulin natural homologue profiles

The calmodulin mutation profile was determined by using the supplementary dataset Dataset_S01.xlsx provided by Halling, D.B. *et al.* (https://doi.org/10.1073/pnas.1600385113), where the accession number and sequence converted to FASTA format for alignment. The same command line procedures used to obtain HA (sub)type A profiles were used to obtain the calmodulin mutation profile.

Design analysis

Native sequence recovery

The reported native sequence recovery for designs was calculated as the ratio of the native amino acid bit score to all bit scores for each designed position within the ten lowest-scoring models of each designed native PDB model, or $aa_{nat\_frequency} = \frac{bit\_score_{native}}{bit\_score_{all}}$, with

```
python2.7 calc_nat_seq_recovery.py \\
--native 3c6e_relaxed.pdb --nmodels 10 \\
--res denvE.resfile 3c6e_msd_d.tab
```

The above script returns both the native sequence recovery of each designed position as well as the average native sequence recovery. In Fig 3 and in comparisons of sequence recovery to RMSD100, the average native sequence recovery is reported. Otherwise, the native sequence recovery of each designed position, or residue, is reported as a percentage.

The reported native sequence recovery for PSI-BLAST, IVR, and calmodulin profiles was calculated similarly, except that the frequency, not bit score, was used to calculate the native sequence recovery, with $aa_{nat\_frequency} = \frac{frequency_{native}}{frequency_{all}}$. It should be noted that for positions that contained aligned gaps, the $frequency_{all} \neq 1$. The following scripts require a header with the columns labeled for amino acid type, and the first column the native amino acid —labels should be one letter amino acid codes. The first script returns both position-specific native sequence recovery frequencies and average native sequence recovery, and is specific to the *.ascii file format. The second script returns only the average native sequence recovery.

```
python calc_pssm_nat_recovery.py \\
-f 3c6e.ascii -o 3c6e_pssm_res_recovery.csv
python calc_msa_nat_recovery.py \\
-f H3N2_align.profile -o H3N2_align_nsr.csv
```

Profile variance

Design mutation preferences, or profiles, were compared to PSI-BLAST profiles by calculating the sum of squared mutation frequency differences between PSI-BLAST and either RECON MSD or SSD profiles, and then normalized by the length of the aligned sequences. The reported average total variance in Fig

3B and Fig 5B represents the average sum of mutation preference differences squared an individual residue exhibits between two profiles, with a score of 0 indicating that the mutation profiles are identical. Unlike when calculating native sequence recovery, comparison of profiles were not calculated for each individual conformation. Instead, the mutation tolerances of each aligned position were averaged using all conformations within the ensemble first before calculating the total variance. Average total variance was calculated using the following

```
python profile_variances.py \\
--reference_profile denvE.ascii \\
--comparison_profile denvE_msd_d.tab \\
--variation_output denvE_RECON_var.csv
```

Testing for equality of variances between two profiles was achieved by using the scipy.stats.levene function, with the center set to median.

```
python Levene_test_for_equal_variances.py \\
profile1.tab profile2.tab
```

Amino acid exchangeability

In the manuscript we used the term *amino acid exchangeability* to represent the averge frequency the native, or *i*, amino acid is replaced with a non-native, or *j*, mutation. Average mutation frequencies, including native amino acid conservation frequencies, were calculated by averaging each of the twenty $i \rightarrow j$ mutation frequencies for each *i* amino acid, using the script below. This script requires a space-delimited file containing a matrix of $n \times m$, with *n* being the length of all designed positions and *m* being the mutation profile of the native amino acid, with the first column containing the *i* amino acid (one-letter code), and a header of the *j* amino acid frequency columns. Average mutation frequencies were calculated for each conformation to generate a 20×20 matrix of average $i \times j$ frequencies. In the manuscript, we report the average of all $i \times j$ frequencies, which was calculated by taking the cumulative average of all $i \rightarrow j$ frequencies within PSI-BLAST, RECON MSD, and SSD profiles.

```
python per_restype_mutations.py 3c6e_msd_d.tab
```

From the average $i \rightarrow j$ mutation frequencies, all $i \rightarrow j$ frequencies where $j = i$ were excluded to calculate the mean amino acid exchangeability and mean native amino acid exchangeability reported in Fig 4B and Fig 4C. The reported mean amino acid exchangeability represents the average frequency any native amino acid is exchanged for a non-native amino acid; both the mean of amino acid exchangeability rates of all PSI-BLAST, RECON MSD, and SSD profiles and comparison of means were calculated using R.

```
‘‘‘{r}
Fig4Bdata <- read.csv(”exchangeability_density.csv”, \\
header = T, sep = ”,”)

PSIexchange <- subset(exchange.density, Profile == ”PSI-BLAST”)
RECONexchange <- subset(exchange.density, Profile == ”RECON”)
SSDexchange <- subset(exchange.density, Profile == ”SSD”)

wilcox.test(PSIexchange$Design, RECONexchange$Design,\\
```

~~~
alternative = ”two.sided”)
wilcox.test(PSIexchange$Design, SSDexchange$Design, \\
alternative = ”two.sided”)
t.test(PSIexchange$Design, RECONexchange$Design, paired=T, \\
alternative = ”two.sided”)
t.test(PSIexchange$Design, SSDexchange$Design, paired=T, \\
alternative = ”two.sided”)

describeBy(exchange.density, group = exchange.density\\
$Profile, mat = T)
```
~~~

The mean native amino acid exchangeability rates represents the average frequency a particular native amino acid is exchanged for a non-native amino acid. In Fig 4C, the mean native amino acid exchangeability is reported as the difference between PSI-BLAST mean native amino acid exchangeability rates and either RECON MSD or SSD rates. Individual $i \rightarrow j$ exchangeability frequencies obtained cumulatively from PSI-BLAST, RECON MSD, and SSD are reported in S3 Fig. Kendall τ_β and linear regression models were built using R.

~~~
```{r}
RECONexchange.corr <- read.csv(”RECONexchangeability_ \\
correlation.csv”, header = T, sep = ”,”)
RECONexchange.lm <- lm(Design ~ PSIBLAST, \\
data = RECONexchange.corr)
Kendall(RECONexchange.corr$Design, RECONexchange.corr$PSIBLAST)
summary(RECONexchange.lm)

SSDexchange.corr <- read.csv(”SSDexchangeability_ \\
correlation.csv”, header = T, sep = ”,”) \\
SSDexchange.lm <- lm(Design ~ PSIBLAST, data = SSDexchange.corr)
Kendall(SSDexchange.corr$Design, SSDexchange.corr$PSIBLAST)$
summary(SSDexchange.lm)
```
~~~

Calculation of $RMSD_{da}$ and contact proximity deviation

As in the case of RSV F protein, even though designs contained the same number of residues, not all conformations contained equivalent chain breaks. Therefore, for residues that form either the N- or C-termini of a chain in any conformation were given an $RMSD_{da}$ score of 0. Otherwise, the dihedral angle deviation of a single residue was calculated as described in the **Methods** section.

~~~
python find_dihedral_deviation.py \\
--list_of_pdb_files 1oan_relaxed.pdb\\
1ok8_relaxed.pdb 3c5x_relaxed.pdb \\
3c6e_relaxed.pdb 3j27_relaxed.pdb\\
3j2p_relaxed.pdb --output_file denvE_rmsdda.csv
~~~

Contact proximity deviation was calculated for all aligned positions within an ensemble using the following:

```
python find_contact_deviations.py \\
−−pdb_list 1oan_relaxed.pdb \\
1ok8_relaxed.pdb 3c5x_relaxed.pdb \\
3c6e_relaxed.pdb 3j27_relaxed.pdb \\
3j2p_relaxed.pdb −−deviation_matrix \\
denvE_contact_dev.csv \\
−−contact_deviation denvE_contact_tally.csv
```

The output within `denvE_rmsdda.csv` and `denvE_contact_tally.csv` were transposed and combined into a single file containing each aligned residue within all eight protein ensembles. A z-score was calculated for each residue's $RMDS_{da}$ and contact proximity score within a single ensemble to normalize scores for all eight ensembles. To calculate the dependency of native sequence recovery on either $RMSD_{da}$ or contact proximity deviation, the average conservation frequency of the native amino acid sequence within each ensemble was used as the reported native residue sequence recovery. Either $RMSD_{da}$ or contact proximity deviationA Kendall's τ_{β} coefficient was calculated using the combined z-scores of all eight ensembles

Plots

The following R scripts were used to generate the figures reported in the manuscript. Note that astericks indicating significance were added using Adobe Illustrator after the initial figure was generated.

Figure 3

```
‘‘‘{r}
Fig3Adata <− read.csv(”benchmark_NSR.csv”, \\
header = T, sep = ”,”)

Fig3A <− ggplot(subset(Fig3Adata, \\
  Minimization != ”Unminimized”), \\
  aes(x = Design, y = Percent_Nat_Seq))
+ geom_boxplot(aes(fill = Design), \\
  position=position_dodge(width=0.8))
+ scale_fill_manual(name = ””, \\
  values=c(”black”, ”#1e90ff”, ”#ff901e”))
+ labs(x = ””, y = ”\nNative Sequence Recovery (%)”)
+ scale_y_continuous(expand=c(0,0), \\
  limits = c(0, 130),breaks=seq(0, 100, by=25))
+ theme(legend.position = ”none”)

Fig3Bdata <− read.csv(”Profile_variability.csv”, \\
header = T, sep = ”,”)
```

~~~
  Fig3Bdata$Protein2 <− factor(Fig6Bdata\$Protein, \\
  labels = c(”5′−nucleotidase”, ”Adenylate kinase”, ”CagL”, \\
  ”Calmodulin”,”Dengue E protein”, ”Influenza HA2”, ”GroEL”, \\
  ”RSV F protein”))

  Fig3B <− ggplot(subset(Fig2Bdata, Minimization == ”Relaxed”),
    aes(x = Design, y = FreqVariability, color=Design, \\
    shape=Protein2))
  + geom_point()
  + geom_path(aes(group=Protein2), color=”#909090”)
  + scale_color_manual(values = c(”#1e90ff”,”#ff901e”), \\
    name = ””, guide=F)
  + scale_shape_manual(name = ”Benchmark Case”, \\
    values = c(17, 0, 4, 8, 9, 11, 13, 15, 2, 3))
  + labs(y = ”\nNormalized Variability\n \\
    from PSI−BLAST profile”, x=””)
  + theme(legend.position = ”right”, \\
    legend.justification = ”center”,
    legend.direction = ”horizontal”, \\
    legend.box = ”vertical”)
  + guides(shape=guide_legend(ncol = 1, \\
    byrow=TRUE, title.position = ”top”))
  + ylim(0, 1.25)

  Fig3 <− plot_grid(Fig3A, Fig3B, nrow = 1, \\
  rel_widths = c(0.75, 1),
  labels = c(”A”, ”B”))
  Fig3
  ```
~~~

Figure 4

Fig 4A PSI-BLAST average mutation frequencies

~~~
```{r}
  PSIBLAST <− read.csv(”n10_AA_freq.csv”, \\
  header = T, sep = ”,”)

  PSIBLAST$Native <− factor(PSIBLAST$Native, \\
  c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
  ”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

  PSIBLAST$Mutate <− factor(PSIBLAST$Mutate, \\
  c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
  ”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

  PSIBLAST <− ggplot(PSIBLAST, aes(Native,Mutate))
~~~

```
+ geom_tile(aes(fill=frequency), color=”black”)
+ scale_fill_gradientn(name=”Mutation\nFrequency”, \\
  colours=c(”white”,”#ff901e”, ”#1e90ff”,”black”), \\
  limits = c(0,1), guide = guide_legend(reverse = T), \\
  breaks = c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0))
+ labs(title = ”PSI−BLAST”)
+ xlab(bquote(’Native ’ ~ AA[x]))
+ ylab(bquote(’Average mutation frequency of ’ ~ AA[x]))
+ theme(legend.position = ”left”, \\
  legend.justification = ”center”)

# Isolate legend for final figure
Fig4Alegend <− get_legend(PSIBLAST)

# Remove legend from panel
PSIBLASTplot <− ggplot(PSIBLAST, aes(Native,Mutate))
+ geom_tile(aes(fill=frequency), color=”black”)
+ scale_fill_gradientn(name=”Frequency”, \\
  colours=c(”white”,”#ff901e”, ”#1e90ff”,”black”), \\
  limits = c(0,1), \\
  guide = guide_legend(reverse = T), \\
  breaks = c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0))
+ labs(title = ”PSI−BLAST”)
+ xlab(bquote(’Native ’ ~ AA[x]))
+ ylab(bquote(’Average mutation frequency of ’ ~ AA[x]))
+ theme(legend.position = ”none”)
‘‘‘
```

Fig 4A RECON average mutation frequencies

```
‘‘‘{r}
RECON <− read.csv(”rMSD_AA_freq.csv”, \\
header = T, sep = ”,”)

RECON$Native <− factor(RECON$NativeAA, \\
c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

RECON$Mutate <− factor(RECON$MutateAA, \\
c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

RECONplot <− ggplot(RECON, aes(Native,Mutate))
+ geom_tile(aes(fill=frequency), color=”black”)
+ scale_fill_gradientn(name=””, \\
  colours=c(”white”, ”#ff901e”, ”#1e90ff”,”black”), \\
  limits = c(0,1), guide = guide_legend())
+ labs(title = ”RECON”)
+ xlab(bquote(’Native ’ ~ AA[x]))
+ ylab(bquote(’Average mutation frequency of ’ ~ AA[x]))
+ theme(legend.position = ”none”)
‘‘‘
```

Fig 4A SSD average mutation frequencies

```
‘‘‘{r}
SSD <- read.csv(”rSSD_AA_freq.csv”, \\
header = T, sep = ”,”)

SSD$Native <- factor(SSD$NativeAA, \\
c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

SSD$Mutate <- factor(SSD$MutateAA, \\
c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

SSDplot <- ggplot(SSD, aes(Native,Mutate))
+ geom_tile(aes(fill=frequency), color=”black”)
+ scale_fill_gradientn(name=””, \\
  colours=c(”white”,”#ff901e”, ”#1e90ff”,”black”), \\
  limits = c(0,1), guide = guide_legend())
+ labs(title = ”SSD”)
+ xlab(bquote(’Native ’ ~ AA[x]))
+ ylab(bquote(’Average mutation frequency of ’ ~ AA[x]))
+ theme(legend.position = ”none”)
‘‘‘
```

Fig 4B

```
‘‘‘{r}
Fig4Bdata <- read.csv(”exchangeability_density.csv”, \\
header = T, sep = ”,”)

Fig4Bdata$Profile <- factor(exchange.density$Profile,
labels = c(”PSI-BLAST”, ”RECON”, ”SSD”))

Fig4B <- ggplot(Fig4Bdata, aes(x=Profile,y=Design, \\
  fill=Profile))
+ geom_boxplot()
+ scale_fill_manual(name=””, \\
  values = c( ”black”,”#1e90ff”,”#ff901e”), guide=F)
+ labs(y=”\nExchangeability”, x=””)
+ theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

~~~
+ ylim(0, 0.25)
‘‘‘
~~~

Fig 4C

~~~
‘‘‘{r}
Fig4Cdata <- read.csv(”AA_freq_avgdeviationCompare_noDesign.csv”,
header = T, sep = ”,”)

Fig4Cdata$Native <- factor(nonNativeDiff$NativeAA, \\
c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

Fig4Cdata$profile <- factor(nonNativeDiff$Profile, \\
c(”PR”, ”PS”))

Fig4Cdata$profile <- factor(nonNativeDiff$profile, \\
labels = c(”PSI-BLAST - RECON      ”, ”PSI-BLAST - SSD      ”))

Fig4C <- ggplot(Fig4Cdata, aes(x = Native, y = NonNative, \\
  color=profile, shape=profile)) + geom_point(size=3)
+ scale_color_manual(values = c(”#454545”, ”#909090”), name = ””)
+ scale_shape_manual(name=””, values = c(15, 17))
+ xlab(bquote(’Native ’ ~ AA[average]))
+ ylab(TeX(”$\\Delta \\bar{AA}_{exchangeability}$”))
+ ylim(-0.035, 0.035)
+ geom_hline(yintercept = 0, linetype=”dashed”)
+ theme(legend.position = ”bottom”, \\
  legend.justification = ”center”, \\
  legend.box = ”vertical”, \\
  legend.text = element_text(size = 12), \\
  legend.title = element_blank())
+ guides(shape=guide_legend(nrow=1,byrow=TRUE, \\
  title.position = ”top”))
‘‘‘
~~~

Fig 4

~~~
‘‘‘{r}
Fig4A <- plot_grid(Fig4Alegend, PSIBLAST, RECON, SSD,
nrow = 1, rel_widths = c(0.4,1,1,1), \\
labels = c(”A”,””,””,””))

Fig4BC <- plot_grid(Fig4B, Fig4C, nrow = 1, \\
rel_widths = c(0.65, 1),
labels = c(”B”,”C”))
+ theme(plot.margin=unit(c(5.5,5.5,11,5.5), ”pt”))
~~~

~~~
Fig4 <- plot_grid(Fig4A, Fig4BC, ncol = 1)

Fig4
```
~~~

Figure 5

Fig 5A and 5C

~~~
```{r}
Fig5ACdata <- read.csv(”functional_profile_variances.csv”, \\
header = T, sep = ”,”)

Fig5ACdata$Protein <- factor(funcprofilevar$Protein, \\
labels = c(”Calmodulin”, ”Influenza A HA2”))

calmodulin <- subset(Fig5ACdata, Protein==”Calmodulin”)

HA2 <- subset(Fig5AC, Protein==”Influenza A HA2”)

calmodulinvar <- ggplot(calmodulin, \\
  aes(group=Protein, x=Comparison, \\
  y=Deviation, fill=Comparison))
+ geom_bar(stat = ”identity”, color=”black”)
+ scale_fill_manual(values=c(”#1e90ff”,”#ff901e”),name = ””)
+ labs(y = ”\nRMSD from Design Profile”, x=””)
+ theme(legend.position = ”none”)
+ ylim(0, 1)
+ facet_grid(~ Protein)

HA2var <- ggplot(HA2, aes(group=Protein, x=Comparison, \\
  y=Deviation, fill=Comparison))
+ geom_bar(stat = ”identity”, color=”black”)
+ scale_fill_manual(values=c(”#1e90ff”,”#ff901e”),name = ””)
+ labs(y = ”\nRMSD from Design Profile”, x=””)
+ theme(legend.position = ”none”)
+ ylim(0, 1)
+ facet_grid(~ Protein)
```
~~~

Fig 5B and 5D

~~~
```{r}
calmodulin_resdev <- ggdraw()
+ draw_image(”calmodulin_resdev.png”)

HA2_resdev <- ggdraw()
~~~

~~~~
+ draw_image(”influenza_pre_resdev.png”)
‘‘‘
~~~~

Fig 5

~~~~
‘‘‘{r}
plot_grid(calmodulinvar, calmodulin_resdev, \\
HA2var, HA2_resdev, \\
nrow = 2, rel_widths = c(0.4, 1), labels = ”AUTO”)
‘‘‘
~~~~

Figure 6

Fig 6A

~~~~
‘‘‘{r}
Fig6Adata <− read.csv(”HA_var_dist_matrix.csv”, \\
header = T, sep = ”,”)

Fig6Adata.matrix <− as.matrix(Fig6Adata[,−c(1)])

rownames(Fig6Adata.matrix) <− Fig6Adata$Subtype

Fig6A.dendro <− as.dendrogram(hclust(d = \\
dist(x = Fig6Adata.matrix)))

HAdendsort <− dendsort(Fig6A.dendro)

plot(HAdendsort, type = ”triangle”)
‘‘‘
~~~~

Fig 6B

~~~~
‘‘‘{r}
Fig6Bdata <− read.csv(”HA2_subtype_variances.csv”, \\
header = T, sep = ”,”)

Fig6Bdata$Subtype <− factor(HA2subtypesdata$Subtype, \\
c(”H3”,”H3N2”,”H4”,”H7”,”H1”,”H2”))

Fig6B <− ggplot(Fig6Bdata, \\
  aes(x=Subtype, y=RMSD,  fill=Comparison))
+ geom_bar(stat = ”identity”, color=”black”, \\
  position=position_dodge())
+ scale_fill_manual(values = c(”#1e90ff”,”#ff901e”), \\
  name = ”IVR MSA profile RMSD with respect to”)
+ labs(y = ”\nRMSD”, x=”IVR MSA Profile”) + ylim(0, 1)
+ theme(legend.position = ”bottom”)
~~~~

```
‘‘‘
```

Figure 7

```
Fig 7A
‘‘‘{r}
Fig7Adata <- read.csv(”benchmark_relaxed_NSR.csv”, \\
header = T, sep = ”,”)

Fig7Adata$maxRMSD100 <- cut2(Fig7Adata$MaxRMSD100, g=3)

Fig7A <- ggplot(relaxnsr, \\
    aes(x = maxRMSD100, y = Percent_Nat_Seq))
+ geom_boxplot(aes(color=Design, fill=Design), alpha=0.3)
+ labs(x = ”\nMaximum RMSD100 (Å)”, y = ”\nNSR (\%)”)
+ scale_color_manual(name = ””, \\
    values = c(”black”,”dodgerblue”,”#ff901e”))
+ scale_fill_manual(name = ””, \\
    values = c(”black”,”dodgerblue”,”#ff901e”))
+ facet_wrap(~ Design)
+ theme(legend.position = ”none”, \\
    axis.text.x = element_text(size = 8, \\
    angle = 45, hjust = 1))
+ ylim(0,130)
‘‘‘
```

Fig 7B

```
‘‘‘{r}
Fig7BCdata <- read.csv(”proximity_diRMSD_bydesign.csv”, \\
header = T, sep = ”,”)

Fig7BC$Measure <- factor(Fig7BCdata$Measure, \\
labels = c(”Contact Proximity\n \\
Deviation (Å)”, ”Dihedral Angle\nDeviation (rad)”))

Fig7Bdata <- subset(Fig7BCdata,Measure == \\
”Contact Proximity\nDeviation (Å)”)

Fig7Bdata$deviation <- cut2(Fig7Bdata$Deviation, g=3)

Fig7B <- ggplot(Fig7Bdata, \\
aes(x=deviation, y=AvgPNSR, color=Design))
+ geom_boxplot(aes(color=Design, fill=Design), alpha=0.3)
+ scale_color_manual(name=””, \\
    values = c(”black”,”#1e90ff”,”#ff901e”))
```

```
+ scale_fill_manual(name=””, \\
  values = c(”black”,”#1e90ff”,”#ff901e”))
+ labs(x=”\nContact Proximity Deviation z−score (Å)”, \\
  y=”\nResidue NSR (\%)”)
+ theme(axis.text.x = element_text(angle = 45, \\
  size = 10, hjust = 1))
+ ylim(0,130)
+ facet_wrap(~ Design)
+ theme(legend.position = ”none”, \\
  axis.text.x = element_text(size = 8, \\
  angle = 45, hjust = 1))
‘‘‘
```

Fig 7C

```
‘‘‘{r}
Fig7Cdata <− subset(Fig7BCdata,Measure == \\
”Dihedral Angle\nDeviation (rad)”)

Fig7C$deviation <− cut2(Fig7Cdata$Deviation, g=3)

Fig7Cdata <− ggplot(Fig7Cdata, \\
  aes(x=deviation, y=AvgPNSR, color=Design))
+ geom_boxplot(aes(color=Design, fill=Design), alpha=0.3)
+ scale_color_manual(name=””, \\
  values = c(”black”,”#1e90ff”,”#ff901e”))
+ scale_fill_manual(name=””, \\
  values = c(”black”,”#1e90ff”,”#ff901e”))
+ labs(x=”\nDihedral Angle RMSD z−score (rad)”, \\
  y=”\nResidue NSR (\%)”)
+ theme(axis.text.x = \\
  element_text(angle = 45, size = 10, hjust = 1))
+ facet_wrap(~ Design)
+ theme(legend.position = \\
  ”none”, axis.text.x = element_text(size = 8, \\
  angle = 45, hjust = 1))
+ ylim(0,130)
‘‘‘
```

Figure 8

```
‘‘‘{r}
Fig8data <− read.csv(”benchmark_energies_relaxed\\
−w−templates.csv”,
header = T, sep = ”,”)

Fig8data$Design <− factor(Fig8data$Design, \\
labels = c(”RECON”,”SSD”,”Native”))

Fig8data$Dataset2 <- factor(Fig8data$Dataset, \\
labels = c(”5′-nucleotidase”, ”Adenylate kinase”, ”CagL”, \\
”Calmodulin”,”Dengue E protein”,”Influenza HA stem”, \\
”GroEL subunit”, ”RSV F protein”))

Fig8 <- ggplot(Fig7data, aes(x = Design, y = ResScore))
+ geom_violin(aes(fill = Design))
+ labs(x = ””, y = ”Mean Residue Score (REU)”)
+ scale_fill_manual(name=””, \\
  values = c(”#1e90ff”,”#ff901e”,”grey50”))
+ facet_wrap(~ Dataset2, ncol = 4)
+ theme(axis.text.x = \\
  element_blank(), axis.ticks.x = element_blank())
+ ylim(-2.5, -1.0)
‘‘‘
```

Figure 9

Fig 9A

```
‘‘‘{r}
Fig9Adata <- read.csv(”benchmark_energies_relaxed\\
-w-templates.csv”,\\
header = T, sep = ”,”)

Fig9Adata$maxRMSD100 <- cut2(Fig9Adata$MaxRMSD100, g=3)

Fig9Adata$Design <- factor(Fig9Adata$Design, \\
labels = c(”RECON”,”SSD”,”Native”))

Fig9Adata$Design <- factor(Fig9Adata$Design, \\
c(”Native”,”RECON”,”SSD”))

Fig9Adata.subset <- subset(Fig9Adata, Design != ”Native”)

Fig9A <- ggplot(Fig9Adata.subset, \\
aes(x=maxRMSD100, y=ResScore))
+ geom_boxplot(aes(color=Design, fill=Design), alpha=0.3)
+ labs(x = ”\nMaximum RMSD100 (Å)”, \\
  y = ”\nMean Residue Total Score (REU)”)
+ scale_color_manual(name = ””, \\
  values = c(”#1e90ff”,”#ff901e”))
+ scale_fill_manual(name = ””, \\
  values = c(”#1e90ff”,”#ff901e”))
+ facet_wrap(~ Design)
+ theme(legend.position = ”none”, \\
  axis.text.x = element_text(size = 6, \\
  angle = 45, hjust = 1), \\
  axis.title = element_text(size = 10), \\
  title = element_text(size = 12))
+ ylim(-5,2.5)
‘‘‘
```

Fig 9B

```
‘‘‘{r}
Fig9Bdata <- read.csv(”proximity_diRMSD_bydesignonly.csv”, \\
header = T, sep = ”,”)

Fig9Bdata$Measure <- factor(Fig9Bdata$Measure, \\
labels = c(”Contact Proximity\nDeviation (Å)”, \\
”Dihedral Angle\nDeviation (rad)”))

Fig9Bdata <- subset(Fig8Bdata, \\
Measure == ”Dihedral Angle\nDeviation (rad)”)

Fig9Bdata$deviation <- cut2(Fig9Bdata$Deviation, g=3)

Fig9B <- ggplot(Fig9Bdata, \\
  aes(x=deviation, y=ResEnergy, color=Design))
+ geom_boxplot(aes(color=Design, fill=Design), \\
  alpha=0.3)
+ scale_color_manual(name=””, \\
  values = c(”#1e90ff”, ”#ff901e”))
+ scale_fill_manual(name=””, \\
  values = c(”#1e90ff”, ”#ff901e”))
+ labs(x=”\nDihedral Angle RMSD z-score (rad)”, \\
  y=”Residue Score (REU)”)
+ facet_wrap(~ Design)
+ theme(legend.position = ”none”, \\
  axis.text.x = element_text(size = 6, \\
  angle = 45, hjust = 1), \\
  axis.title = element_text(size = 10), \\
  title = element_text(size = 12))
+ ylim(-5, 2.5)
‘‘‘
```

Fig 9C

```
‘‘‘{r}
Fig9Cdata <- read.csv(”proximity_diRMSD_bydesignonly.csv”, \\
header = T, sep = ”,”)
```

~~~
Fig9Cdata$Measure <- factor(Fig9Cdata$Measure, \\
labels = c(”Contact Proximity\nDeviation (Å)”, \\
”Dihedral Angle\nDeviation (rad)”))

Fig9Cdata <- subset(Fig8Cdata,Measure == \\
”Contact Proximity\nDeviation (Å)”)

Fig9Cdata$deviation <- cut2(Fig9Cdata$Deviation, g=3)

Fig9C <- ggplot(Fig9Cdata,  \\
  aes(x=deviation, y=ResEnergy, color=Design))
+ geom_boxplot(aes(color=Design, fill=Design), alpha=0.3)
+ scale_color_manual(name=””, \\
  values = c(”#1e90ff”, ”#ff901e”))
+ scale_fill_manual(name=””, \\
  values = c(”#1e90ff”, ”#ff901e”))
+ labs(x=”\nContact Proximity Deviation (Å)”, \\
  y=”Residue Score (REU)”)
+ facet_wrap(~ Design)
+ theme(legend.position = ”none”, \\
  axis.text.x = element_text(size = 6, \\
  angle = 45, hjust = 1), \\
  axis.title = element_text(size = 10), \\
  title = element_text(size = 12))
+ ylim(-5, 2.5)
```
~~~

S1 Fig

S1 Fig A

~~~
```{r}
S1Adata <- read.csv(”benchmark_NSR.csv”, \\
header = T, sep = ”,”)

S1AFig <- ggplot(S1Adata, \\
  aes(x = Minimization, y = Percent_Nat_Seq))
+ geom_boxplot(aes(fill = Design), \\
  position = position_dodge(width=0.8))
+ facet_grid(cols = vars(Design), \\
  scales = ”free_x”, space = ”free_x”)
+ scale_fill_manual(name = ””, \\
  values = c(”black”,”#1e90ff”, ”#ff901e”))
+ labs(x = ””, y = ”\nNative Sequence Recovery (%)”)
+ scale_y_continuous(expand=c(0,0), limits = c(0,130), \\
  breaks=seq(0,100,by=25))
~~~

```
+ theme(legend.position = ”none”)
‘‘‘
```

S1 Fig B

```
‘‘‘{r}
S1Bdata <- read.csv(”Profile_variability.csv”, \\
header = T, sep = ”,”)

S1Bdata$Protein2 <- factor(S1Bdata$Protein, \\
labels = c(”5′-nucleotidase”,”Adenylate kinase”, \\
”CagL”,”Calmodulin”,”Dengue E protein”, \\
”Influenza HA2”,”GroEL”,”RSV F protein”))

S1BFig <- ggplot(S1Bdata, aes(x = Design, \\
  y = FreqVariability, color=Design,\\
  shape=Protein2))
+ geom_point()
+ geom_path(aes(group=Protein2), color=”#909090”)
+ scale_color_manual(values = c(”#1e90ff”, ”#ff901e”), \\
  name = ””, guide=F)
+ scale_shape_manual(name = ”Benchmark Case”, \\
  values = c(17,0,4,8,9,11,13,15,2,3))
+ facet_grid(~ Minimization)
+ labs(y = ”\nNormalized Variability\nfrom PSI-BLAST profile”)
+ theme(axis.title.x = element_blank(), \\
  legend.position = ”bottom”, \\
  legend.justification = ”center”, \\
  legend.direction = ”horizontal”, \\
  legend.box = ”vertical”)
+ guides(shape=guide_legend(nrow=4,byrow=TRUE, \\
  title.position = ”top”))
+ ylim(0, 1.25)
‘‘‘
```

S1 Fig

```
‘‘‘{r}
S1Fig <- plot_grid(S1AFig, S1BFig, nrow = 1, \\
rel_widths = c(1,0.7),\\
labels = c(”A”,”B”))
‘‘‘
```

S2 Fig

```
‘‘‘{r}
S2data <- read.csv(”AA_freq_deviationCompare.csv”, \\
header = T, sep = ”,”)

S2data$Native <- factor(S2data$NativeAA, \\
c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

S2data$Mutate <- factor(S2data$MutateAA, \\
c(”G”,”A”,”V”,”L”,”I”,”M”,”F”,”W”,”P”,”S”, \\
”T”,”C”,”Y”,”N”,”Q”,”D”,”E”,”K”,”R”,”H”))

S2data$profile <- factor(S2data$Profile, \\
c(”PR”,”PS”,”Design”))

S2data$profile <- factor(S2data$profile,\\
labels = c(”PSI-BLAST - RECON”,”PSI-BLAST - SSD”, \\
”RECON - SSD”))

S2Fig <- ggplot(S2data, aes(Native,Mutate))
+ geom_tile(aes(fill=FrequencyDiff), color=”black”)
+ scale_fill_gradientn(name=TeX(”$\\Delta$ Freq”), \\
  colours=c(”#994d00”,”white”,”#004d99”,”black”), \\
  limits = c(-0.3, 0.6), \\
  guide = guide_legend(reverse = T), \\
  breaks = c(-0.3,-0.2,-0.1,0.0,0.1,0.2,0.3,0.4,0.5,0.6))
+ facet_grid(~ profile)
+ xlab(bquote(’Native ’ ~ AA[x]))
+ ylab(TeX(”$\\Delta$ Average mutation frequency of $AA_x$”))
+ theme(legend.position = ”left”, \\
  legend.justification = ”center” )
‘‘‘
```

S3 Fig

S3 Fig A

```
‘‘‘{r}
RECONexchange.corr <- read.csv(”RECONexchangeability_\\
correlation.csv”, header = T, sep = ”,”)

RECONexchange.lm <- lm(Design ~ PSIBLAST, \\
data = RECONexchange.corr)

RECONexchange.model <- augment(RECONexchange.lm)

RECONexchange.model$predicted <- predict(RECONexchange.lm)

RECONexchange.model$residuals <- residuals(RECONexchange.lm)
```

~~~
RECONexchange.residuals <- ggplot(RECONexchange.model,\\
aes(x=PSIBLAST,y=Design))
+ geom_point(aes(color = abs(residuals), \\
  alpha = abs(residuals)))
+ xlim(0, 0.175)
+ ylim(0, 0.175)
+ geom_abline(intercept = 0, slope = 1, color=”#909090”)
+ labs(x=”PSI-BLAST”, y=”\nRECON”)
+ scale_color_continuous(low = ”#1e90ff”, \\
  high = ”black”, limit=c(0,0.1))
+ guides(alpha=F, color=F)

SSDexchange.corr <- read.csv(”SSDexchangeability_\\
correlation.csv”, header = T, sep = ”,”)

SSDexchange.lm <- lm(Design ~ PSIBLAST, \\
data = SSDexchange.corr)

SSDexchange.model <- augment(SSDexchange.lm)

SSDexchange.model$predicted <- predict(SSDexchange.lm)

SSDexchange.model$residuals <- residuals(SSDexchange.lm)

SSDexchange.residuals <- ggplot(SSDexchange.model,\\
aes(x=PSIBLAST,y=Design))
+ geom_point(aes(color = abs(residuals), \\
  alpha = abs(residuals)))
+ xlim(0, 0.175)
+ ylim(0, 0.175)
+ geom_abline(intercept = 0, slope = 1, \\
  color=”#909090”)
+ labs(x=”PSI-BLAST”, y=”\nSSD”)
+ scale_color_continuous(low = ”#ff901e”, \\
  high = ”black”, limit=c(0,0.1))
+ guides(alpha=F, color=F)

```
~~~

S3 Fig B

~~~
```{r}
RECONexchange.lm <- lm(Design ~ PSIBLAST, \\
data = RECONexchange.corr)
~~~

```
indexlabel <- read.csv(”IndexLabel.csv”, header = T)

N <- nrow(RECONexchange.lm$model)

df <- data.frame(Index = 1:N, \\
dfstats = dfbetas(RECONexchange.lm))

df$group <- factor(ifelse(df$dfstats.PSIBLAST > 0.1026, \\
1,ifelse(df$dfstats.PSIBLAST < -0.1026 , 1,0)))

df <- cbind(df, indexlabel)

RECON.dfbeta <- ggplot(df, \\
aes(Index, dfstats.PSIBLAST))
+ geom_point(size=0.25, aes(color=group))
+ geom_segment(aes(Index,xend=Index,0,\\
  yend=dfstats.PSIBLAST,color=group),data=df)
+ scale_color_manual(values=c(”#1e90ff”, ”black”))
+ labs(y=”\nDFBETA”)
+ guides(color = F)
+ ylim(-1,1)
+ geom_label_repel(data = subset(df, group==”1”), \\
  aes(label=Labelindex), \\
  nudge_y = 0.1, direction=”y”,segment.color = ”grey80”, \\
  segment.size = 0.5, size = 1.5)

SSDexchange.lm <- lm(Design ~ PSIBLAST, data = SSDexchange.corr)

N <- nrow(SSDexchange.lm$model)

df <- data.frame(Index = 1:N, \\
dfstats = dfbetas(SSDexchange.lm))

df$group <- factor(ifelse(df$dfstats.PSIBLAST > 0.1026, \\
1,ifelse(df$dfstats.PSIBLAST < -0.1026 , 1,0)))

df <- cbind(df, indexlabel)

SSD.dfbeta <- ggplot(df,aes(Index, dfstats.PSIBLAST))
+ geom_point(size=0.25, aes(color=group))
+ geom_segment(aes(Index,xend=Index, 0,\\
  yend=dfstats.PSIBLAST,color=group),data=df)
+ scale_color_manual(values=c(”#ff901e”, ”black”))
+ labs(y=”\nDFBETA”)
+ guides(color = F)
+ ylim(−1,1)
+ geom_label_repel(data = subset(df, group==”1”), \\
  aes(label=Labelindex), \\
  nudge_y = 0.1, direction=”y”, segment.color = ”grey80”, \\
  segment.size = 0.5, size = 1.5)
‘‘‘
```

S3 Fig

```
‘‘‘{r}
S3Fig <− plot_grid(RECONexchange.residuals, \\
SSDexchange.residuals, \\
RECON.dfbeta, SSD.dfbeta, \\
nrow = 2, ncol = 2, labels = c(”A”,””,”B”,””))
‘‘‘
```

S4 Fig

```
‘‘‘{r}
S4data <− read.csv(”HA_var.csv”, \\
header = T, sep = ”,”)

S4data$Subtype <− factor(S4data$Subtype, \\
c(”H3N2”,”H3”,”H4”,”H7”,”H1”,”H2”))

ggplot(S4data, aes(x=ResNum, y=ResDev, color=Subtype))
+ geom_point()
+ scale_color_manual(name=”Subtype”, \\
  values = c(”#901eff”,”#982fea”,”#a548cc”,\\
  ”#ba7298”,”#ca9370”, ”#e3c532”))
+ labs(x=”Residiue Number (2HMG, chain F)”, \\
  y=”RMSD of design profile to IVR subtype profile”)
+ facet_grid(Subtype ~ Design)
+ theme(legend.position = ”bottom”, \\
  legend.justification = ”center”, \\
  legend.direction = ”horizontal”, \\
  legend.box = ”vertical”)
+ theme(legend.position = ”none”)
‘‘‘
```

S5 Fig

S5 Fig A

```
‘‘‘{r}
S5data <− read.csv(”proximityXcontact.csv”, \\
header = T, sep=”,”)
```

```
S5AFig <− ggplot(S5data, aes(x=diRMSD, y=Tally))
+ geom_point(color=”white”, size=0.25)
+ geom_hex(bins=25)
+ scale_fill_gradientn(name=”Residue Count”, \\
  colours=c(”#E8E8E8”,”#a8a8a8”,”#909090”,”#454545”,\\
  ”#222222”,”black”), guide = guide_legend())
+ labs(x=”Dihedral Angle\nDeviation (rad)”, \\
  y=”\nContact Proximity\nDeviation (Å)”)
+ theme(legend.position = ”bottom”, \\
  legend.justification = ”center”, \\
  legend.direction = ”horizontal”, \\
  legend.box = ”vertical”, \\
  legend.text = element_text(size = 8), \\
  legend.key.height = grid::unit(0.33,”cm”), \\
  legend.key.width = grid::unit(0.33,”cm”))
  + guides(fill=guide_legend(nrow = 1, byrow = T, \\
  title.position = ”top”, \\
  title.hjust = 0.5), color=F)
‘‘‘
```

S5 Fig B

```
‘‘‘{r}
S5data <− read.csv(”proximityXcontact.csv”, \\
header = T, sep=”,”)

S5BFig <− ggplot(S5data, aes(x=ZdiRMSD, y=ZTally))
+ geom_point(color=”white”, size=0.25) + geom_hex(bins=25)
+ scale_fill_gradientn(name=”Residue Count”, \\
  colours=c(”#E8E8E8”,”#a8a8a8”,”#909090”,\\
  ”#454545”,”#222222”,”black”), \\
  guide = guide_legend(), \\
  breaks = c(10,25,50,75,100,150))
+ labs(x=”Dihedral Angle\nDeviation z−score (rad)”, \\
  y=”\nContact Proximity\nDeviation z−score (Å)”)
+ theme(legend.position = ”bottom”, \\
  legend.justification = ”center”, \\
  legend.direction = ”horizontal”,
  legend.box = ”vertical”, \\
  legend.text = element_text(size = 8),\\
  legend.key.height = grid::unit(0.33,”cm”), \\
  legend.key.width = grid::unit(0.33,”cm”))
+ guides(fill=guide_legend(nrow = 1, byrow = T, \\
  title.position = ”top”, \\
  title.hjust = 0.5), color=F)
```

‘ ‘ ‘

APPENDIX B

SUPPLEMENTARY FIGURES FOR THE RECON MSD BENCHMARK

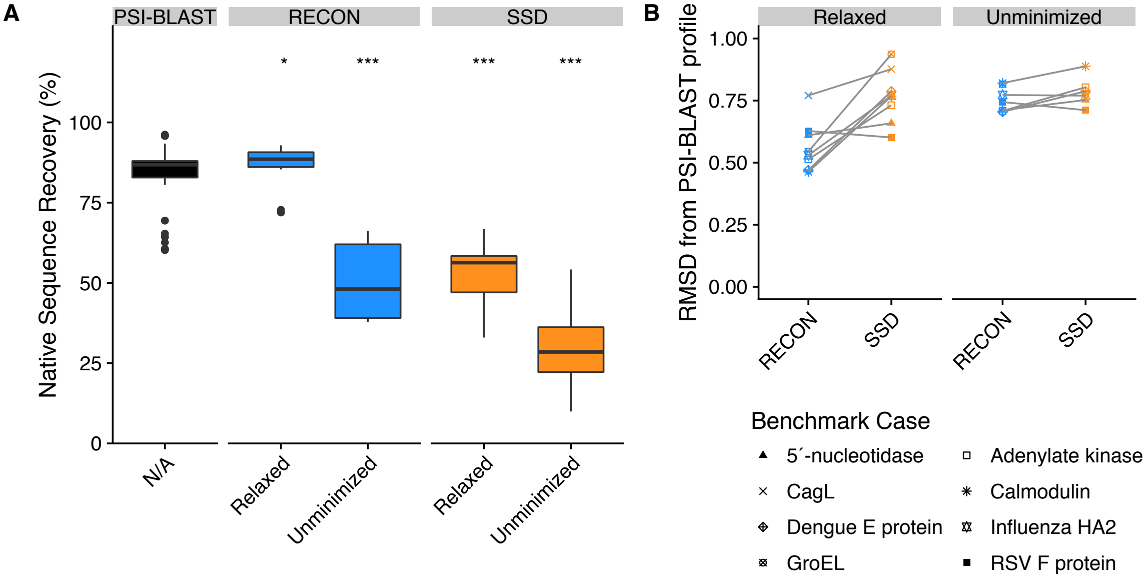


Figure B.1.: Design native sequence recovery and mutation profile variability comparisons to PSI-BLAST profiles using relaxed and unminimized starting models. *(A) Comparison of total native sequence recovery of relaxed and unminimized RECON MSD and SSD designs to PSI-BLAST sequence profiles generated using the native sequence. Asterisks indicate the significance of difference of means of each design in comparison to the PSI-BLAST profile, with a z-test p-value < 0.01 represented by one asterisk, and a p-value < 0.00001 by three asterisks. (B) Mutation frequency variances of designs in comparison to a PSI-BLAST profile, normalized by protein length. The y-axis values represent the average variability of mutation profiles for each designed residue in relation to a PSI-BLAST profile, as described in Fig 3.*

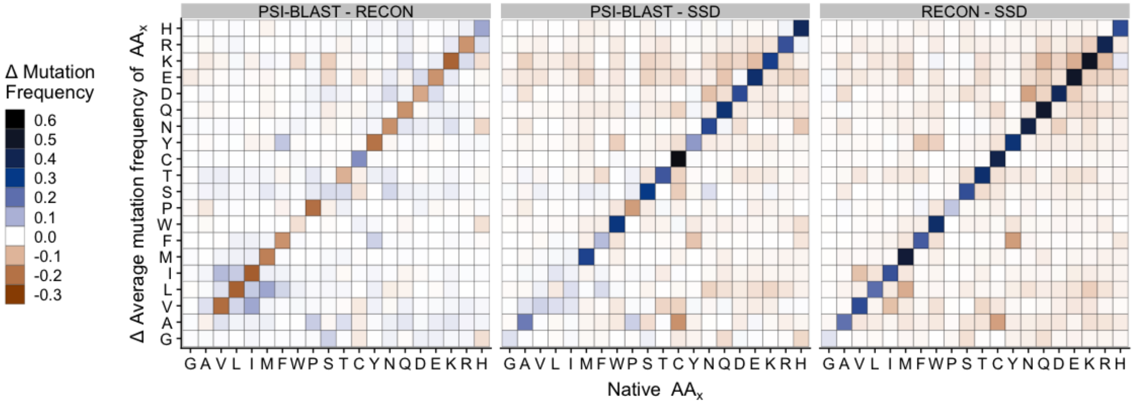


Figure B.2.: Difference in average amino acid exchangeability between sequence profiles. *The x axis represents the original amino acid. The y axis represents the difference in average mutation frequencies between two profiles, which are noted above each grid. Along the diagonal axis, indicating native sequence conservation, values less than zero (oranges) signify that the latter profile was more highly conserved, and values greater than zero (blues to black) signify that the native residue was less conserved in the latter profile. Not along the diagonal, values less than zero indicate that exchangeability of the native residue to the indicated residue along the y axis was higher in the latter profile, whereas values greater than zero indicate that exchangeability was lower in the latter profile.*

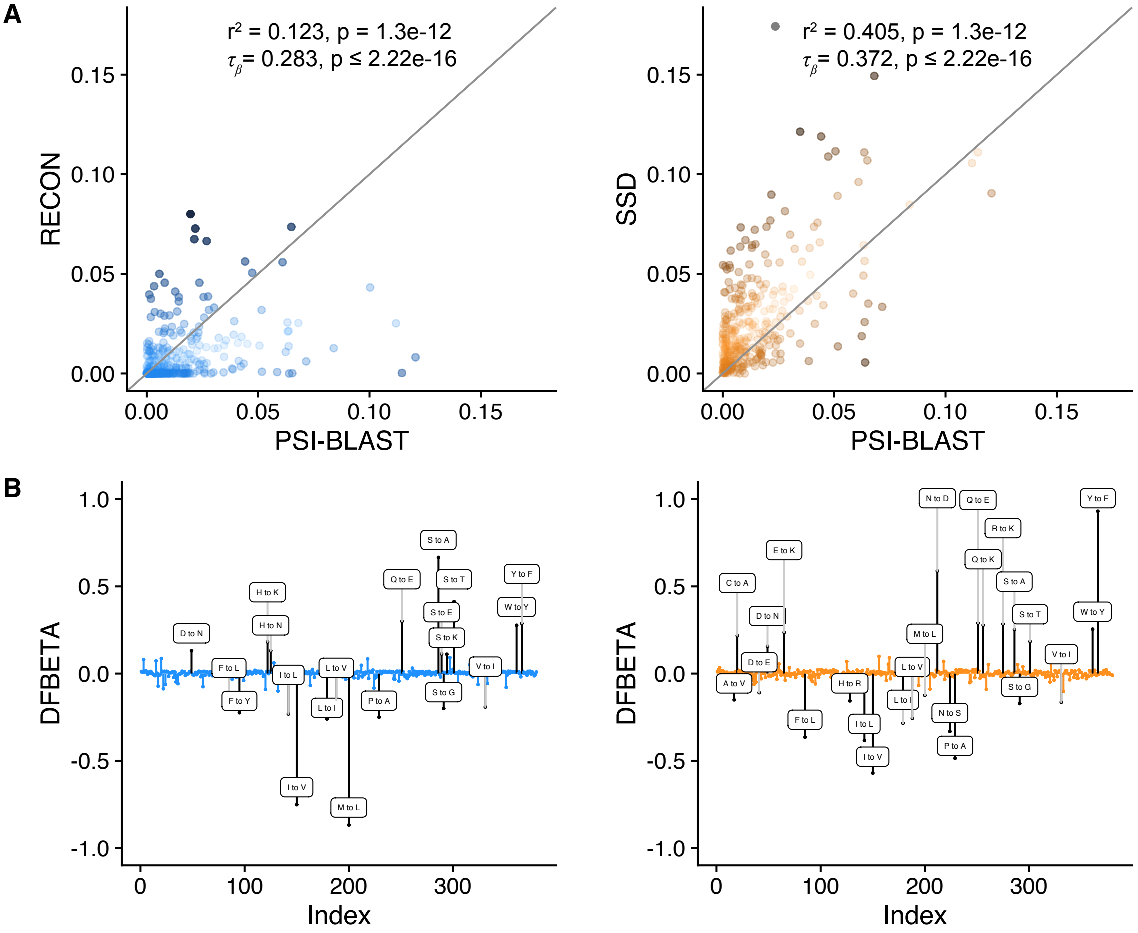


Figure B.3.: Comparison of individual exchangeability rates. *(A) Scatterplots of each exchangeability rate as observed in a PSI-BLAST profile compared to design profile. Both the x and y axes represent the exchangeability frequency of a native amino acid to a specific, non-native amino acid, with PSI-BLAST exchangeability rates along the x axis, and design exchangeability rates along the y axis. For reference, a grey line drawn is drawn along where the exchangeability rates would be equal between a PSI-BLAST and design profile. Both an adjusted r^2 and τ_β value is provided, along with the associated two-sided p-value. Lighter points are found along the linear regression model, and darker points represent outliers. (B) Measures of influence for individual exchangeability rate. The index listed along the x axis refers to each exchangeability rate, indexed in order alphabetically. For reference, in the first nineteen indices, the first index refers to the A to C mutation frequency, followed by the next eighteen indices that correspond with A to D through Y mutation frequencies. The measure of influential observation, or DFBETA index, is represented along the y axis. The height and direction of each bar corresponds with the change in regression model correlation coefficient without that particular observation. Influential outliers that have $> |\frac{2}{\sqrt{N}}|$ index value, or ±0.106 threshold, are colored in black and are labeled with the associated mutation.*

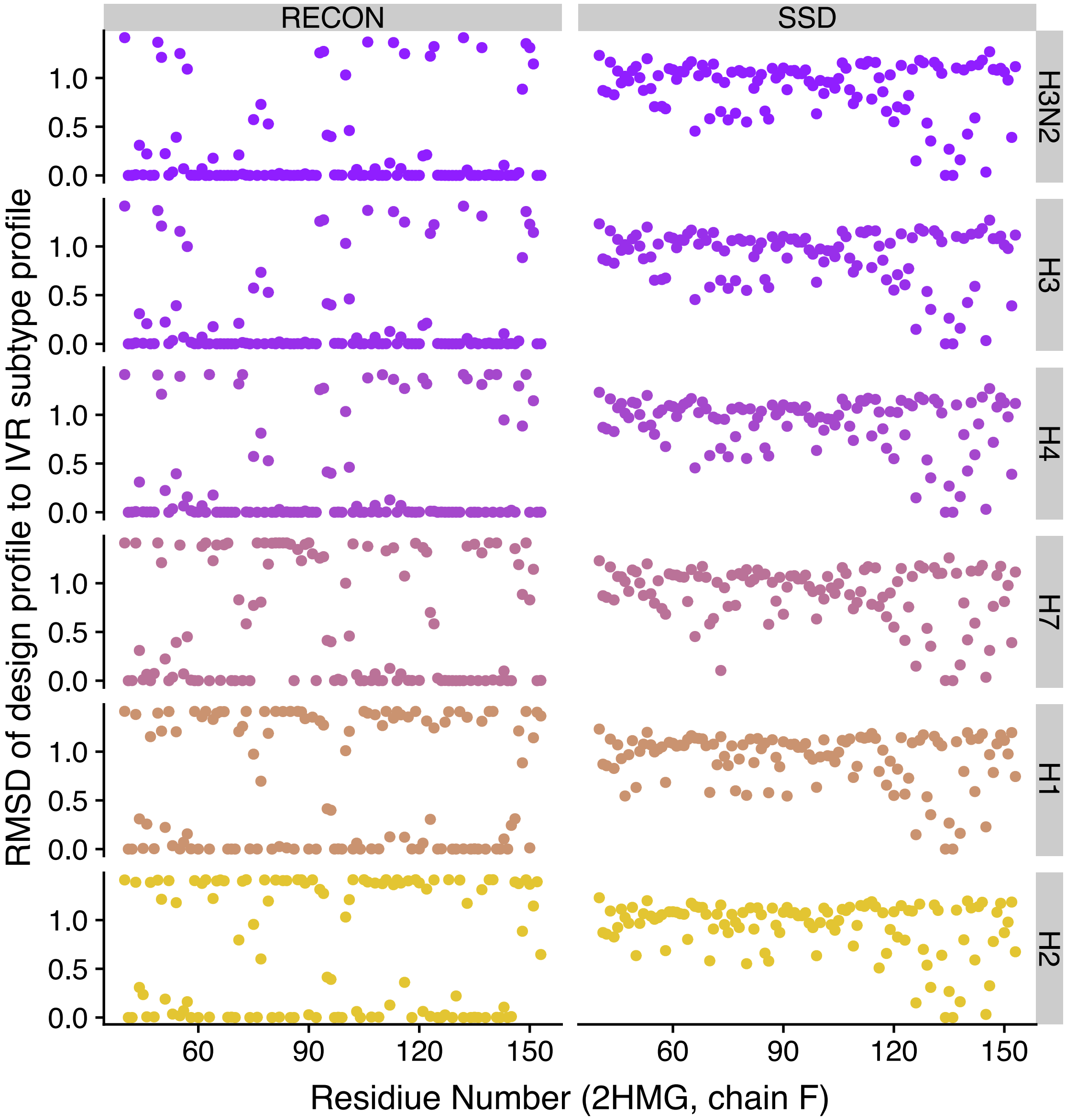


Figure B.4.: Root mean square deviation of residue mutation preferences between influenza A subtype multiple sequence alignments and their RECON MSD and SSD profiles. *Each IVR subtype mutation profile was generated by multiple sequence alignment of HA2 sequences within the IVR database, subdivided by HA subtype including H1, H2, H3, H4, and H7. Because the designed sequence used only an H3N2 HA2 backbone, the H3N2 subtype was included in addition to H3. Only positions that align to the native sequence used for design were included within the profile. HA2 subsequences are separated and ordered by similarity to H3N2, from highest similarity on the top. The x axis each aligned position of the HA2 sequence, corresponding to the H3N2 residue numbering of PDB ID 2HMG, chain F. The y axis is the root mean square deviation (RMSD) of each residue's subtype-specific profile within the multiple sequence alignment with respect to RECON MSD, on the left, and to SSD on the right.*

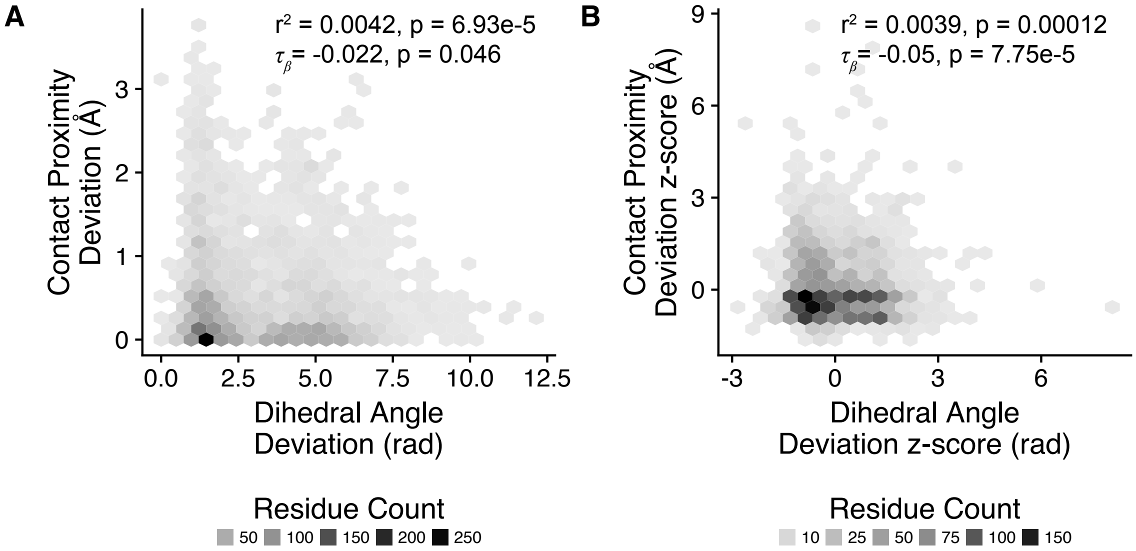


Figure B.5.: Correlation of dihedral angle RMSD and $C_\beta - C_\beta$ ***distance deviation.*** *(A) The x-axis represents dihedral RMSD, measured in radians, and the y-axis represents contact proximity deviation, measured in Å. The hex bins shaded in grey are the number of residues within the deposited PDB structure have have both a* $C_\beta - C_\beta$ *distance deviation and dihedral angle RMSD within a bin. (B) Axes represent same metrices as in Panel A, normalized by z-score.*

APPENDIX C

SUPPLEMENTARY FIGURES FOR THE IDENTIFICATION AND CLUSTERING OF MINIMAL CONFORMATIONAL B-CELL EPITOPES

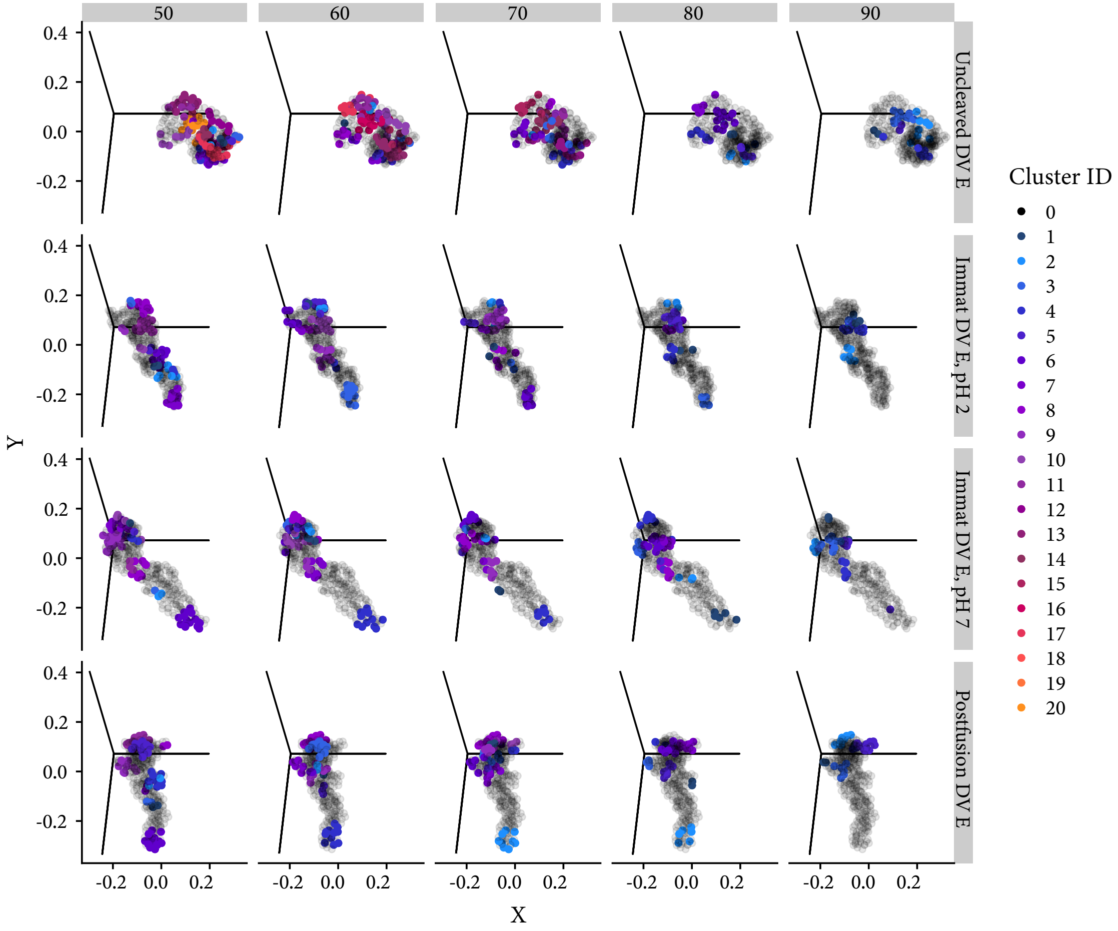


Figure C.1.: Clustered DV E residues identified by high contact proximity deviation, total energy scores, and community detection. *The x, y, and z axes represent the C_α atom coordinates of each residues with the pre- or post-fusion conformations, indicated on the right grey panels. C_α coordinates not considered for clustering are shown in transparent grey, whereas clustered residues are designated by an opaque color.*

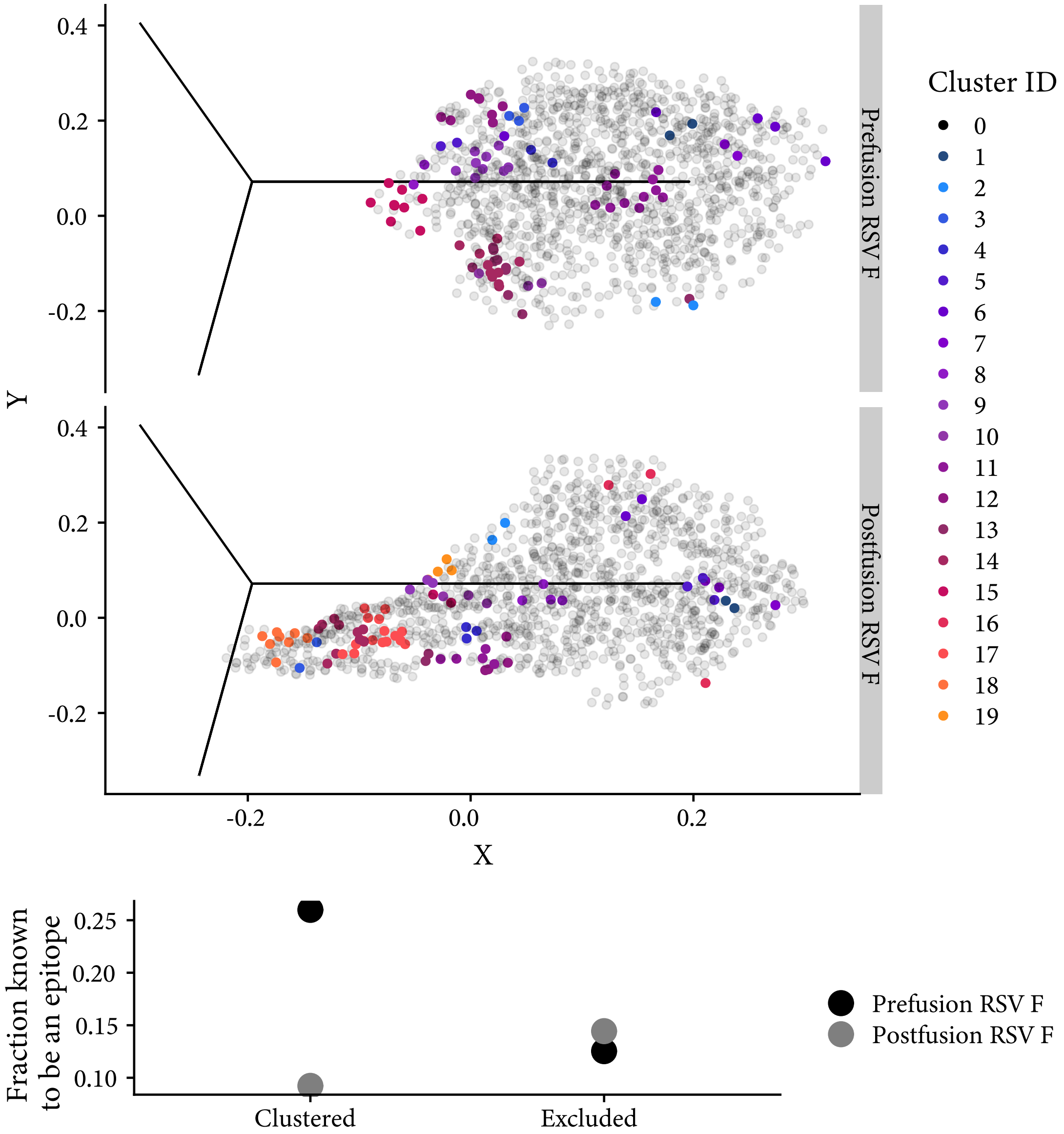


Figure C.2.: Clustered RSV F residues identified by high contact proximity deviation, total energy scores, and community detection and their enrichment for positively identified epitopes. *(Top) The x, y, and z axes represent the C_α atom coordinates of each residues with the pre- or post-fusion conformation. C_α coordinates not considered for clustering are shown in transparent grey, whereas clustered residues are designated by an opaque color. (Bottom) The bottom panel panel depicts the number of residues identified (Clustered) or not identified (Excluded) as epitopes by the clustering approach described in the last paragraph of section III.2.2 on page 45.*

APPENDIX D

STRUCTURAL BASIS FOR NONNEUTRALIZING ANTIBODY COMPETITION AT ANTIGENIC SITE II OF THE RESPIRATORY SYNCYTIAL VIRUS FUSION PROTEIN

This chapter is based on the publication "Structural basis for nonneutralizing antibody competition at antigenic site II of the respiratory syncytial virus fusion protein". Marion F. Sauer performed FPLC purification negative stain electron microscopy, particle picking, class averaging to provide the structural basis of binding angle of the 14N4-RSV 18537 B post-fusion F complex. Although not included in the manuscript, Marion F. Sauer also performed negative stain electron microscopy on additional Fab complexes with RSV DS-Cav1 pre-fusion RSV 18537 B post-fusion conformations, including Fabs 13A8 and 3J20. Due to the poor resolution, the negative stain electron microscopy class averages were insufficient in providing any structural data of these Fab-RSV F protein complexes. However, the observation of binding angle deviation in these initial experiments contributed to the development of the working hypothesis within "Structural basis for nonneutralizing antibody competition at antigenic site II of the respiratory syncytial virus fusion protein".

Palivizumab was the first antiviral mAb approved for therapeutic use in humans, and remains a prophylactic treatment for infants at risk for severe disease because of RSV. Palivizumab is an engineered humanized version of a murine mAb targeting antigenic site II of the RSV F protein, a key target in vaccine development. There are limited reported naturally occurring human mAbs to site II; therefore, the structural basis for human antibody recognition of this major antigenic site is poorly understood. Here, we describe a nonneutralizing class of site II-specific mAbs that competed for binding with palivizumab to postfusion RSV F protein. We also describe two classes of site II-specific neutralizing mAbs, one of which escaped competition with nonneutralizing mAbs. An X-ray crystal structure of the neutralizing mAb 14N4 in complex with F protein showed that the binding angle at which human neutralizing mAbs interact with antigenic site II determines whether or not nonneutralizing antibodies compete with their binding. Fine-mapping studies determined that nonneutralizing mAbs that interfere with binding of neutralizing mAbs recognize site II with a pose that facilitates binding to an epitope containing F surface residues on a neighboring protomer. Neutralizing antibodies, like motavizumab and a new mAb designated 3J20 that escape interference by the inhibiting mAbs, avoid such contact by binding at an angle that is shifted away from the nonneutralizing site. Furthermore, binding to rationally and computationally designed site II helix-loop-helix epitope-scaffold vaccines distinguished neutralizing from nonneutralizing site II antibodies.

D.1. Introduction

RSV is a highly contagious human pathogen, infecting the majority of infants before age 2 y, and is the leading cause of viral bronchiolitis and viral pneumonia in infants and children.[226,227] RSV remains a top priority for vaccine development, as thousands of deaths are recorded worldwide each year because of complications from infection.[228] To date, there is no licensed RSV vaccine. A major focus of RSV vaccine development has been inclusion of the RSV F protein, a class I fusion glycoprotein that is synthesized as a precursor and cleaved into two disulfide-linked fragments upon maturation into a trimer.[229] Although the RSV virion contains two additional surface proteins, the highly-glycosylated attachment (G) protein and the small hydrophobic protein, the F protein is highly conserved among strains of RSV strains and is the major target of protective neutralizing antibodies.

The F protein is known to adopt at least two major conformations: the metastable prefusion conformation and the postfusion conformation. Following attachment of the virion to a cell by the G protein, the F protein undergoes a dramatic structural rearrangement, resulting in fusion of the viral and cell membranes, and in cultured cells causes formation of cell syncytia. Four major neutralizing antigenic regions have been identified to date in the F protein, generally designated antigenic sites I, II, IV, and Ø, with the latter present only in the prefusion conformation. Site II is the target of palivizumab,[230] a prophylactic treatment licensed for use in high-risk infants during the RSV season. An RSV F protein subunit vaccine candidate comprising aggregates of the postfusion conformation of RSV F is being tested currently in clinical trials,[231] and serum antibody competition with palivizumab has been proposed as a potential serologic correlate of immunity for that vaccine.[232,233] We and others have isolated and studied RSV F-specific mAbs using murine hybridomas,[234] sorted macaque B cells,[224] and transformed human B cells or human antibody gene phage-display libraries.[235,236] Examples include mAbs 101F,[234] D25,[68] and the next-generation site II mAb motavizumab.[237] However, there are no reported naturally occurring human mAbs to site II, and palivizumab is an engineered humanized version of the murine mAb 1129.[238] Therefore, the repertoire of human antibodies interacting with site II and the structural basis for their recognition of this major antigenic site is poorly understood.

To characterize the human immune response to the RSV F protein, we isolated and characterized human mAbs targeting the RSV F protein, and in particular focused discovery efforts on antigenic site *II*. Defining the structural basis for interaction of site II-specific antibodies revealed new insights into the complexity of this site and diverse modes of recognition that determined whether or not site II competing human antibodies neutralize RSV.

D.2. Results

D.2.1. Antibody Isolation, Binding, and Neutralization

We isolated nine human mAbs from four human donors targeting the postfusion RSV F protein using human hybridoma technology (16). Transformed B cells generated from the B cells of adult human donors were screened by ELISA for reactivity to the RSV A2 F protein. Reactive cells were fused with myelomas to create hybridoma cell lines and plated in a 384-well plate. After 7 to 10 d, culture supernatants were screened for binding to recombinant, postfusion RSV A2 F protein. Cells from positive wells were expanded, respectively, into single wells in a 96-well culture plate using culture medium containing CpG, Chk2 inhibitor II, and irradiated heterologous human PBMCs. After 1 wk, culture supernatants were screened by ELISA for binding to recombinant, postfusion RSV A2 F protein. Clonal hybridomas were obtained by single-cell flow cytometric sorting, and isotyping analysis of purified mAbs showed them to be primarily of the IgG1 subclass (table D.1 on the next page). To assess whether the mAbs possessed neutralizing activity, purified mAbs were tested by a plaque reduction neutralization assay using RSV strain A2. As expected, serum from two donors neutralized RSV. Of

				Neutralization (IC_{50}; ng mL^{-1})	Binding to F protein for indicated strain (EC_{50}; ng mL^{-1})			
Donor	Monoclonal antibody	IgG subclass	Light chain	RSV A2	RSV A2	RSV A2 DS-Cav1	RSV A2 SC-TM	RSV 18537 B
2	4E7	1	λ	>	19	>	110	21
2	10F13	1	κ	>	17	66	93	21
1	14C16	1	κ	>	19	110	95	20
3	4B6	3	λ	>	24	>	130	24
1	9J5	1	κ	>	30	>	150	40
1	12I1	1	λ	>	26	>	250	33
1	14N4	1	κ	695	78	70	57	57
4	13A8	1	κ	55	82	62	52	64
2	3J20	1	κ	377	84	60	48	50
Control mAbs	Motivuzimab	1	κ	123	30	37	28	35
	101F	1	κ	402	50	62	80	45
	D25	1	κ	21	>	89	72	>

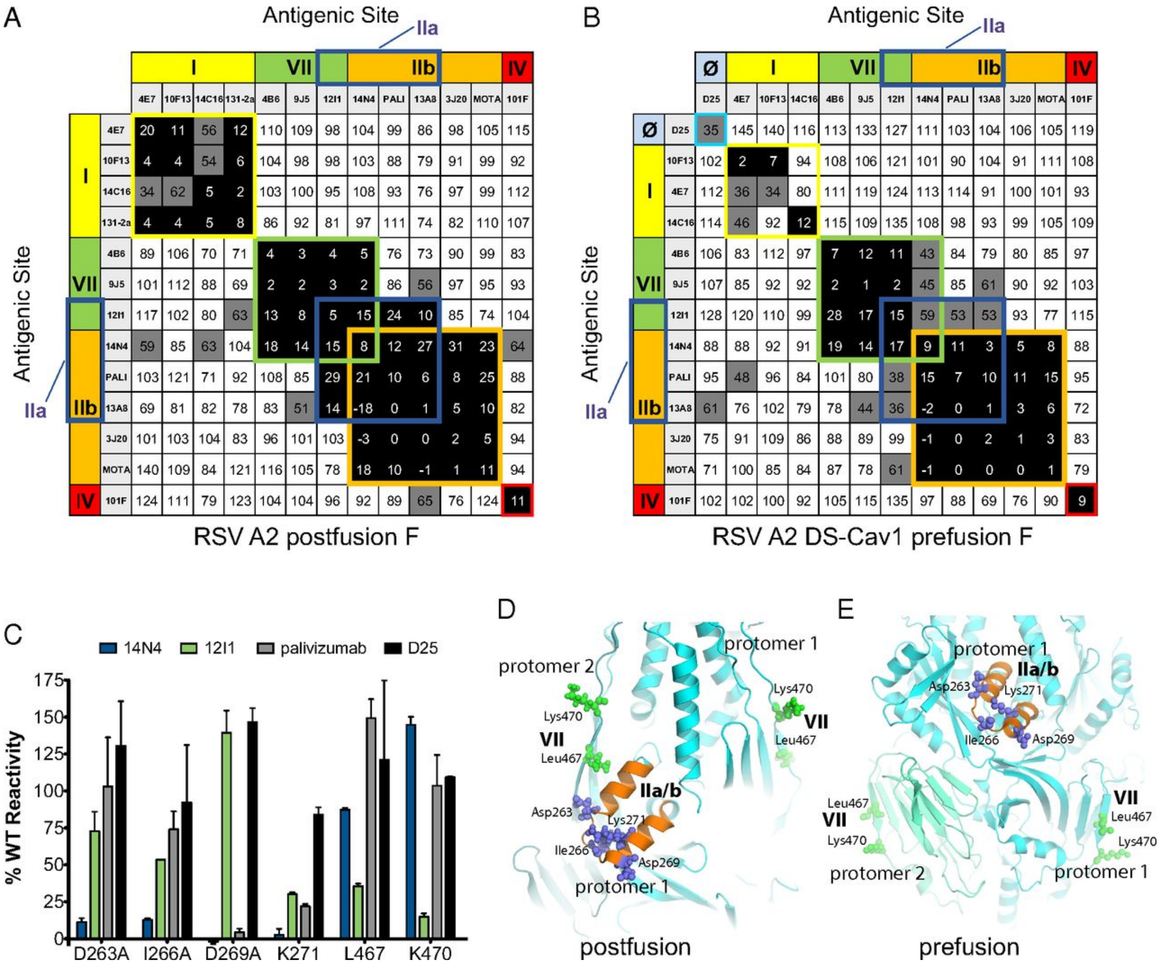
Table D.1.: Isotype, binding and neutralization features of nine new RSV F-specific human mAbs or control mAbs. EC_{50} *values correspond to the concentration at which half-maximum signal was obtained in ELISA, based on optical density at* 405 nm. *Neutralization values were determined using a plaque-reduction assay, where the* IC_{50} *corresponds to the mAb concentration at which* 50 % *plaque reduction was observed. > indicates no signal was detected below* 100 μg mL^{-1} *in neutralization assays and* 20 μg mL^{-1} *in ELISA binding assays; DS-Cav1 and SC-TM represent prefusion stabilized RSV F.*

the mAbs isolated, 14N4, 13A8, and 3J20 neutralized virus, whereas the remaining mAbs failed to show neutralization activity when tested at concentrations up to 100 μg mL^{-1}. These three neutralizing mAbs had IC_{50} values less than 1 μg mL^{-1} (table D.1 and Fig. S1). Recombinantly expressed site II mAb motavizumab (14), and previously described mAbs to site IV (101F)[234] and site Ø (D25),[68] were also tested for comparison. Mab 13A8 possessed potency similar to that of motavizumab and D25. mAbs were tested for binding by ELISA to postfusion or prefusion-stabilized disulfide-cavity filling (DsCav1) or single chain-triple mutant (SC-TM) RSV strain A2 F proteins[14,239] and postfusion F from RSV strain 18537 B (table D.1). Determination of EC_{50} values revealed that the three neutralizing mAbs bound to both prefusion and postfusion F proteins with equal affinity, agreeing with the conservation of the antigenic site II epitope between pre- and postfusion RSV F (table D.1). Furthermore, we did not detect major differences between binding to purified DSCav1 or SC-TM prefusion-stabilized F protein variants, suggesting the conformation of these antigens is similar at site II. Although the remaining mAbs did not neutralize RSV, EC_{50} values for binding in ELISA to postfusion F protein were similar for the neutralizing and nonneutralizing mAbs. These data suggest that the binding location or pose, rather than the affinity, is the critical determinant for RSV neutralization in this set of mAbs. MAbs 4E7, 4B6, 9J5, and 12I1 favored the postfusion conformation, based on differences in binding to stabilized prefusion versus postfusion F protein. Serum from two donors was also tested for binding, and no significant differences were observed among the two.

D.2.2. Epitope Binning Reveals the Complexity of Site II

To determine putative binding sites for the isolated mAbs, real-time competitionbinding studies were conducted with his-tagged RSV F proteins coupled to antipenta-his biosensor tips. We included recombinant forms of the previously described RSV mAbs 101F (site IV), 131-2a (site I),[240] palivizumab (site II), and motavizumab (site II) for comparative purposes in the competition-binding study on postfusion and prefusion F, because the epitopes for those mAbs have been defined previously. A complex array of five distinct competition-binding groups was observed for binding to postfusion F (figure D.1 on the next page). The groups containing mAbs binding to antigenic sites I, II, and IV were identified using the control mAbs. Three mAbs targeted site I, a neutralizing epitope present near the membrane proximal region of the F protein. However, none of these mAbs possessed neutralizing activity. The previously reported murine mAb 131-2a exhibits a low level of neutralizing activity,[68] but recognition of this epitope by human mAbs was not associated with neutralization, suggesting antigenic site I is not a major target of the human neutralizing antibody response. The remaining mAbs competed with antibodies directed to antigenic site II. Three mAbs (4B6, 9J5, 12I1) competed with site II-specific antibodies, yet failed to neutralize RSV, suggesting they do not bind in the correct orientation or they do not contact the full complement of critical amino acid residues in the site. The three neutralizing mAbs 14N4, 13A8, and 3J20 competed for binding to postfusion F with both palivizumab and motavizumab, as would be expected for mAbs targeting antigenic site II, yet subtle differences were observed among the competition patterns. mAb 3J20 differed from the other two by competing only with other neutralizing mAbs. The most potent mAb, 13A8, showed 50 % competition with the nonneutralizing mAb 9J5 and directly competed with 12I1. Interestingly, mAb 14N4 directly competed with all three nonneutralizing mAbs, forming a block of four mAbs containing both neutralizing and nonneutralizing mAbs. Furthermore, intermediate onedirectional competition was observed for 14N4 with site I mAbs 4E7 and 14C16.

Based on these data, it is apparent that mAbs competing for antigenic site II constitute at least three groups, which we designated antigenic sites IIa and IIb for neutralizing poses, and site VII for the nonneutralizing site. Antigenic site VII is represesented by the nonneutralizing mAb 12I1. Antigenic site IIb, containing mAb 3J20 and motavizumab, is a discrete competition group containing only neutralizing mAbs. Antigenic site IIa is an intermediate site, distinguished from site IIb as competing with both neutralizing and nonneutralizing mAbs, and is recognized by mAbs 14N4, 13A8, and palivizumab. Further differences in competition patterns within the site IIa group of mAbs were observed, as 14N4 competes with all three nonneutralizing mAbs, 13A8 competes with two, and palivizumab competes with one, suggesting a gradient of binding poses occur at antigenic site IIa between sites VII and IIb. We also tested competition using prefusion F (DSCav1) as the immobilized antigen, and included the prefusionspecific mAb D25 for comparison (figure D.1 on the following page). Although site VII mAbs do not bind well to prefusion F protein by ELISA, we observed significant binding in biolayer interferometry experiments, allowing competition studies to be conducted with prefusion F. A similar pattern of three distinct groups was observed for antigenic site II in prefusion F; however, competition at site IIa was weaker among mAbs in the group, suggesting sites VII and IIa may be further apart in the prefusion than in the postfusion conformation. Such a complex array of competition-binding groups was unexpected, because the site II mAb palivizumab, which is used in prophylactic treatment, also bidirectionally competed with the nonneutralizing mAb 12I1. A palivizumabcompetition assay designed to detect the presence of site II antibodies in immune serum by competing with palivizumab[232,233] has been proposed as a correlate of immunity for an RSV postfusion F protein vaccine candidate.



		I				VII			IIb					IV
		4E7	10F13	14C16	131-2a	4B6	9J5	12I1	14N4	PALI	13A8	3J20	MOTA	101F
I	4E7	20	11	56	12	110	109	98	104	99	86	98	105	115
	10F13	4	4	54	6	104	98	98	103	88	79	91	99	92
	14C16	34	62	5	2	103	100	95	108	93	76	97	99	112
	131-2a	4	4	5	8	86	92	81	97	111	74	82	110	107
VII	4B6	89	106	70	71	4	3	4	5	76	73	90	99	83
	9J5	101	112	88	69	2	2	3	2	86	56	97	95	93
	12I1	117	102	80	63	13	8	5	15	24	10	85	74	104
IIb	14N4	59	85	63	104	18	14	15	8	12	27	31	23	64
	PALI	103	121	71	92	108	85	29	21	10	6	8	25	88
	13A8	69	81	82	78	83	51	14	-18	0	1	5	10	82
	3J20	101	103	104	83	96	101	103	-3	0	0	2	5	94
	MOTA	140	109	84	121	116	105	78	18	10	-1	1	11	94
IV	101F	124	111	79	123	104	104	96	92	89	65	76	124	11

		Ø	I			VII			IIb					IV
		D25	4E7	10F13	14C16	4B6	9J5	12I1	14N4	PALI	13A8	3J20	MOTA	101F
Ø	D25	35	145	140	116	113	133	127	111	103	104	106	105	119
I	10F13	102	2	7	94	108	106	121	101	90	104	91	111	108
	4E7	112	36	34	80	111	119	124	113	114	91	100	101	93
	14C16	114	46	92	12	115	109	135	108	98	93	99	105	109
VII	4B6	106	83	112	97	7	12	11	43	84	79	80	85	97
	9J5	107	85	92	92	2	1	2	45	85	61	90	92	103
	12I1	128	120	110	99	28	17	15	59	53	53	93	77	115
IIb	14N4	88	88	92	91	19	14	17	9	11	3	5	8	88
	PALI	95	48	96	84	101	80	38	15	7	10	11	15	95
	13A8	61	76	102	79	78	44	36	-2	0	1	3	6	72
	3J20	75	91	109	86	88	89	99	-1	0	2	1	3	83
	MOTA	71	100	85	84	87	78	61	-1	0	0	0	1	79
IV	101F	102	102	100	92	105	115	135	97	88	69	76	90	9

Figure D.1.: Epitope binning and saturation alanine scanning mutagenesis for mAbs binding RSV F protein in the postfusion (A) or DS-Cav1 prefusion (B) conformations. *Data indicate the percent binding of the competing antibody in the presence of the primary antibody, compared with the competing antibody alone. Cells filled in black indicate full competition, in which ≤33 % of the uncompeted signal was observed, intermediate competition (gray) if signal was between 33 % and 66 %, and noncompeting (white) if signal was ≥ 66 %. Antigenic sites are highlighted at the top and side based on competition-binding with the control mAbs D25 (site Ø), 131-2a (site I), palivizumab (PALI) or motavizumab (MOTA) (site II), or 101F (site IV). Competition for antigenic site II mAbs formed three groups, corresponding to site VII (green border), IIa (blue border), or IIb (orange border). Competition with nonneutralizing mAbs was less pronounced in the prefusion conformation. (C) Binding values for isolated mAbs 14N4 and 12I1 with palivizumab or D25 control mAbs. The mAb reactivity for each RSV F mutation was calculated relative to that of wild-type RSV F. Error bars indicate the measurement range. (D) The residues important for binding of 14N4 (blue) or 12I1 (green) are mapped on the RSV F trimeric structure as spheres. Residues important for 14N4 and 12I1 binding are very distant on the same protomer, yet are in close contact through quaternary interactions at the protomer 1–protomer 2 interface, leading to competition between neutralizing mAb 14N4 and nonneutralizing mAb 12I1. (E) Quaternary interactions between antigenic sites IIa and VII were less pronounced in the prefusion conformation, as site IIa is farther away from site VII on the same and adjacent protomers.*

Indeed, we repeated the competition using published palivizumab competition assay protocols,[232] where biotinylated palivizumab was spiked into control mAbs, as well as donor serum. As expected, we observed donor serum neutralized RSV and competed with palivizumab at low dilutions. Furthermore, mAbs 14N4 and 12I1 both competed with palivizumab, with 12I1 showing competition only on postfusion F, similar to the competition data in figure D.1 on the previous page.

Based on the data described, it appears motavizumab and 3J20-like mAbs may be better candidates for this purpose, as competition with these mAbs is observed only with neutralizing mAbs, but the palivizumab-competing antibody population contains a proportion of nonneutralizing mAbs. To determine if the nonneutralizing mAb 12I1 blocked neutralization of palivizumab or 14N4, we incubated mAb 12I1 with virus initially before applying the neutralizing mAbs. No significant difference was observed between those samples incubated with 12I1 and control mAbs. This finding is expected as 12I1 favors the postfusion conformation (table D.1 on page 119), which allows membrane fusion by the F protein before 12I1 binding. Thus, the site VII mAbs do not inhibit neutralization, yet are likely produced in response to a postfusion F immunogen, and also affect the palivizumab competition assay.

D.2.3. Saturation Alanine Scanning Mutagenesis

To better understand the complexity of antigenic site II and the specificity of mAbs recognizing the site, we performed saturation alanine scanning mutagenesis to identify residues critical for the binding of the neutralizing mAb 14N4 or nonneutralizing mAb 12I1. Residues Asp263, Ile266, Asp269, and Lys271 were critical for 14N4 binding (Panel C in figure D.1 on the previous page). Interestingly, we previously identified a Ile266Met mutation when generating monoclonal antibody-resistant mutant (MARM) virus by in vitro selection using the RSV F targeting human Fab19[235] isolated from a phage-display library. Based on the X-ray crystal structure of the RSV F protein (Panel D in figure D.1 on the preceding page), Ile266 is positioned at the bottom of the antigenic site II helix– loop–helix motif and is pointed toward the inner protein core, suggesting the residue disrupts the antigenic motif by allosteric effects. In the same study,[235] selection with several murine mAbs produced MARM viruses with Lys272Asn, and similarly, selection with palivizumab in vitro or in vivo, generated similar MARM viruses with the following mutations: Lys272Met, Lys272Gln, and Asn268Ile.[241,242] The Lys272Gln MARM virus completely resisted prophylactic palivizumab treatment.[243] Unexpectedly, mutagenesis scanning for the site VII mAb 12I1 revealed critical residues over 40 Å away in the RSV F monomer: Leu467 and Lys470 (Panels C in figure D.1 on the previous page). Although the site VII mAb 12I1 and site IIa mAb 14N4 competed for binding, the critical residues for binding were quite different, with site VII residues falling on the 47 Å extended loop connecting the lower structured portion to the helix bundle in a single protomer of F in postfusion conformation (Panel D in figure D.1 on the preceding page). However, when the F protein is viewed as a trimeric structure, all residues in antigenic sites VII and IIa come in close proximity through quaternary interactions. Antigenic site IIa in one protomer of F in the trimer is within 13 Å of antigenic site VII on an adjacent protomer. Although a quaternary epitope for RSV F has been described for the mAb AM14,[244] the site VII/IIa mAb competition is the first described example of quaternary interactions contributing to nonneutralizing mAb competition with a neutralizing mAb. In the prefusion conformation (Panel E in figure D.1 on the previous page), antigenic sites VII and IIa are farther apart than in the postfusion form. Antigenic site IIa is equidistant from site VII on the same and the adjacent protomer. This difference confirms the observation in the epitope binning studies in which competition on prefusion F between antigenic sites IIa and VII was less pronounced than in the postfusion conformation. The intermediate level of competition for binding to the prefusion form of F between sites VII and IIa mAbs was consistent for mAbs 14N4, 13A8, and palivizumab.

D.2.4. Structure of the 14N4-Fab–RSV F Complex

Because 14N4 is a unique mAb, competing not only with palivizumab but also with nonneutralizing mAbs, we next sought to determine the structural basis for competition of 14N4 with other mAbs recognizing site II. The 14N4 fragment antigen-binding region (14N4-Fab) was crystallized in spacegroup P 1 21 1 and the structure was solved to 2.0 Å with Rwork/Rfree = $\frac{19.5}{21.0}$ % . 14N4-Fab then was incubated with postfusion RSV A2 F, and the complex was isolated by size-exclusion chromatography and crystallized in spacegroup P 42 21 2. After screening with numerous cryoprotectants and attempts at data collection at room temperature, the best X-ray diffraction of the complex was to 4.1 Å. The crystal structures of postfusion RSV F and 14N4 variable and constant Fab regions were used in molecular replacement to solve the structure of the complex with Rwork/ Rfree = $\frac{25.6}{28.2}$ %, refined using noncrystallographic symmetry (NCS) torsion and reference-model restraints. Separate searches were needed for the variable and constant regions of the 14N4-Fab region as the constant region was shifted 56° from the apo–14N4-Fab structure, an observation likely attributed to crystal packing, as the constant region makes contacts to the next asymmetric unit. The asymmetric unit is composed of the RSV F trimer with three 14N4-Fab molecules, one at each protomer of RSV F (Panel A in figure D.2 on the following page. Electron density for the RSV F protein and the three 14N4-Fab variable regions was well defined, especially at each interface between the two molecules. To confirm binding at antigenic site II in RSV strain 18537 B, we complexed 14N4 with RSV 18537 B postfusion F and class-averages determined from negative-stain EM images indicated the position and orientation of the 14N4-Fab molecules were similar to those in the X-ray crystal structure (Panel A in figure D.2 on the next page). The HCDR3 of 14N4-Fab nestles between the two helices in the antigenic site II motif, where several hydrophobic residues exist. Residues in the RSV F structure important for binding based on alanine scanning mutagenesis are highlighted in Panel B of figure D.2 on the following page, where they make key interactions with 14N4-Fab. Asp263 is within hydrogen bonding distance of the backbone Gly56 on 14N4, and Lys271 likely interacts with the HCDR3 by hydrogen bonding with Thr107 (Panel B in figure D.2 on the next page). Furthermore, the light-chain also appears important for binding, because Asn99 and Ser37 of the LCDR1 are in close contact with Asp269. Lys272 is near of the LCDR2 Asp57, although this residue was not critical for binding in mutagenesis scanning experiments. As expected, interactions were not observed for Ile266, as this residue is buried at the base of the helix–loop–helix motif.

Compared with the structure of motavizumab in complex with the site II peptide, striking differences were observed. Overlaying at antigenic site II, the motavizumab angle of binding is significantly different, as it is shifted 42° from the 14N4 binding region in the direction away from the 12I1 site VII (Panel C in figure D.2 on the following page). This structural difference correlates with the lack of competition between antigenic site IIb mAbs motavizumab and 3J20, and the antigenic site VII nonneutralizing mAbs binding at Leu467 and Lys470. 14N4 could indeed block the binding of 12I1, because its binding pose is predicted to be shifted just 27° from site VII. However, motavizumab is shifted away from site IIa enough to prevent competition with mAb 12I1. Considering critical binding interactions, we noted that motavizumab hydrogen bonds to Asp263 using Asp54 (HCDR2) distantly, to Lys272 with Asp50 (LCDR2), and Asp269 using Ser92 (LCDR3) (Panel D in figure D.2 on the next page). Interestingly, motavizumab bypasses Lys271, leaving no residues in the vicinity with which to interact. This positioning causes a shift away from site VII, as the majority of the interactions are

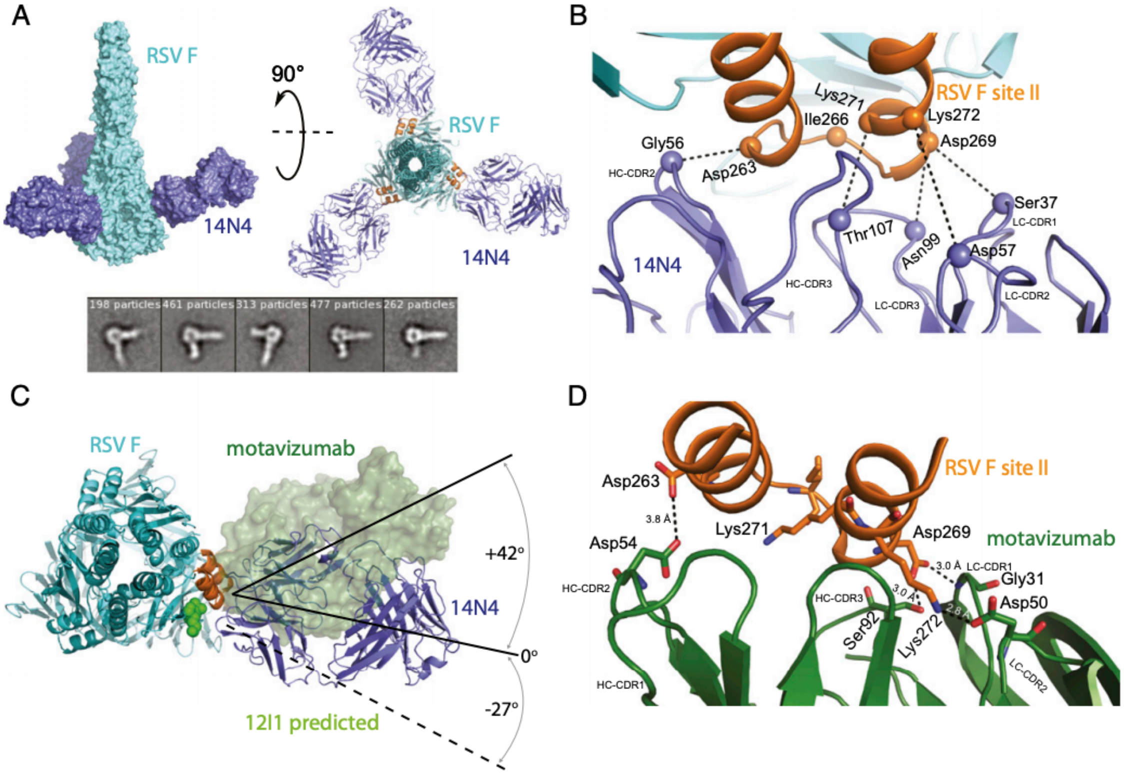


Figure D.2.: The complex of 14N4-Fab with RSV F. *(A) X-ray crystal structure of 14N4-Fab (blue) in complex with postfusion RSV strain A2 F protein (cyan). The overall structure is displayed in surface form and rotated 90° in cartoon form. 14N4-Fab bound RSV F at each protomer in the trimeric structure. EM class averages with RSV 18537 B are also displayed, confirming the binding location of 14N4-Fab. The side length of panels is 32.7 nm. (B) Chemical interactions between Fab 14N4 and RSV strain A2 F protein. Several key hydrogen bonds are important for molecular recognition. (C) Overlay of the complex with the motavizumab–site II peptide complex (PDB ID code 3IXT). Motavizumab is shown in green surface form, RSV F in cyan, and 14N4-Fab in blue. The antigenic site II region is colored in orange, and residues important for site VII binding are shown as spheres in light green. Motavizumab binds antigenic site II at a different orientation than 14N4-Fab, allowing it to be free of interactions with site VII. (D) Interactions between motavizumab and the antigenic site II peptide (PDB ID code 3IXT). Lys271 does not interact with motavizumab, unlike its role in the 14N4–RSV F complex.*

involved on the right helix, rather than the left helix, where only hydrophobic interactions exist with the motavizumab HCDR3.

D.2.5. Human Antibodies Bind Scaffold-Based Immunogens

Attempts to generate a vaccine against RSV have been largely unsuccessful, and the presence of nonneutralizing mAbs competing with neutralizing mAbs may contribute to this problem. We and others have recently reported structure-based designed vaccine candidates for presenting the site II immunogen. Strategies included a stable trihelix scaffold protein purpose-built to support the helix–loop–helix motif of antigenic site II (FFL_001),[224] a Fab-based scaffold for site II,[245] and also a strategy in which the RSV F site II was grafted onto the metapneumovirus (MPV) F protein (RPM-1) to generate a chimeric protein capable of inducing a cross-reactive immunogenic response[246] (Panel A in figure D.3 on the following page). Each of these three epitope-based scaffolds induced at least partial immune responses in mice to RSV F, and the FFL_001 vaccine candidate induced reasonable titers of neutralizing mAbs from immunized macaques. We tested binding by ELISA of the three neutralizingsite II human mAbs 14N4, 13A8, and 3J20 to FFL_001 and RPM-1 and found that they did bind, as did palivizumab and motavizumab used as positive controls (Panel B in figure D.3 on the next page). EC_{50} values for binding of the mAbs to the scaffolded epitopes were similar to those obtained for the RSV F protein, suggesting antigenic site II is the primary region necessary for human mAb binding. This finding also is consistent with the X-ray crystallography and EM structural data for the 14N4-Fab–RSV F complex. Interestingly, binding was not detected for the nonneutralizing mAb 12I1 or other antigenic site VII mAbs to either FFL_001 or RPM-1 scaffold proteins. Therefore, binding to the scaffolded epitopes distinguishes neutralizing from nonneutralizing site VII competing antibodies. Surface plasmon resonance revealed very low KD values for the three neutralizing mAbs (Panel D in figure D.3 on the following page, suggesting limited residues are needed for Fab binding to antigenic site II, a finding consistent with the X-ray structure of 14N4-Fab with RSV F, as no molecular contacts were observed outside site II. However, additional interacting residues may be present in 14N4 binding to prefusion RSV F. Binding was not detected to a mutated FFL_001 control.

To confirm the binding location for 14N4 to the FFL_001 scaffolded epitope, we performed hydrogen-deuterium (HD) exchange mass spectrometry (Panel A in figure D.4 on page 127. We mapped the majority of the 14N4- Fab region, and the peptides with the largest decrease in deuterium exchange in the bound state were localized to the HCDR3 loop, with a limited effect in the LCDR2. This finding iscompared with 14N4 and motavizumab (Panel C in figure D.4 on page 127. mAb 17HD9 is positioned further left than 14N4, close to antigenic site VII, suggesting that 17HD9 would compete with 12I1 and other site VII mAbs. Indeed, we observed such competition between recombinantly expressed mAb 17HD9 and site VII mAbs.

mAb 14N4 uses VH3-53 and JH4 gene segments to encode the expressed antibody. Because of the paucity of human antibodies that target RSV antigenic site II, it was unclear if this mAb is unique among human donors, or if 14N4-like mAbs exist that do compete with nonneutralizing mAbs in the general population. To help answer this question, we searched a database of 50 million antibody heavy-chain variable sequences obtained from 31 adult human subjects, and found similar sequences in 31 individuals that used VH3-53 and JH4 gene segments and shared 85 % similarity in the HCDR3. When the HCDR3 identity cut-off for matching was extended to 100 %, the majority of sequence matches remained. These sequence homology data suggest that 14N4-like mAbs may be common in the human population, and the presence of nonneutralizing mAbs competing with neutralizing mAbs

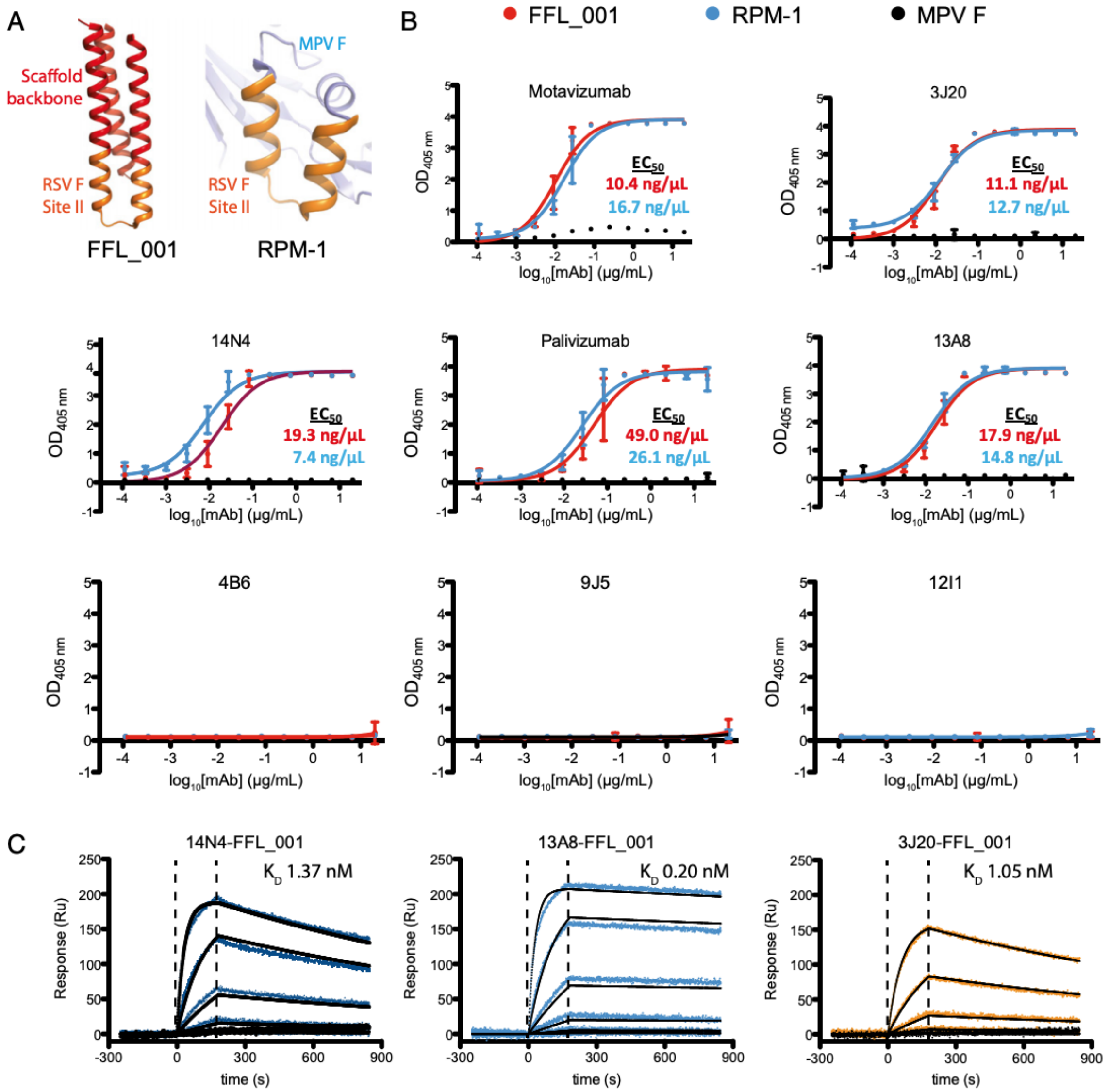


Figure D.3.: Human mAbs bind to synthetic immunogens. *(A) X-ray structure of FFL_001 displayed in red with RSV antigenic site VII shown in orange (PDB ID code 4JLR). A model of RPM-1 shows the region surrounding the corresponding antigenic site VII in the MPV F protein (blue, PDB ID code 4DAG), and RSV antigenic site II is overlaid in orange. (B) ELISA binding curves for three human mAbs 14N4, 13A8, and 3J20 along with antigenic site VII mAbs motavizumab and palivizumab. Binding curves for FFL_001 are in red and for RPM-1 are in blue. Binding to MPV F protein is shown in black. EC50 values are displayed for each, in corresponding colors. Error bars indicate 95% confidence intervals. (C) Surface plasmon resonance of 14N4, 13A8, and 3J20 Fabs binding to FFL_001 with calculated KD values displayed. Colored data points are overlaid with the curve fit line in black. Dotted lines indicate the start of association and dissociation steps.*

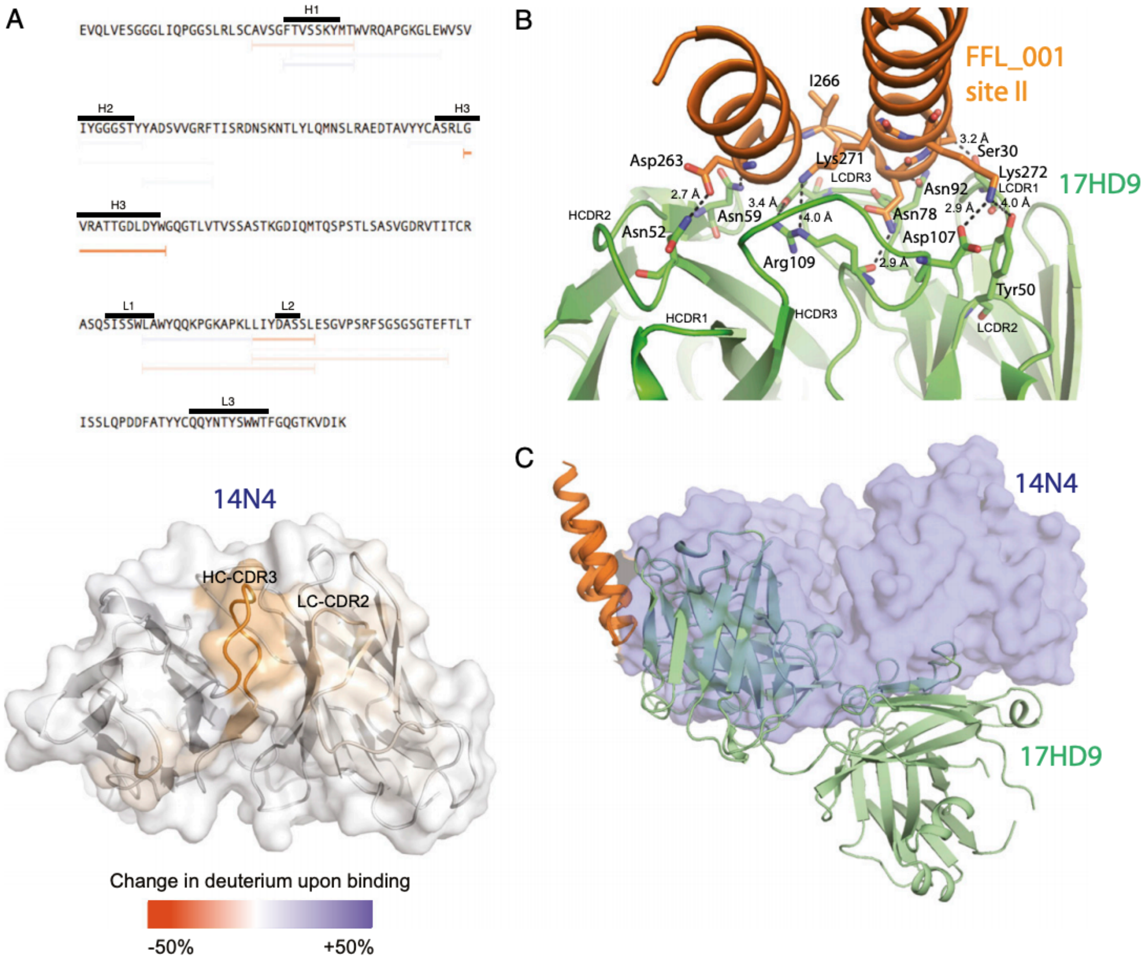

A
H1
EVQLVESGGGLIQPGGSLRLSCAVSGFTVSSKYMTWVRQAPGKGLEWVSV
H2
H3
IYGGGSTYYADSVVGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCASRLG
H3
VRATTGDLDYWGQGTLVTVSSASTKGDIQMTQSPSTLSASVGDRVTITCR
L1
L2
ASQSISSWLAWYQQKPGKAPKLLIYDASSLESGVPSRFSGSGSGTEFTLT
L3
ISSLQPDDFATYYCQQYNTYSWWTFGQGTKVDIK
14N4
HC-CDR3
LC-CDR2
Change in deuterium upon binding
-50%
+50%
B
FFL_001
site II
I266
Asp263
Lys271
3.2 Å
Ser30
Lys272
LCDR1
17HD9
LCDR3
Asn92
2.7 Å
3.4 Å
Asn59
4.0 Å
Asn78
2.9 Å
4.0 Å
HCDR2
Asn52
Asp107
Arg109
2.9 Å
Tyr50
HCDR1
HCDR3
LCDR2
C
14N4
17HD9

Figure D.4.: HD exchange with FFL_001 and comparison with mab 17HD9. *(A) HD exchange protection of 14N4 upon scaffold binding. Each antibody-derived peptide was monitored for deuterium incorporation in the presence or absence of the scaffold protein. Peptides are colored according to the difference in incorporated deuterium atoms in the bound vs. unbound form, with a large reduction in incorporation indicating a putative binding site (orange). Values from the* 30 min *deuteration time point are shown. The HD exchange profile of 14N4-derived peptides is mapped onto the 14N4 Fab structure. (B) Interactions between the macaque Fab 17HD9 and FFL_001 (PDB ID code 4N9G). (C) Overlay of 14N4 with antigenic site II and 17HD9 with FFL_001. 14N4 is displayed as surface form in blue, and 17HD9 in pale-green (PDB ID code 4N9G). 17HD9 interacts with the lower loop of antigenic site II along with both helices, whereas 14N4 interacts only with the two helices.*

may be a common feature in human RSV immune responses.

D.3. Discussion

Although palivizumab has been used as a prophylactic treatment for high-risk infants during RSV season for nearly two decades, no vaccine is currently approved for protection against RSV. Vaccine strategies have been proposed that focus on the 150 kDa postfusion RSV F trimeric protein to elicit an immune response, yet antibody production is directed toward both protective and nonprotective epitopes. We have shown in the described human mAbs evidence for substantial neutralizing/nonneutralizing mAb competition binding at antigenic site II. Considering the competition patterns, antigenic site II was delineated into two subsites based on epitopes on adjacent protomers of the RSV F trimer, and a new region, site VII, was characterized as a nonneutralizing antigenic site that competes with site II. Based on the X-ray structure of 14N4 in complex with RSV F, subtle changes in the binding pose can cause substantial effects in competing antibodies. Although the competition was described here for RSV, these data may inform general vaccine design, as nonneutralizing antibody production is a common occurrence during viral infection. Furthermore, studying the B-cell response of vaccinated individuals in clinical trials will assist in determining the extent of neutralizing/nonneutralizing mAb competition in human sera.

Competition between 14N4 and 12I1 mAbs on postfusion F is readily observed, as the 12I1 site VII is in close proximity to antigenic site IIa. However, the competition was less pronounced in the prefusion conformation, as sites VII and IIa are not in close proximity before the pre- to postfusion rearrangement. Because 12I1 favors the postfusion conformation (Table 1), vaccine strategies involving prefusion F may be more beneficial to avoiding the competing interactions at antigenic site II. Indeed, 12I1 was likely generated against the RSV F postfusion conformation, and these 12I1-like mAbs may not have been isolated if prefusion F was used in the initial B-cell isolation. Future experiments detailing the mAb response to prefusion F will be beneficial in determining the overall impact of the competition with nonneutralizing mAbs. When assessing vaccine efficacy using competition with palivizumab, nonneutralizing antibody competition with palivizumab must be taken into account, especially in vaccine candidates using postfusion RSV F. We further propose using motavizumab or other 3J20-like mAbs rather than palivizumab in serum antibody competition-binding assays to monitor neutralizing mAbs, as motavizumab competes only with neutralizing mAbs.

As an alternative to full-length RSV F as a vaccine strategy, our data support the concept of using scaffold-based epitopes for immunization against RSV. For example, FFL_001 avoids the potential for nonneutralizing 12I1-like mAb production to compete for binding with neutralizing 14N4-like mAbs, because onlythe neutralizing epitope is present for an immune response, unlike RSV F, where the 12I1 site VII is on an adjacent protomer. Binding to RPM-1 also provides insight into the neutralizing site II epitope, because homologous residues exist in the MPV protein near site VII, yet nonneutralizing RSV-specific antibodies do not bind RPM-1. Thus, these scaffold-based immunogens can be used to identify neutralizing mAbs targeting site II, instead of intact RSV F, which also binds nonneutralizing antibodies. As potential vaccines, epitope-scaffold immunogens would not induce site VII mAbs, likely producing only neutralizing mAbs to antigenic site II.

In summary, careful study of the fine specificity of new human antibodies to the RSV F antigenic site II revealed important structural features that inform next-generation vaccine design and testing, and provide potently neutralizing candidate prophylactic human mAbs.

D.4. Materials and Methods

D.4.1. ELISA for Binding to RSV F Protein

For recombinant protein capture ELISA, 384-well plates were treated with 2 µg/mL of antigen for 1 h at 37 °C or overnight at 4 °C. Following this procedure, plates were blocked for 1 h with 2 % (wt/vol) milk supplemented with 2 % (vol/vol) goat serum. Primary mAbs and culture supernatants were applied to wells for 1 h following three washes with PBS-T. Plates were washed with PBS-T four times before applying 25 µL secondary antibody (goat anti-human IgG Fc; Meridian Life Science) at a dilution of 1:4000 in blocking solution. After 1 h incubation, the plates were washed five times with PBS-T, and 25 µL of phosphatase substrate solution (1 mg mL^{-1}) phosphatase substrate in 1 M Tris aminomethane (Sigma) was added to each well. The plates were incubated at room temperature before reading the optical density at 405 nm on a Biotek plate reader. The palivizumab competition assay ELISA was conducted by coating ELISA plates with the desired 2 µg mL^{-1} of the desired antigen. Next, serially diluted competing mAbs spiked with 50 ng mL^{-1} biotinylated palivizumab were added to the plates. Alternatively, serially diluted serum was spiked with 50 ng mL^{-1} biotinylated palivizumab. Control wells contained PBS with 50 ng mL^{-1} biotinylated palivizumab. Palivizumab was biotinylated using the EZ-Link NHS PEG4 Biotinylation Kit (ThermoFisher) following the manufacturer's protocol. After 1-h incubation, the plates were washed with PBS-T and streptavidin-HRP (ThermoFisher) diluted 1:4000 in blocking solution was applied for 1 h. After a washing step, plates were incubated with one-step Ultra TMB solution (ThermoFisher). The reaction was stopped by adding an equal volume of 1 M HCl. Plates were read on a Biotek plate reader at 450 nm.

D.4.2. Human Hybridoma Generation

Participation of healthy human adult subjects was approved by the Vanderbilt University Institutional Review Board, and blood samples were obtained only after informed consent. PBMCs were isolated from human donor blood samples using Ficoll-Histopaque density gradient centrifugation. Approximately 10 million PBMCs were mixed with 17 mL of ClonaCell-HY Medium A (StemCell Technologies), 8 µg mL^{-1} of CpG (phosphorothioate-modified oligodeoxynucleotide ZOEZOEZZZZZOEEZOEZZZT (Invitrogen), 3 µg mL^{-1} of Chk2 inhibitor II (Sigma), 1 µg mL^{-1} of cyclosporine A (Sigma), and 4.5 mL of filtered supernatant from a culture of B95.8 cells (ATCC VR-1492) containing Epstein-Barr virus and plated in a 384-well plate. After 7 to 10 d, culture supernatants were screened for binding to recombinant, postfusion RSV strain A2 F protein and FFL_001. Cells from positive wells were expanded into single wells in a 96-well culture plate using culture medium containing 8 µg mL^{-1} CpG, 3 µg mL^{-1} Chk2 inhibitor II, and irritated heterologous human PBMCs (Nashville Red Cross). After 1 wk, culture supernatants were screened by ELISA for binding to recombinant, postfusion RSV A2 F protein and FFL_001. Cells from positive wells were fused with HMMA2.5 myeloma cells by electrofusion.[247] Fused cells were plated in 384- well plates in growth medium containing 100 µM hypoxanthine, 0.4 µM aminopterin, 16 µM thymidine (HAT Media Supplement, Sigma), and 7 µM ouabain (Sigma). Hybridomas were screened after 2 wk for mAb production by ELISA, and cells from wells with reactive supernatants were expanded to 48- well plates for 1 wk before being screened again by ELISA, and then subjected to single-cell fluorescence-activated sorting. After cell sorting into 384-well plates containing Medium E (StemCell Technologies), hybridomas were screened by ELISA before expansion into both 48-well and 12-well plates.

D.4.3. Human mAb and Fab Production and Purification

Hybridoma cells lines were expanded in Medium E until 80 % confluent in 75 cm^2 flasks. For antibody production, cells from one 75 cm^2 cell culture flask were collected with a cell scraper and expanded to four 225 cm^2 cell culture flasks in serum-free medium (Hybridoma-SFM, Gibco). After 21 d, supernatants were sterile filtered using 0.45 μm pore size filter devices. For antibody purification, HiTrap MabSelectSure columns (GE Healthcare Life Sciences) were used to purify antibodies using the manufacturer's protocol. To obtain Fab fragments, papain digestion was used (Pierce Fab Preparation Kit, Thermo Scientific). Fab fragments were purified by removing IgG and Fc contaminants using a HiTrap MabSelectSure followed by purification with an anti-CH1 column (GE Healthcare Life Sciences).

D.4.4. Production and Purification of Recombinant RSV F Protein RSV mAbs, and Epitope Immunogens

Plasmids encoding cDNAs for RSV subgroup A strain A2 or subgroup B strain 18537 prefusion (DS-Cav1) and postfusion F protein constructs (a gift from Barney Graham, Viral Pathogenesis Laboratory, National Institutes of Health, Bethesda) were expanded in Escherichia coli DH5α cells and plasmids were purified using Qiagen Plasmid Maxiprep kits (Qiagen). Prefusion-stabilized RSV F SC-TM was synthesized (Genscript). Plasmids encoding cDNAs for the the protein sequences of mAb 101F and mAb D25 were synthesized (Genscript), and heavy- and light-chain sequences were cloned into vectors encoding human IgG1 and λ or κ lightchain constant regions, respectively. MAb 131-2a protein was obtained from Sigma. Commercial preparations of palivizumab (Medimmune) were obtained from the pharmacy at Vanderbilt University Medical Center. For each liter of protein expression, 1.3 mg of plasmid DNA was mixed with 2 mg of polyethylenimine in Opti-MEM I + GlutaMAX cell culture medium (Fisher). After 10 min, the DNA mixture was added to HEK293 cells at 1 × 106 cells per milliliter. The culture supernatant was harvested after 6 d, and the protein was purified by HiTrap Talon crude (GE Healthcare Life Sciences) column for RSV F protein variants or HiTrap MabSelectSure columns for mAbs, following the manufacturer's protocol. 14N4-Fab heavy and light variable region DNA was synthesized (Genscript) and cloned into vectors containing human CH1 and kappa sequences. 14N4-Fab was expressed in Expi293 (Invitrogen) cells using Expifectamine 293 (Invitrogen) following the manufacturer's protocol. Recombinant Fab was purified using anti-CH1 Capture Select column (GE Healthcare Life Sciences). FFL_001, FFL_001 mutant proteins, and RPM-1 were expressed and purified as described previously.[224,246] mAb 17HD9 was expressed in expi293F cells following the manufacturer's protocol, and using the vectors described previously.[224]

D.4.5. RSV Plaque Neutralization Experiments

mAbs isolated from hybridoma supernatants were incubated 1:1 with a suspension of infectious RSV strain A2 for 1 h. Following this process, confluent HEp-2 cells, maintained in Opti-MEM I+GlutaMAX (Fisher) supplemented with 2 % (vol/vol) FBS at 37 °C in a CO_2 incubator, in 24-well plates, were inoculated with 50 μL of the antibody:virus or serum:virus mixture for 1 h. After the hour, cells were overlaid with 1 mL of 0.75 % methylcellulose dissolved in Opti-MEM I + GlutaMAX. Cells were incubated for 4 d after which the plaques were visualized by fixing cells with 10 % (vol/vol) neutral-buffered formalin and staining with Crystal violet. Plaques were counted and compared with a virus control. Data were analyzed with Prism software (GraphPad) to obtain IC_{50} values. To determine competition with 12I1, virus was first mixed with 40 μg mL^{-1} 12I1 for 1 h. The virus:12I1 mixture was overlaid onto serial dilutions of 14N4 and palivizumab for 1 h. The rest of the process was completed as described above.

D.4.6. Assays for Competition-Binding

After obtaining an initial baseline in kinetics buffer (ForteBio; diluted 1:10 in PBS), 10 μg mL^{-1} of his-tagged RSV F protein was immobilized onto antipenta-his biosensor tips for a biolayer interferometry instrument (Octet Red, ForteBio) for 120 s. The baseline signal was measured again for 60 s before biosensor tips were immersed into wells containing 100 μg mL^{-1} primary antibody for 300 s. Following this process, biosensors were immersed into wells containing 100 μg mL^{-1} of a second mAb for 300 s. Percent binding of a second mAbs in the presence of the first mAb was determined by comparing the maximal signal of the second mAb after the first mAb was added to the maximum signal of the second mAb alone. mAbs were considered noncompeting if maximum binding of the second mAb was ≥66 % of its uncompeted binding. A level between 33 % and 66 % of its uncompeted binding was considered intermediate competition, and ≤33 % was considered competing.

D.4.7. Antibody Epitope Mapping

Shotgun mutagenesis epitope mapping of anti– RSV F antibodies was performed using an alanine scanning mutagenesis library for RSV F protein (hRSV-A2; NCBI ref # FJ614814), covering 368 surface-exposed residues identified from crystal structures of both the prefusion and postfusion conformations of RSV F. An RSV F expression construct was mutated to change each residue to an alanine (and alanine residues to serine). The resulting 368 mutant RSV F expression constructs were sequence confirmed and arrayed into a 384-well plate (one mutation per well).

Library screening was performed essentially as described previously (27). The RSV F alanine scan library clones were transfected individually into human HEK-293T cells and allowed to express for 16 h before fixing cells in 4 % (vol/vol) paraformaldehyde (Electron Microscopy Sciences) in PBS plus calcium and magnesium. Cells were incubated with mAbs, diluted in 10 % (vol/vol) normal goat serum (NGS), for 1 h at room temperature, followed by a 30 min incubation with 3.75 μg mL^{-1} Alexa Fluor 488-conjugated secondary antibody (Jackson ImmunoResearch Laboratories) in 10 % NGS. Cells were washed twice with PBS without calcium or magnesium and resuspended in Cellstripper (Cellgro) plus 0.1 % BSA (Sigma-Aldrich). Cellular fluorescence was detected using the Intellicyt high-throughput flow cytometer (Intellicyt). Before library screening, to ensure that the signals were within the linear range of detection, the optimal screening concentrations for each mAb were determined using an independent immunofluorescence titration curve against cells expressing wild-type RSV F.

Antibody reactivity against each mutant protein clone was calculated relative to wild-type protein reactivity by subtracting the signal from mocktransfected controls and normalizing to the signal from wild-type proteintransfected controls. Mutations within clones were identified as critical to the mAb epitope if they did not support reactivity of the test mAb, but supported reactivity of other antibodies. This counter-screen strategy facilitates the exclusion of RSV F protein mutants that are misfolded or have an expression defect. The detailed algorithms used to interpret shotgun mutagenesis data are described elsewhere.[110]

D.4.8. Crystallization and Structure Determination of 14N4-Fab and 14N4-Fab–RSV F

Recombinant 14N4-Fab was concentrated to 10 mg mL^{-1} and a crystal was obtained in Hampton Index HT screen condition 20 % (wt/vol) PEG 3350, 50 mM zinc acetate. The crystal was harvested directly from the screening tray, cryoprotected in the mother liquor with 20 % (vol/vol) glycerol, and data were collected using a Bruker Microstar microfocus rotating-anode X-ray generator equipped with a Bruker Proteum PT135 CCD area detector, and Proteium2 software (Bruker-AXS). Data were processed with XPREP[248] to 2.0 Å. The structure of 14N4-Fab were determined by molecular replacement in Phaser[249] using the separate constant and variable domain models from PDB ID code 4Q9Q. The model was improved through iterative refinements in Phenix[249] and manual building in Coot,[250] guided by composite omit maps.

To crystallize 14N4 in complex with RSV F, both hybridoma-cleaved 14N4 and RSV A2 F were buffer-exchanged in excess into 50 mM Tris pH 7.5, 50 mM NaCl. 14N4-Fab was mixed in excess with RSV A2 F postfusion protein and incubated at 37 °C for 2 h. Following this, the sample was subjected to sizeexclusion chromatography (S200, 16/300; GE Healthcare Life Sciences) in 50 mM Tris pH 7.5, 50 mM NaCl. The complex was concentrated to 10 mg mL^{-1} and crystals were obtained in Hampton Crystal Screen HT in 2 M ammonium sulfate, 5 % (vol/vol) 2-propanol. Approximately 40 crystals were screened for diffraction, and numerous cryoprotectants were tried; however, the best diffraction obtained was to 4.1 Å using the mother liquor with 20 % (vol/vol) glycerol as a cryoprotectant. X-ray diffraction data were collected at the Advanced Photon Source LS-CAT beamline 21-ID-F. Data were indexed and scaled using XDS.[251] A molecular replacement solution was obtained in Phaser[249] using RSV A2 F protein trimer PDB ID code 3RRR and the structure of 14N4-Fv region. Significant density, albeit shifted from the apostructure, was observed for the constant region, and a solution could be obtained in Phaser with the constant region. The structure was refined using group B-factors, coordinates, NCS restraints, and 14N4-Fab and PDB ID code 3RRR as reference models restraints. The density around the 14N4–RSV F interface was well defined and CDR loops matched well with the apo–14N4 structure.

D.4.9. Negative-Stain Electron Microscopy

14N4-Fab was mixed in excess with RSV 18537 B postfusion F protein and incubated at 37 °C for 1 h. Following this, the complex was purified by size-exclusion chromatography (S200, 16/300; GE Healthcare Life Sciences) in 50 mM Tris pH 7.5, 50 mM NaCl. Carboncoated copper grids were overlaid with the complex at 5 µg mL^{-1} for 3 min. The sample was washed in water twice and then stained with 0.75 % uranyl formate for 1 min. Negative-stain micrographs were acquired using an FEI Tecnai F-20 transmission EM scope and a Gatan 4k × 4k CCD camera using 50,000× magnification at a defocus of −1.5 µm. Micrographs were rescaled by a factor of two resulting in a final image with 4.36 Å px^{-1}. Particles were picked manually using EMAN Boxer[252] with a box size of 75 pixels and pixel size of 5.25 nm px^{-1}. Reference-free 2D classification was performed using Spider.[253]

D.4.10. Surface Plasmon Resonance

Binding experiments using surface plasmon resonance were carried out on a ProteON XPR36 instrument (Bio-Rad). For this experiment, we used GLC sensor chips (Bio-Rad). To determine detection of Fab binding, FFL_001 was captured using the anti-his mAb (Immunology Consultants Laboratory, Clone 7B8). Mutated FFL_001 (R33C, N72Y, K82E) was used as a binding control. Fabs were injected as analytes in running buffer HBSEP+ (Teknova) with 1 mg mL^{-1} BSA at a flow rate of 50 µL min^{-1}. The surface was regenerated with 0.85 % phosphoric acid (Bio-Rad), four injections, 15 s contact time each. We analyzed data using Proteon Manager software (Bio-Rad, v3.1.0.6). Binding responses were double referenced against interspot and reference channel. We fit the data with the Simple Binding Langmuir model.

D.4.11. HD Exchange Mass Spectrometry

Deuterium exchange was initiated by addition of 6.6 µL 14N4 Fab (2.0 mg mL^{-1}) and 3.3 µL of either scaffold (1.1 mg mL^{-1}) or water into 40 µL exchange buffer (100 mM NaCl, 20 mM Tris·HCl, pH 7.5) made in D_2O. For a nondeuterated control, the reaction was performed in the same buffer made in water. The reaction was allowed to proceed for 15, 30, or 60 min, and was quenched by addition of 50 µL quenching buffer (0.2 % formic acid, 200 mM TCEP, 4 M urea, pH 2.45). The reaction was placed on ice, and 6.6 µL of porcine gastric pepsin (20 mg mL^{-1}) (Sigma-Aldrich) was added. Protease digestion was allowed to proceed for 5 min on ice, after which 100 µL was used for HPLC separation and mass spectrometric analysis. Each time point was performed in triplicate and the results averaged for analysis. The individual peptides were separated and analyzed for deuterium incorporation using a Rheodyne 7010 manual injector (Sigma-Aldrich) connected to a ThermoFinnigan Surveyor HPLC. Peptides were separated using Phenomenex 50 × 2.1 mm C18 reverse-phase column at 100 µL min^{-1}. Separation was performed using a 5–65 % acetonitrile/H_2O gradient over 25 min, with 0.1 % formic acid added to each buffer. The sample loop and column, as well as the chromatographic buffers, were completely submerged in an ice-water slurry to prevent excessive back exchange of deuterium atoms into the solvent. Mass spectra were recorded using a ThermoFinnigan LTQ XL ion trap mass spectrometer using positive ion electrospray ionization. The mass spectrometer was set to scan in the m/z range of 300–2, 000, with the first 2 min of elution diverted to waste to eliminate early-eluting salts. For deuterium-exchange experiments, data were collected in MS1 mode. For peptide identification the same chromatography gradient was used, with the mass spectrometer run in data-dependent mode collecting seven scan events using collusion-induced dissociation fragmentation with a collision energy of 25 V. Peptide identification was done using PEAKS software (v7.0, Bioinformatics Solutions). Peptides were searched using a parent mass error tolerance of 0.5 Da and a fragment mass error tolerance of 0.5 Da, using nonspecific enzymatic cleavage and a charge state of 1–4. Posttranslation modifications of methionine oxidation and asparagine/glutamine deamidation were considered in peptide identification. Peptides were matched against a database consisting of 14N4 heavy and light chains, as well as porcine pepsin. Only peptides with $a - 10logP$ score of 35.3 or better were selected for deuterium exchange analysis, corresponding to a 0.05 false-discovery rate. Of all peptides identified, 15 with consistent signal and optimal coverage of all CDR loops were selected for deuterium-exchange analysis. The centroid mass of each peptide was calculated for each time point and compared with the nondeuterated control to calculate the extent of deuterium incorporation. The shift in mass compared with nondeuterated control was normalized by the theoretical upper limit of deuteration for each peptide to obtain the percent deuteration. Deuterium incorporation for an individual residue was calculated as a weighted average of all fragments containing the residue, weighted by the inverse of the peptide length. This normalization strategy has been used successfully to convert deuterium exchange values to a per-residue basis for structural visualization.[254]

D.5. Acknowledgements

We thank members of the J.E.C. laboratory, especially Nurgun Kose, Robin Bombardi, Rebecca Lampley, and Dr. Gopal Sapparapu for technical assistance; Dr. Jinhui Dong for helpful discussions regarding structure determination; and Dr. Joel Harp for technical assistance. A portion of the experiments described herein used the Vanderbilt robotic crystallization facility, which was supported by NIH Grant S10 RR026915. J.J.M. was supported by Grant T32 AI 07474, National Institute of Allergy and Infectious Diseases, NIH. The project was supported by the National Center for Advancing Translational Sciences Grant 2 UL1 TR000445-06. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.