Computational design and optimization of protein-protein interactions to engineer novel binders of Influenza Hemagglutinin

Aaron A. Chevalier

A dissertation

Submitted in partial fulfillment of the

Requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

David Baker, Chair

Jesse Bloom

Patrick Stayton

Program Authorized to Offer Degree:

Department of Bioengineering

University of Washington

Abstract

Computational design and optimization of protein-protein interactions to engineer novel binders of Influenza Hemagglutinin

Aaron A. Chevalier

Chair of Supervisory Committee:

Professor David Baker

Biochemistry

Influenza is a serious public health concern and new therapeutics that protect against this highly adaptable virus are urgently needed. For this dissertation my efforts were focused on creating and improving *de novo* designed small proteins that bind to the influenza surface protein Hemagglutinin (HA) and mimic the binding interaction of neutralizing antibodies. These designed proteins were aimed at a highly conserved stem region targeted by some neutralizing antibodies that can inhibit viral membrane fusion. While parts of the stem region are highly conserved within the two main Influenza groups (I and II) differences between the groups make engineering a broad intergroup binder difficult. New high throughput experimental and computational methods were developed which allowed for the testing and design of tens of thousands of new proteins to achieve these goals. Furthermore, newly developed proteins were designed to be small and hyperstable in order to be more ideal therapeutics.

Acknowledgements

Upon first entering the Baker Lab I was very fortunate to have been mentored by Tim Whitehead and Sarel Fleishman. My entire graduate career has been a direct extension of the work they pioneered 5 years ago and it is because of them that I design Influenza binding proteins today.

One of the biggest advantages I have working in a large lab is the tremendous amount of both computational and experimental support received from my colleagues. I have had the luxury of being surrounded by many passionate intelligent people that have guided my development these past 5 years. Their contributions cannot be counted and the work in this thesis would never have come to fruition without their continual support.

For computational tutelage in learning the art of Rosetta design I would specifically like to thank Darwin Alonso, Chris Bahl, Jacob Bale, TJ Brunette, Javier Castellanos, Jorge Fallas, Alex Ford, Neil King, David La, Tom Linsky, Vikram Mulligan, Lucas Nivon, Daniel Silva, Lei Shi, Gabriel Rocklin, Will Sheffler, and Yifan Song.

For experimental work Cassie Bryan, Jeremy Mills, James Moody, Jorgen Nelson, Fabio Parmeggiani, Erik Procko, Justin Siegel, Eva-Maria Strauch, and Shawn Yu have all been of great help.

Our lab is kept running by our amazing technicians Jasmine Gallaher, Inna Goreshnik, Rashmi Ravichandran. The Protein Production Facility and Lauran Carter have made many proteins for me saving me many days of work.

Our structural collaborators in Ian Wilson's Lab have always gone out of their way to keep me stocked with Hemaggluinin protein. Our animal collaborators in Deb Fuller's Lab and Merika Treants have shown what designed protein therapeutics are capable of. Lance Stewart and the Institute of Protein Design continually push me to dream big with my goals.

I would also like to thank the members of my Doctoral Advisory Committee, Jesse Bloom, Pat Stayton, Roland Strong, and Paul Yager for taking time from their busy schedules to provide support and advice about various topics of my research.

David Baker has made this all possible and it has been a great honor to work in his lab. I thank him very much for the opportunity.

Lastly, and most of all, I'd like to thank my family for supporting me through the years.

Introduction

Four years ago two *de novo* computationally designed binders were published that bound to the stem region of influenza Hemagglutitin[1]. At the time they were the first demonstration of computationally redesigned naturally occurring proteins binding a target epitope with atomic level accuracy. However the two binders HB36 and HB80, though extensively optimized experimentally, reached an affinity barrier well above the neutralizing antibodies that motivated their creation. This hurdle prevented us, at the time, from determining if *de novo* protein binders were truly alternatives to antibodies and ultimately a new pathway to the creation of viable therapeutics. We set out then to explore methods to increase their affinity, discover their antiviral therapeutic potential, and to design the next-generation of binders inspired by the first two.

The over-arching methodological theme of this thesis was the application of high-throughput molecular biology and directed evolution techniques into the computational protein-protein interface design (PPID) pipeline. This was first applied to lower the affinity of HB80 and HB36 to antibody like levels. Next-generation sequencing was used in combination with designed protein libraries and directed evolution to determine the fine grained fitness landscapes of the binders. These sequence-function maps were then used to find hard to evolve combinations of mutations that allowed both proteins to bind with dissociation constants below 1nM. Through *in vitro* testing we showed that these high affinity variants were capable of eliciting very similar viral inhibitory effects to neutralizing antibodies. These results were published in 2012 in Nature Biotechnology titled *Optimization of affinity, specificity, and function of designed influenza inhibitors using deep sequencing*[2] and are the entirety of Section 1 and 2 here.

Though not discussed at length in thesis initial attempts at using the enhanced binders as therapeutics *in vivo* were hindered when pharmacokinetic (PK) characterization of the HB36 and HB80 binders showed rapid clearance from blood serum. These results were most likely due to renal clearance from the small size of the proteins and degradation due to serum proteases. The PK data was reinforced when the binders showed no protection against viral challenge *in vivo*. However, we found when administered intra-nasally in both mouse and ferret models the HB36 binder could be quite effective as both a therapeutic and prophylactic against H1 viral challenge. These results are currently being written up and submitted as *Prophylactic and Therapeutic Protection Against Influenza by a Computationally Designed Protein* to Nature Medicine lead by our collaborators in the Fuller Lab. The results suggest that by delivering the binders directly to the lungs they can remain active and at high enough concentrations to be effective. However, the binders were only effective against half the circulating strains of Influenza A, still suffered from stability and solubility issues, and were too large to synthesize chemically. We chose to utilize this information, and by applying some of the high-throughput techniques pioneered in Section I, build a new PPID pipeline to generate *de novo* proteins that did not suffer these shortcomings.

Within the framework of Rosetta the PPID pipeline has always involved an initial phase of modeling and design of proteins *in silico,* experimentally testing a number of those models for binding, and finally optimizing a subset of those working binders through mutation. Rosetta computational design has an inherent advantage of being easily scalable as the process is suitably efficient; simply increasing computational resources can generate tens to thousands to millions of potential designs. This ability to generate design models should continue to expand as high-performance computing systems continue to decrease in cost and increase in accessibility. Thus the bottleneck in the ability to iterate over the PPID pipeline, to quickly generate new highly

functional proteins, are limitations in the number the proteins that can be tested quickly and easily. By utilizing highly parallel oligo synthesis methods and NGS we show in Section III-IV a 100 fold increase in the number of designs that can be tested. We also show that by using *de novo* designed scaffold proteins we can engineer binders that have more ideal therapeutic properties.

The ability to generate new small hyperstable binder designs is dependent on the number of suitable starting scaffold proteins available. The protein data bank (PDB) contains only a small number of proteins that would be suitable scaffolds for protein therapeutics, limiting the number of binder designs that can be constructed. However, by generating scaffolds computationally within the framework of Rosetta, we can escape the limitations of the PDB and generate orders of magnitude more potential designs. We can imbue these scaffolds with the desired physical traits (size/stability/structure) from their inception; such that the final binders are more ideal therapeutics. We show in Section III-IV a new method, which combines *de novo* scaffold generation with high-throughput testing, to create binders against the stem region of Group I and II Influenza hemagglutinin.

Though the methods described here have been applied to create protein binders against the influenza hemagglutinin stem region, they were developed to be generalizable and robust enough to be applied to other protein target surfaces. The combination of high-throughput design and testing should be a powerful tool for future advancements in protein interface design. Also, we believe the new hyperstable proteins developed have great potential to be a new class of protein-based therapeutics and diagnostics.

Table of Contents

Section 1: Optimization of affinity, specificity, and function of designed Influenza inhibitors using next generation sequencing

Abstract

We show that comprehensive sequence-function maps obtained using next generation sequencing can be used to reprogram interaction specificity and to leapfrog over bottlenecks in affinity maturation by combining many individually small contributions not detectable in conventional approaches. We use this approach to optimize two computationally designed H1N1 Influenza Hemagglutinin (HA) inhibitors and, in both cases, obtain variants with sub-nanomolar binding affinity for H1. The most potent of these, a 51-residue protein, is broadly cross-reactive against all group 1 HAs, including human H2 HA, and neutralizes H1N1 viruses with potency that rivals several human monoclonal antibodies, demonstrating that computational design followed by comprehensive energy landscape mapping can generate proteins with potential therapeutic utility.

Background and Motivation

Influenza is a serious public health concern, and new therapeutics are urgently needed that protect against this highly adaptable virus. We recently reported the de novo design of two proteins that, after affinity maturation using error-prone PCR, bind with nanomolar affinity to a conserved epitope on the stem of influenza hemagglutinin that is the target of broadly neutralizing antibodies[1]. One of these designed binders, HB80.3, inhibited the pH-induced conformational change necessary for influenza virus infectivity and so was a promising candidate for generating a broad spectrum antiviral agent against influenza, but further screening

failed to isolate higher affinity variants. We hypothesized that further improvement of activity could require a combination of multiple small contributions from mutations that might individually be difficult to identify. To identify such sequence variants and obtain a complete map of their contributions to binding in these designed proteins, we extended a recently described approach for mapping binding interfaces using high-throughput sequencing to encompass much larger sets of positions[3,4].

Results

We investigated the contributions to binding at all 51 positions in HB80.3 and 53 positions surrounding the experimentally determined binding surface (out of 93 possible) in the designed binder HB36.4 (**Table 2.1**, **Figure 2.1**). To ensure adequate statistics with such a large number of positions and compensate for short sequencing read lengths, which allow coverage of only a subset of the interrogated positions , we utilized libraries in which each member contained a single substitution (a complete set of amino-acid variants were generated at each position, and the individual position libraries were then combined). Using yeast display[5] and fluorescence-activated cell sorting (FACS), populations were collected from each library that bound to either SC1918/H1 (H1) or VN2004/H5 (H5) HA subtypes under sorting conditions of varying stringency (details are in **Figure 2.2** and **Tables 2.2-3**). From each selected population, plasmid DNA was extracted and the mutant genes PCR amplified and then sequenced in two segments using Illumina GA-II 76-bp paired-end deep sequencing.

Analysis of the unselected libraries showed that near complete sequence coverage was achieved: the HB36.4 library contained 1053 of the possible 1061 single amino-acid substitutions, and the HB80.3 library, 1013 of the 1021 possibilities. In each selected population, the ~1000 unique amino-acid sequence variants were sampled with a median depth

of coverage of over 300 per variant and little sequencing error **(Figure 2.1a-c, Figures 2.3-5, Tables 2.2-3)**. The median number of DNA reads per population was 1,534,424, and the minimum 1,049,035. In libraries sorted solely for display on the yeast surface, the variant frequencies were surprisingly similar to those in the unselected population, suggesting that even aberrantly folded proteins make it to the surface despite the yeast secretion quality control system, perhaps due to the small size of the displayed proteins (**Figure 2.6**).

The enrichment ratios (defined as the $\log_2$ ratio of the frequencies of a variant in the selected vs. the unselected population) provide a measure of the effect of each amino-acid substitution on binding. Under ideal conditions (free equilibration of fluorescently labeled HA among the different clones, equal growth rates of all clones, etc.), this measure would be directly proportional to the change in free energy of binding resulting from the substitution. These conditions are not likely to be perfectly met in the experiment, but several lines of evidence suggest that the measure is a reasonable proxy. The enrichment ratios are nearly identical for synonymous mutations (**Figure 2.7**) and correlate with independent affinity measurements on individual variants using yeast surface display titrations **(Table 2.4**), In experiments in which clones with widely ranging in vitro affinities were mixed and then subjected to yeast display selection, the highest affinity clone rapidly took over the population **(Figure 2.8).** Finally, as noted below the enrichment ratio is broadly consistent with the structures of the designed complexes.

Maps of the enrichment ratios for H1 HA binding of each of the ~1000 single amino-acid substitutions in HB36.4 and HB80.3 suggest that most substitutions are neutral or deleterious **(Figure 1.1 a,c)**; the computationally designed interfaces in this respect are similar to naturally occurring interfaces as found in previous large-scale mapping experiments of protein

sequence/function[6–9]. The positions where very little sequence variation is tolerated are either in the core of the protein or directly at the designed interface (**Figure 1.1 b,d)** with the starting designed amino acid being almost always favored (**Figure 1.1 e,f**). In HB36.4, few substitutions were tolerated for the binding hotspots Phe49 and Trp57, and, in HB80.3, the hotspot residues Phe13 and Tyr40 are also strongly conserved. Overall, the enrichment ratios are consistent with the design models of both interfaces and the crystal structure of the HB36.3 interface[10].

Energy Function Improvement

More detailed analysis of the enrichment ratios provides a comprehensive view of the binding energy landscapes of computationally designed interfaces, which differ from naturally evolved interfaces in not being optimized by countless generations of natural selection. These data provide an unprecedented opportunity to identify and remedy the shortcomings in the computational model that underlies the design calculations. We tested the energy function used in the design calculations by attempting to recapitulate computationally the experimental maps using a simple model which accounts for the effects of mutations on the free energy of both folding and binding ($P_{binding}$ = probability_of_folding * probability_of_binding_if_folded; see **Figure 1.2** legend and **Methods**)[11,12]. While the model partially discriminates deleterious substitutions from neutral ones, it does not identify beneficial substitutions (**Figure 1.2a,b**); this result is expected since any substitutions that are favorable according to the design model would have been incorporated in the original design. Many of the newly identified beneficial mutations likely increase electrostatic complementarity at the interface periphery, including substitutions to basic residues in the vicinity of acidic patches on the HA surface (e.g. P66K/R on HB36.4 and G12K/R on HB80.3) (**Figure 1.2 c-d**). Long-range electrostatics were not modeled in the original design calculations because of difficulties in computationally efficient and accurate

modeling of these interactions and hence these beneficial substitutions were missed. To remedy this shortcoming, we incorporated into the energy function used in the calculations a rapidly computable static Poisson-Boltzmann electrostatics model (see **Methods**), which results in improved recapitulation of the beneficial electrostatic substitutions (**Figure 1.2 a,b**) and better overall recapitulation of the experimental results (for quantitative comparison see **Table 2.5**). As described in **Figure 2.9**, the model also improves recapitulation of the free energy changes brought about by mutation in the completely independent Barnase-Barstar complex. Continuum electrostatics calculations have been applied to modeling protein-protein interactions previously[13,14]; our implementation is particularly well suited to calculations on large numbers of mutations because it employs a single full Poisson-Boltzmann solution for the potential of the fixed target in all calculations, which makes computations rapid and reduces noise due to changing boundary conditions. More generally, the large number (~2000) of experimental data points generated by the approach was invaluable for guiding robust improvement of the forcefield; the much smaller datasets generated by conventional methods can be all too readily overfit.

Specificity Switch

Achieving binding specificity among structurally related ligands has proven challenging in protein engineering; this is typically approached by alternating negative selection steps with positive selection, but negative selection can be problematic and the iteration can make the approach laborious[15]. The energetic differences revealed by the experimental maps can be exploited to achieve binding specificity by identifying substitutions that are neutral or enriched in one population and depleted in another. The SC1918/H1 (H1) or VN2004/H5 (H5) HA subtypes differ only by a handful of conservative substitutions at the target surface, making

engineering for specificity quite challenging. Comparative analysis of the HB36.4 H1 and H5 HA medium-stringency binding maps (**Figure 1.3a**) uncovered the single substitution I58E, which is completely depleted in the H5 binding population, but not at all depleted in the H1 binding population (in the bound complex, position 58 binds close to a region in which H1 and H5 differ; see **Figure 2.10**). HB36.4 I58E bound H1 HA, but showed no binding of H5 HA at the maximum concentration tested, where the net change in specificity is over 30-fold (**Figure 1.3b**). Comparison of the energy landscapes mapped by next generation sequencing thus allows reprogramming of interaction specificity, in this case providing a route to the development of subtype-specific influenza binders for clinical diagnosis.

Combining Enriched Substitutions

The enrichment landscapes also provide a route forward to obtain higher affinity variants by combining individually small beneficial effects that may not be detectable by conventional directed evolution selections. To investigate whether the substitutions that were enriched in the selections for HA binding can be combined to produce higher affinity binders and if the contributions of the individual substitutions are additive, we created libraries consisting of 12 variable positions and 4,600,000 unique variants for HB36.4 and 9 variable positions with a total diversity of 300,000 unique variants for HB80.3 by allowing, at each position, the starting residue type and the beneficial substitutions with more than 4-fold enrichment (**Table 2.6**). We carried out Illumina sequencing of the HB80.3 library before and after selection for H1 HA binding, and compared the enrichments of each pair of substitutions at the 9 variable positions to those expected if the mutational effects were purely additive. A strong overall correlation was observed between the experimentally determined enrichment of pairs and the prediction based on the effects of the individual mutations (**Figure 2.11**), but a statistical model that

distinguishes between direct (positions i and j covary) and indirect (positions i and k covary because both covary with j) covariance using a maximum-likelihood approach (see **Methods**) found statistically significant co-variances between several positions (**Figure 2.12**)[16]. Because the effects were not strictly additive, we carried out 4 additional yeast display sorts for increased H1 HA binding affinity and slower off rates (see **Methods**), and determined the sequences of selected clones in the enriched population. The likelihood of these selected sequences using the maximum likelihood model based on the round 1 next generation sequencing data increased when the observed co-variances were included (**Figure 2.13)**; we anticipate that next generation sequencing of more complex libraries followed by model fitting including co-variances will allow creation of more active variants in situations where the size of the library makes exhaustive experimental characterization impossible.

A subset of the enriched HB80.3 and HB36.4 variants (**Tables 2.7-9**) were expressed in *E. coli* with an N-terminal FLAG tag and a C-terminal His tag and purified by affinity chromatography. The binding affinities for HA of six of the variants that were soluble and monomeric were determined by surface Plasmon resonance. The highest affinity of the HB36 variants, F-HB36.5 (F- denotes an N-terminal FLAG tag), differs at 8 positions from the starting sequence and binds SC1918/H1 HA with a binding dissociation constant (K_d) of 890 pM, 28-fold lower than HB36.4, and a reduced off-rate (k_{off}) of 0.0015 s^{-1}. The best of the HB80.3 variants, F-HB80.4, which harbors 5 mutations compared to HB80.3 (**Figure 2.14)** has a K_d of 600 pM, which is 25-fold lower than HB80.3, and a k_{off} of 0.0022 s^{-1}, 10-fold slower than F-HB80.3 (**Figure 1.3c**). Three of the five substitutions in HB80.4 likely improve long-range electrostatics (G12R, A35R, S42R). Incorporation of these three substitutions alone (construct F-HB80.4.1)

yields a K_d of 1.2nM and a k_{off} of 0.0056 s^{-1} (**Figure 2.15),** showing that much, but not all, of the binding improvements are due to the contributions from charge-charge interactions.

Structural Determination

To investigate the molecular determinants of recognition of the improved design variant, we sought to determine the x-ray structure of F-HB80.4 in complex with the SC1918 H1 HA ectodomain. We succeeded in determining the structure at 2.7 Å resolution. After molecular replacement using only the 1918/H1 HA structure as the search model (PDB 3GBN)[17], clear electron density was observed for the inhibitor. F-HB80.4 binds the target HA region in the orientation predicted by the designed model, with the main recognition helix packed in the hydrophobic groove between helix A and the N-terminal segment of HA1 (**Figure 1.4a-b**). The overall backbone conformation of F-HB80.4 agrees well with the electron density maps, but atomic displacement parameters (B-values) are elevated and a few features, such as some side chains, are not apparent for residues that are distant from the F-HB80.4-HA interface, presumably due to conformational plasticity in F-HB80.4 or some heterogeneity in binding (**Figures 2.16-18 and Table 2.13**). However, the main contact helix on F-HB80.4 is well ordered and, after refinement, electron density was apparent for most of the key contact residues on F-HB80.4, including Phe13, Ile17, Ile21, Phe25 and Tyr40. Taken together, the crystal structure of F-HB80.4, as well as that of the previously solved HB36.3, are in excellent agreement with the designed interface, with no significant deviations at any of the contact positions. This agreement between the design model and the crystal structure is quite encouraging given the early stage of de novo protein interface design -- HB80.4 not only interacts with the hydrophobic cleft in HA recognized by HB36[10], but also interacts with the A helix and N-terminal segment of HA1 through the designed hotspot residue Tyr40 that

recapitulates the similar interaction of Tyr98 in CR6261 and Tyr102 in the broadly neutralizing antibody F10[18]

Evaluation of the binding affinity of F-HB80.4 against a panel of group I HAs by biolayer inferometry showed that it is more cross-reactive than the starting HB80.3 and many neutralizing antibodies targeting the same surface on HA, such as CR6261 (**Figure 1.4c**). In addition to binding all of the Group 1 HA's recognized by antibody CR6261 (H1, H2, H5, H6, H9, H13, and H16), F-HB80.4 also binds to H12 HA, which neither CR6261 nor HB80.3 do[9]. Most remarkably, F-HB80.4 binds human H2 HAs with high affinity. H2N2 viruses were responsible for the deaths of ~1 million people during the 1957 pandemic, and these viruses continued to circulate in humans until 1968. The Ile45Phe substitution in the HA2 subunit found in all human H2 viruses strongly reduces the binding of CR6261 and other V_H1-69 related antibodies[19]. Consequently, CR6261 neutralization of H2 is restricted to avian viruses (with Ile45) and only the recently reported F16v3 antibody has been reported to neutralize all virus subtypes, including human H2 viruses[20]. Despite targeting the same surface recognized by neutralizing antibodies, the high affinity interaction of F-HB80.4 with human H2 underscores a potential advantage of de novo designed binders, as they are likely to bind the target differently than an antibody (e.g., using a helix rather than loops) and can, in some cases, circumvent barriers that pose problems for antibodies, such as that for V_H1-69 antibodies binding H2 viruses. Given their proven capacity for sustained replication and transmission in humans and the lack of widespread immunity to H2N2 viruses in the general population (i.e., people born after 1968 have never been exposed to H2 viruses and immunity among older individuals infected more than 40 years ago has likely declined), the reservoir of H2N2 viruses in birds are a possible source for a future pandemic. Consequently, antivirals with more potent and cross-reactive

activity against the H2 subtype, such as F-HB80.4, would be key components of a comprehensive therapy for influenza.

Binding and Neutralization

Given its high affinity, heterosubtypic binding, and inhibitory activity in biochemical assays (**Figure 2.19**)[10], we tested the neutralization potential of F-HB80.4 against the recent A/California/04/2009 H1N1 virus, which was responsible for the 2009 H1N1 pandemic and is currently established as the predominant circulating strain, as well as the seasonal human flu virus A/H1N1/Hawaii/31/2007. F-HB80.4 showed 50% effective concentrations (EC_{50}s) of 170nM (1.6 mg/mL) and 98nM (0.9 mg/mL) against 25 $TCID_{50}$ (50% tissue culture infective dose) of these viruses (**Figure 1.4d**). These levels of neutralization activity are comparable to those of neutralizing antibodies, which have a 50% inhibitory concentration (IC_{50}) range from 0.1-100ug/mL IgG (for example, the IC_{50} for CR6261 IgG against H1 HA is 9ug/mL (~120nM))[19]. While the therapeutic potential of small binding proteins remains to be proven in humans, F-HB80.4 either alone, as a fusion with an antibody Fc, or as a high avidity oligomer is a promising lead candidate for a next generation of anti viral therapeutics.

Discussion

Deep sequencing of populations undergoing non-purifying selection has been used to experimentally determine fitness landscapes for a heat shock protein[21] and an RNA enzyme[22], and to map interactions for protein-DNA[23,24], protein-peptide[3], and HIV-1 antibody-antigen complexes[25]. These approaches probed sequence changes within a single segment less than the ~80bp that can be covered in an Illumina sequencing run. Our approach using single-site mutagenesis libraries and multiple-segment Illumina sequencing has the advantage of being able to interrogate large stretches of sequence and still allow enrichment ratios to be associated with

specific substitutions. Furthermore, our use of single site mutagenesis libraries allowed complete probing of an extended region (150bp) with relative small starting libraries, which resulted in extensive sampling and robust statistics for the vast majority of the substitutions investigated; as in previous approaches normalization to the starting pools corrected for any initial library bias (from either codon usage or synthesis). Beyond these technical advances, because we applied the method to computationally designed and, hence, non-optimized proteins, our landscapes differ from those observed in previous studies, as there are clear positions where substitutions provide significant enrichment over the initial starting sequence.

Both the HB36.4 and HB80.3 results show that landscapes mapped by deep sequencing can be used to rapidly obtain large increases in binding affinity after conventional directed evolution by PCR mutagenesis has plateaued by combining large numbers of individually small favorable effects. The specific combination of mutations contained within these variants would be very difficult to find by conventional affinity maturation approaches. For example, identification of the F-HB80.4 variant with 5 amino-acid mutations (8 DNA sequence changes) using unbiased libraries would have required screening all 5 amino-acid mutant combinations – a diversity of 7.5E+12 --while the total diversity of the landscape guided library was 10^7-fold lower. The traditional approach of carrying out multiple rounds of selection and then using conventional sequencing to identify the few best clones would also not have arrived at the high affinity variants; only one of the substitutions found in the highest affinity variant was among the most heavily enriched in the population and, therefore, combining the few top mutations found after conventional selection and sequencing would not have lead to the best combined variant. The results also illustrate how the landscapes can be exploited to reprogram interaction specificity for closely related targets (H1 and H5 HA) by examining not just beneficial

mutations, but also neutral and deleterious ones. Finally, our results show how the landscapes generated by next generation sequencing can provide a comprehensive view of the shortcomings in computational protein design and should guide and stimulate the development of more accurate forcefields and more powerful design methods. More generally, integration of next generation sequencing with computational protein design provides a powerful route to inhibitors or binders for, in principle, any surface patch on any desired target of interest. Given a newly arising pathogen, for example, following structure determination and identification of sites of interaction with the host, hot-spot based protein interface design can be used to generate diverse small proteins predicted to block the host interaction surface. With modern oligonucleotide assembly methods, genes for large numbers of designs can be rapidly built and displayed on yeast, where the functional designs can be readily identified by flow cytometry. Complete single-site saturation mutagenesis libraries can then be generated for functional designs and subjected to next generation sequencing before and after one round of selection for increased binding activity. The enriched substitutions can be combined in a final library, and optimized high affinity variants selected from this pool. We anticipate that this combined approach will be widely useful in generating high affinity and specificity binders to a broad range of targets for use in therapeutics, diagnostics and targeting.

Methods

Library Creation

Single-site saturation mutagenesis (SSM) libraries for HB36.4 and HB80.3 were constructed from synthetic DNA by Genscript. Parental DNA sequences are listed in **Table 2.1** with mutagenic region highlighted in red. Yeast EBY100 cells were transformed with library DNA and linearized pETCON[5] using an established protocol[26], yielding 1.4e6 and 3.3e6

transformants for the HB36.4 & HB80.3 SSM libraries, respectively. After transformation, cells were grown overnight in SDCAA media in 30 mL cultures at 30°C, passaged once, and stored in 20 mM HEPES 150 mM NaCl pH 7.5, 20% (w/v) glycerol in 1e7 aliquots at -80°C.

Yeast display selections and titrations

Cell aliquots were thawed on ice, centrifuged at 13,000 rpm for 30 s, resuspended in 1e7 cells per mL of SDCAA media, and grown at 30°C for 6 h. Cells were then centrifuged for 13,000 rpm and resuspended at 1e7 cells per mL SGCAA media and induced at 22°C between 16-24 h. Cells were labeled with either biotinylated Viet/2004/H5 HA or SC1918/H1 HA, washed, secondary labeled with SAPE (Invitrogen) and anti-cmyc FITC (Miltenyi Biotech), and sorted by fluorescent gates as outlined in **Tables 2.2-3** and **Figure 2.2**. Biotinylated HA was produced as previously described[10]. Cells were recovered overnight at 2.5e5 collected cells per mL SDCAA media, whereupon at least 1e7 cells were spun down at 13,000 rpm for 1 min and stored as cell pellets at -80°C before library prep for deep sequencing. Plasmid DNA for individual clones was produced according to the method of Kunkel[27] and yeast display titration was done as previously reported[10,26].

Library prep and sequencing

Between 1-4e7 yeast cells were resuspended in Solution I (Zymo Research yeast plasmid miniprep II kit) with 25 U zymolase and incubated at 37°C for 4 hrs. Cells were then freeze/thawed using a dry ice/ethanol bath and a 42°C incubator. Afterwards, plasmid was recovered using a Zymo Research yeast plasmid miniprep II kit (Zymo Research, Irvine, CA) into a final volume of 30 mL 10 mM Tris-HCl pH 8.0. Contaminant genomic DNA was processed (per 20 mL rxn) using 2 mL ExoI exonuclease (NEB), 1 mL lambda exonuclease (NEB), and 2 mL lambda buffer at 30°C for 90 min followed by heat inactivation of the enzymes

at 80°C for 20 min. Plasmid DNA was separated from the reaction mixture using a Qiagen PCR cleanup kit (Qiagen). Next, 18 cycles of PCR (98°C 10 s, 68°C 30s, 72°C 10 s) using Phusion high fidelity polymerase (NEB, Waltham, MA) was used to amplify the template and add the Illumina adaptor sections. Primers used were population-specific and are listed in **Table 2.10**. The PCR reaction was purified using an Agencourt AMPure XP kit (Agencourt, Danvers, MA) according to the manufacturer's specifications. Samples were quantified using Qubit dsDNA HS kit (Invitrogen) for a final yield of 1-4 ng/uL. Samples were combined in an equimolar ratio; from this pool, 0.32 fmol of total DNA was loaded on 2 separate lanes and sequenced using a Genome Analyzer IIx (Illumina) with appropriate sequencing primers (**Table 2.10**).

Sequencing analysis

Alignment and quality filtering of the sequencing data from raw Illumina reads were treated essentially as described[3]. Sequencing reads were assigned to the correct pool on the basis of a unique 8 bp barcode identifier (**Table 2.10**). All pools were treated identically in sequence analysis and quality filtration. Custom scripts were used to align all paired-end reads with both reads above an average Phred quality score equal or above 20. Paired-end reads were aligned using a global Needleman-Wunsch algorithm, reads without gaps were merged into a single sequence and differences between sequences resolved using the higher quality score for the read.

To investigate amino-acid sequence co-variance, two-body analysis was performed whereby the enrichment ratio for pairs of mutations was compared to the predicted enrichment ratio based on the individual component mutations. The individual enrichment ratios were calculated as the overall normalized probability of finding the mutation in the selected pool, the predicted enrichment for a pair of mutations was the sum of the component mutations enrichment ratios, and the actual enrichment ratio was calculated as the overall normalized

probability of finding that pair of mutations in selected pool. A more rigorous analysis was performed to rank each mutational variant found in the deep sequenced library using a statistical model based on the method of Balakrishnan[14]. In brief, the method constructs a maximum entropy statistical model of the following functional form:

$$P(s) \propto \exp(\sum_i f_i(s_i) + \sum_{(i,j)\in E} f_{ij}(s_i, s_j))$$

where s is a particular 9-mer from the sort1 set, s_i and s_j are the amino acids at the ith/jth positions of this sequence, E is the set of interacting pairs of positions identified by the model and f_i, f_{ij} are model parameters which can be thought of as 1 and 2 body (negative) statistical energies respectively. Thus, each f_i can be thought of as a vector that stores the statistical energies for the possible amino acids at that position while f_{ij} is, analogously, a matrix that stores the statistical energies for the amino acid pairs at positions i and j. These parameters are learned from the data using a maximum likelihood procedure based on LASSO[28]. A baseline model that does not capture sequence co-variation (i.e a model with all f_{ij}'s set to zero) was also learnt from the data. Note that, as expected, the probability of an entire sequence can then be written as the product of probabilities of the amino-acid compositions at each position; i.e, each position of the 9mer is treated independently under the baseline model.

Affinity maturation and specificity

Beneficial mutations predicted to result in higher affinity for SC1918/H1 HA were combined into single libraries for both HB80.3 and HB36.4. The DNA library for each design was constructed from assembly PCR using an Ultramer oligonucleotide (Integrated DNA Technologies, CA) to encode the variable region. Primers and sequences are listed in **Table 2.10**, while the DNA sequence for the libraries is listed in **Table 2.6**. The total library size was

3e5 for HB80.3 and 4e6 for HB36.4 and was transformed into yeast[29], yielding 8e6 and 1.5e7 transformants, respectively. These libraries went through 5 sorts of yeast display selection with increasing stringency against HA1-2 as specified in **Table 2.11**. Promising constructs were subcloned into a custom pET-29-based plasmid (NdeI/XhoI) with an N-terminal FLAG tag and a C-terminal His_6 tag and transformed into *E. coli* Rosetta (DE3) chemically competent cells for expression.

Solubility screening

HB80.3 clones selected from the affinity maturation library were screened by solubility in an *E. coli* expression system using a dot-blot assay. Cells were grown from colonies in deep well plates overnight, and diluted 25-fold into deep well plates at 37°C for 3 h, followed by IPTG induction (1 mM) for 4 h at 37°C. Following induction, cells were separated from spent media by centrifugation at 3,000 x g for 15 min at 4°C and stored as pellets overnight at -20°C. The next morning, plates were thawed on ice for at least 15 min and 200 mL binding buffer (200 mM HEPES, 150 mM NaCl, pH 7.5) was added to each well. The plate was sonicated using the Ultrasonic Processor 96-well sonicator for 3 min at 70% pulsing power and lysate centrifuged for 4000 rpm for 30 min at 4°C. Supernatant at 100-fold dilution was transferred to a dot blot manifold Minifold I (Whatman) and dried onto nitrocellulose membrane for 5 min. The membrane was then labeled with an anti-FLAG HRP conjugated mouse antibody (Sigma, St. Louis, MO) and visualized with DAB substrate (Pierce).

Protein production and purification

Protein expression was induced using the autoinduction method of Studier[30]. Cells were harvested by centrifugation, resuspended into buffer HBS (20 mM Hepes, 150 mM NaCl pH 7.4), and sonicated to release cell lysate. Following clarification by centrifugation, supernatant

was applied to a Talon resin column for purification. Proteins were eluted by step elution at 400 mM imidazole in HBS. Size exclusion chromatography on a Superdex75 column was used as a finishing purification step for HB80.3 variants. Proteins were stored at 4^{o}C for short-term analysis or flash frozen in liquid nitrogen.

Binding analysis

All surface plasmon resonance data were recorded on a Biacore model T100 (Biacore, Uppsula, Sweden). A Biotin CAPture chip (Biacore) was coated with 500 response units (RU) of biotinylated SC1918/H1 HA1-2 ectodomain. All proteins were in buffer HBS-EP with 3 mM EDTA and 0.005%(v/v) P20 surfactant. 238 mL of designed protein was applied at a flowrate of 100 mL/min for 2 min and a dissociation time of 300s with full chip regeneration between each trace. At least five varying concentrations of protein were used to determine kinetic and equilibrium fits. Binding kinetics were determined using a 1:1 Langmuir binding model with Biacore T100 evaluation software and double background-subtracted values.

Biolayer inferometry using an Octet Red (ForteBio, Menlo Park, CA) was used to determine subtype-specific binding for HB80.4 and CR6261. Biotinylated HAs, purified as described[1], were used for these measurements. Briefly HAs at ~10-50 μg/mL in 1x kinetics buffer (1x PBS, pH 7.4, 0.01% BSA, and 0.002% Tween 20) were loaded onto streptavidin coated biosensors and incubated with varying concentrations of HB80.4 in solution. All binding data were collected at 30°C. The experiments comprised 5 steps: 1. Baseline acquisition (60 s); 2. HA loading onto sensor (300 s); 3. Second baseline acquisition (180 s); 4. Association of HB80.4 for the measurement of k_{on} (180 s); and 5. Dissociation of HB80.4 for the measurement of k_{off} (180 s). Five concentrations of HB80.4 were used, with the highest concentration varying, depending on the HA affinity from 50 to 200nM. Baseline and dissociation steps were carried

out in buffer only. Binding kinetics were determined using a 1:1 Langmuir binding model in kinetics data analysis mode using the Fortebio data processing software. The sequences of all biotinylated HAs used in this work are available in Fasta format in **Table 2.12**.

Protease susceptibility assays

Protease susceptibility assays has been done as described[1]. For A/South Carolina/1/1918 (H1N1) HA, each reaction contained ~2.5 μg HA or ~2.5 μg HA and a 5-fold molar excess of F-HB80.4. Significant inhibition was detected with a high ratio of binder to HA, presumably due to the stringency of our assay (1 hour at 37°C at low pH). Little protection was observed when the reaction contained approximated 1 binder per HA protomer.

Computational methods

The Rosetta all atom energy function and design methodology was used to calculate the predicted effect of every possible point mutation in the designed proteins on the free energies of folding and binding. Starting from models of the HB36.4 and HB80.3 complexes which came for the experimentally determined structures for HB36.3 and F-HB80.4[1], each position was singly mutated to all 20 amino-acid identities and for each mutation the structure was optimized by combinatorial repacking of side chains and gradient-based steepest-descent minimization of degrees of freedom on side chain of both sides of the complex and backbone of the designed protein. The complex binding affinity and the unbound stability of the designed monomer were both analyzed using an all-atom energy function dominated by van-der-Waals, hydrogen bonding, and solvation[31]. In binding affinity calculations, the monomers were repacked in the unbound state but backbone degrees of freedom were kept fixed. For monomer stability calculations, a Coulombic model using distance dependent dielectric constant (ε=r) is added to account for intra-molecular electrostatic interactions. The PARSE charges[32] are used for all

residues. The ΔΔG of protein stability and binding energy upon mutation is calculated with both standard van-der-Waals parameters and a reduced repulsive term[33]. Earlier benchmarks showed that this is an efficient approach to identify mutations that introduce van-der-Waals clashes but can be tolerated given more structural flexibility. If ΔΔG decreases by over 5 R.e.u, an additional step of structure optimization is added with standard van-der-Waals parameters, allowing freedom on the rigid body movement between the proteins and side chain and the backbone of both sides of the complex. This additional optimization step leads to more small to large mutations favored in the calculations, decreasing the number of false negatives, but increasing the number of false positives for predicting the favored mutations[12]. This is a desirable behavior for the protocol, as it leads to more favorable mutations that can be tested. This procedure was implemented using the Rosetta macromolecular software package. To model long range electrostatics efficiently and with minimal noise, we calculated the electrostatic potential in the vicinity of the designed proteins due to HA on a grid by solving the PB equation with charges on the atoms in HA, but with all atoms in the designed proteins neutral. The Poisson-Boltzmann equation was solved using APBS[28] with PARSE charges and radii[34] for HA atoms, but no charges for HB atoms and the electrostatic potential generated by HA was calculated on a grid with 0.5Å. The protein is modeled in the low dielectric constant of 4. The solvent is modeled implicitly with high dielectric constant of 80 and salt concentration of 0.15M. The PARSE charges are assigned to HA[32] and the HB design variant is neutral. The PARSE radii are assigned to both HA and HB. The dielectric boundary is defined by the solvent exclusion surface using a probe with a radius of 1.4 Å[35]. The electrostatic interaction energy caused by each point mutation was computed using $E = S*q_i*f$ where f is the electrostatic potential from the grid and q_i are the charges of the atoms on the introduced residues. The energy term is converted to the

Rosetta score function term by 1 kT = 1 Rosetta energy unit (R.e.u.). Detailed RosettaScripts[11] for all computational analyses are available in Supplementary Material. Source code is freely available to academic users through the Rosetta Commons agreement (http://www.rosettacommons.org/).

Isolation of F-HB80.4-SC1918/H1 HA complex for crystallization

Following Ni-NTA purification, SC1918 HA was digested with trypsin (New England Biolabs, 5mU trypsin per mg HA, 16 hours at 17°C) to produce uniformly cleaved (HA1/HA2), and to remove the trimerization domain and His-tag. After quenching the digests with 2mM PMSF, the digested material was purified by anion exchange chromatography (10mM Tris, pH 8.0, .05-1M NaCl) and size exclusion chromatography (10mM Tris, pH 8.0, 150mM NaCl), essentially as previously described for other HAs[10].

To prepare the F-HB80.4/SC1918 complex for crystallization, 1.5 molar excess of F-HB80.4 was mixed with purified SC1918 HA in 10mM Tris pH 8.0, 150mM NaCl at ~2mg/mL. The mixtures were incubated overnight at 4°C to allow complex formation. Saturated complexes were then purified from unbound F-HB80.4 by gel filtration.

Crystallization and structure determination of F-HB80.4-SC1918/H1 HA complex

Gel filtration fractions containing the F-HB80.4/SC1918 complex were concentrated to ~10mg/mL in 10mM Tris, pH 8.0 and 50mM NaCl. Initial crystallization trials were set up using the automated Rigaku Crystalmation robotic system at the Joint Center for Structural Genomics (www.jcsg.org). Several hits were obtained, with the most promising candidates grown in ~15% PEG3350 near pH 7. Optimization of these conditions resulted in diffraction quality crystals. The crystals used for data collection were grown by the sitting drop vapor diffusion method with a reservoir solution (100mL) containing 16% PEG3350, and 100mM Tris pH 7.5. Drops consisting

of 100nL protein + 100nL precipitant were set up at 4°C, and crystals appeared after 3 days. The resulting crystals were cryoprotected by soaking in well solution supplemented with increasing concentrations of ethylene glycol (5% steps, 5min/step), to a final concentration of 25%, then flash cooled and stored in liquid nitrogen until data collection.

Diffraction data for the F-HB80.4-SC1918 /H1 complex were collected at the Advanced Photon Source (APS) General Medicine/Cancer Institutes-Collaborative Access Team (GM/CA-CAT) beamline 23ID-D at the Argonne National Laboratory. The data were indexed in $P2_12_12_1$, integrated using HKL2000 (HKL Research), and scaled using Xprep (Bruker). The structure was solved by molecular replacement to 2.5 Å resolution using Phaser[36]. An unpublished, in house, high-resolution structure of the 1918 HA was used as the initial search model. Examination of the maps at this stage revealed clear positive electron density around the membrane distal end of HA consistent with the expected location and orientation of F-HB80.4. As for HB36.3[10] attempts to place F-HB80.4 by molecular replacement using Phaser were unsuccessful. However, phasing using the HA only yielded maps with continuous density for F-HB80.4, including key side-chain features. This phasing model allowed F-HB80.4 to be fitted into the maps manually and unambiguously. Rigid-body refinement, torsion-angle simulated annealing, and restrained refinement (including TLS refinement, with one group for HA1, one for HA2, and one for F-HB80.4) was carried out in Phenix[37]. Between rounds of refinement, the model was rebuilt and adjusted using Coot[38]. Although we report the structure to a final resolution of 2.7 Å, the crystals diffracted anisotropically to 2.4 Å (along a), 2.5 Å (along b), 2.8 Å (along c) as determined by the diffraction anisotropy server[39]. Data that were truncated and scaled by this server were used for model building. The electron density maps from these 2.7 Å data were of better quality and slightly more easy to interpret than those at a higher resolution of 2.5Å. Data collection statistics

are reported for data with the ellipsoidal truncation applied prior to merging of reflections. The final round of refinement was carried out with data that were ellipsoidally truncated, but with no negative isotropic B-value applied to the data. For the inhibitor F-HB80.4, residues distant from the F-HB80.4-HA interface lacking side-chain electron density were modeled as alanine. The HA head region is well ordered with lower B-values, which increase towards the stem and the inhibitor where there are fewer to no crystal lattice contacts. Final refinement statistics can be found in **Table 2.1**3.

Structural analyses

Hydrogen bonds and van der Waals contacts between F-HB80.4 and SC1918/H1 HA were calculated using HBPLUS[40] and CONTACSYM[41], respectively. MacPyMol (DeLano Scientific)[42] was used to render structure figures and for general manipulations. The final coordinates were validated using the JCSG quality control server (v2.7), which includes MolProbity[43].

Neutralization assay viruses

A/California/04/2009 (pdmH1N1) and A/Hawaii/31/2007 (H1N1) were propagated in Madin-Darby canine Kidney (MDCK) cells (American Type Culture Collection, Manassas, VA) to produce working viral stocks.

Cell culture

MDCK cells were grown in minimum essential medium (MEM) with Earle's Balanced Salts supplemented with 5% fetal bovine serum (Hyclone Laboratories, Logan, UT). Virus amplification for virus stock production was carried out in MEM containing gentamicin (50

μg/ml), porcine trypsin (10 units/ml) and EDTA (1μg/ml)[44]. The antiviral testing was performed in MEM supplemented only with gentamicin (50 μg/ml).

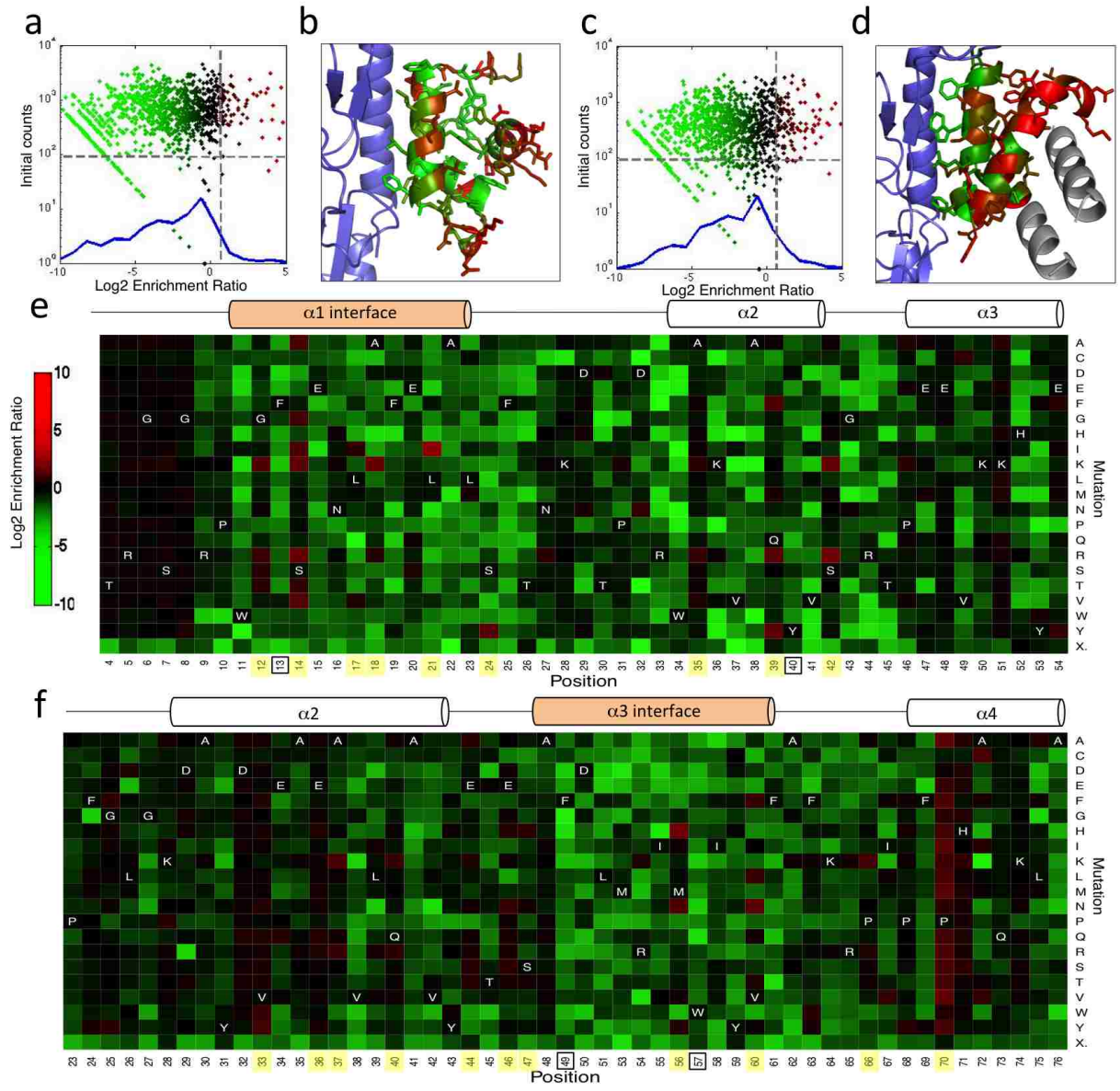
Viral inhibition assays

To calculate the F-HB80.4 concentration-response curve, the peptides were half log diluted in MEM from 10 μM to 0.00032 μM and incubated with 25 $TCID_{50}$ of virus at 37°C with 5 % CO_2 for 1 hour. After incubation, the reaction mixture of each concentration was added to three wells of MDCK cells ($8x10^4$ cells/well) prepared in 96 well plates. Cell controls (uninfected and untreated cells), virus controls (infected and untreated cells), and F-HB80.4 toxicity controls (infected and untreated cells) were included in each test plate. The test was read at day 6 post-inoculation when virus control wells showed 100 % cytopathic effect (CPE). The CPE was evaluated via cell viability through the cellular intake of neutral red (NR) (Thermo Fisher Scientific Inc., Pittsburg, PA)[45]. The NR was used at 0.011% diluted in MEM, the cells were incubated at 37°C with 5 % CO_2 for 2 hours and the plates were read spectrophotometrically.

The EC_{50} for the peptides were obtained by the standardization of the NR results for each of the peptide concentration repetitions against the cell controls (100 % viability) and virus controls (100 % cell death). A plot of the obtained data as percentage of cell viability and percentage of CPE reduction against the peptide concentration was constructed using Excel, 2007. The curve points were also fitted using Excel, 2007[46].

Figures

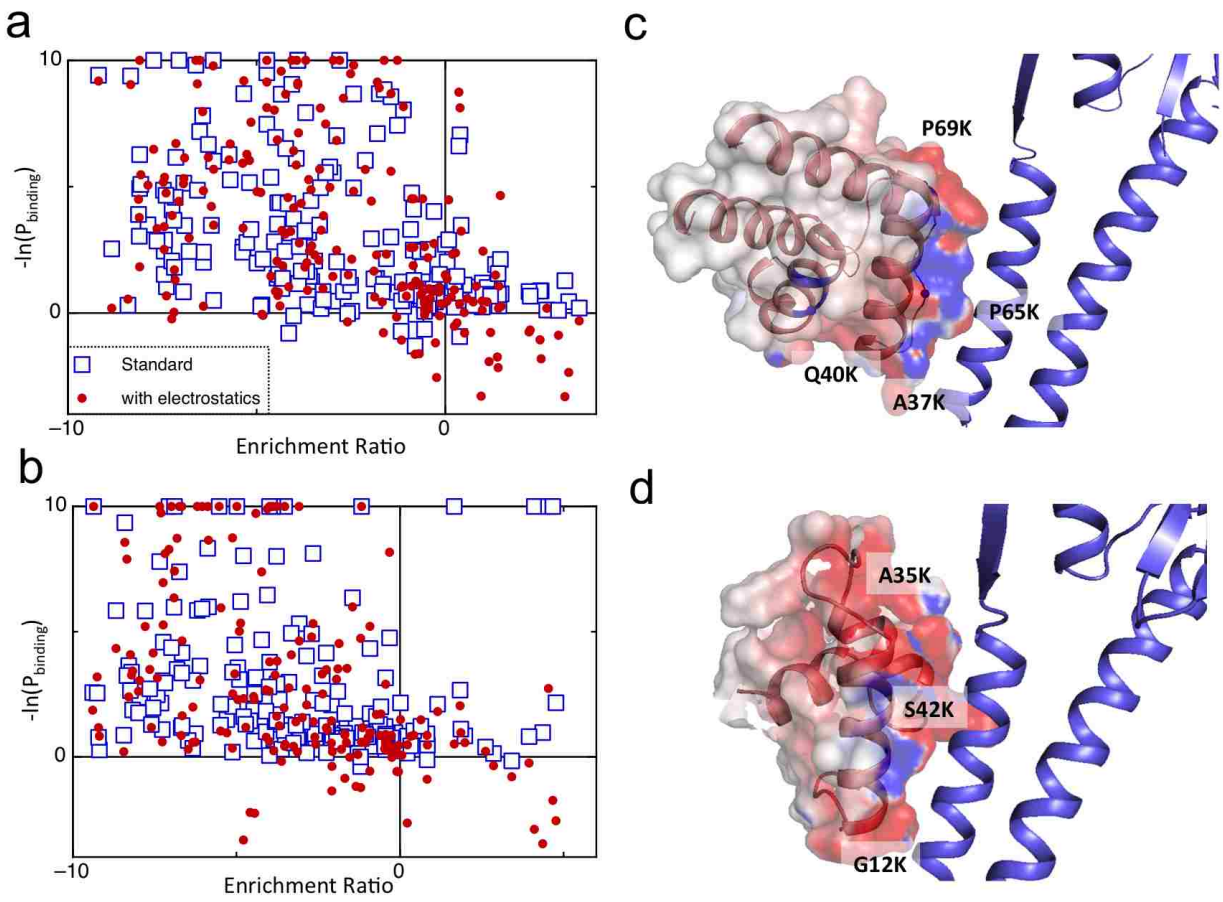
Figure 1.1



Sequence-function landscapes of designed Influenza binding proteins. a, c, Next generation sequencing yields large numbers of independent observations to robustly determine enrichment ratios in stringent binding selections to the H1 HA subtype. Mutations that are heavily depleted are shown in green, while beneficial mutations are indicated in red. Horizontal

dashed line indicates 100 sequence counts for unique non-synonymous substitutions in the library, whereas vertical dashed line demarcates the enrichment ratio of the starting sequence, showing that most substitutions are neutral to deleterious. (a, HB80.3 library; c, HB36.4 library). **b**, d, Model of H1 HA (shown in purple ribbons) bound to HB80.3 (b) and HB36.4 (d). The designed binding proteins are colored by positional Shannon entropy with green indicating positions of low entropy and red those of high entropy. Gray ribbons on HB36.4 indicate positions without deep sequencing data. e, f, Wiring diagrams and heat maps corresponding to H1 HA-binding enrichment ratios under stringent binding selection for all possible single mutations in all 51 positions for **e,** HB80.3 and 53/93 positions for **f,** HB36.4. Starting positions are shown in white font, and the central helix paratope for the design variants are colored in orange in the wiring diagrams. Positions with enrichment greater than 4-fold are colored yellow and were included in the subsequent designed library and black boxes around positions indicate hotspot residues in the original designs.

Figure 1.2



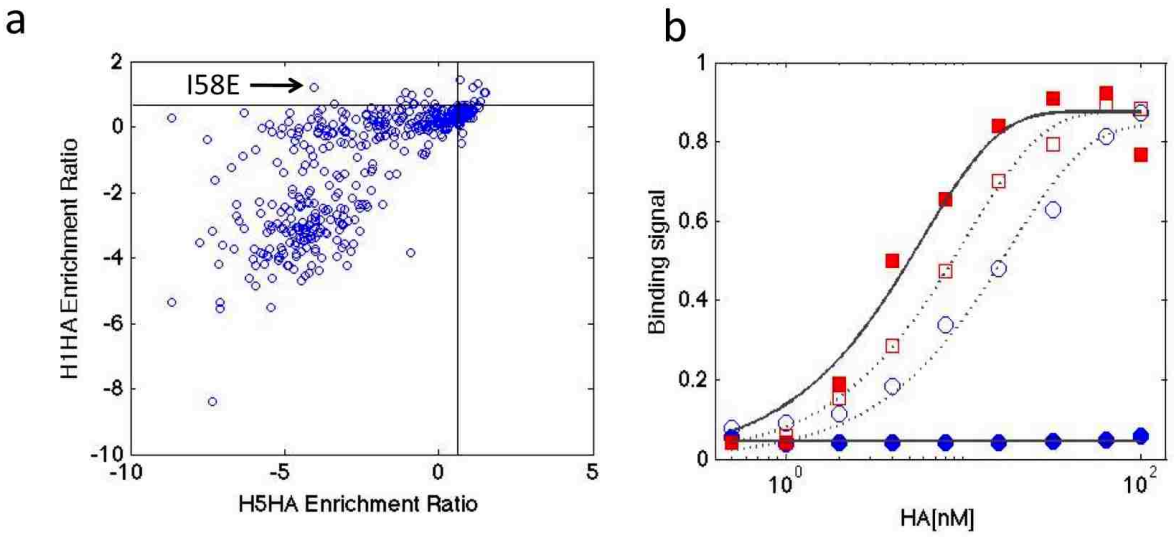
Improvement of computational model by incorporation of long-range electrostatics. The effect of each mutation on binding was computed taking into account both direct effects on the binding interaction and indirect effects on protein stability using

$$-\ln(P_{binding}) = -\ln\left[\frac{e^{\frac{-(\Delta\Delta G_{folding} + \Delta G_o)}{kT}} +}{1 + e^{\frac{-(\Delta\Delta G_{folding} + \Delta G_o)}{kT}}}\right] + \Delta\Delta G_{binding}$$

where $DDG_{folding}$ is the computed change in stability[33], $DDG_{binding}$ is the computed change in binding free energy, and ΔG_0 is the free energy of folding, taken to be 1.0 in the units used here.

The first term accounts for the reduction in the population of the folded state brought about by mutation, the second term, the direct effect of the mutation on the binding interaction. Taking Pbinding to be the Boltzmann weight of the folded bound state in equilibrium with the unfolded and folded but not bound states yielded similar results (data not shown). a, b, Correlation between $P_{binding}$ and the enrichment ratio improves when the Rosetta energy function is supplemented with a long range electrostatics model (see Methods). **a**, HB36.4 and **b**, HB80.3; open blue squares - all-atom Rosetta energy function without the electrostatics term; red closed circles - energy function supplemented with electrostatic interactions computed using the fixed electrostatic field of the target HA (see **Methods**). To highlight the effect of the electrostatic term, only mutations to charged residues (Arg, Lys, Asp, and Glu) are shown. Mutations to neutral residues show a similar correlation; however, there is little difference with and without the electrostatic term. c, d: Electrostatic potential from H1 HA (blue ribbons) mapped onto model of design variant **c**, HB36.4 substitutions A37K, Q40K, P65K and P69K improve electrostatic interactions with HA **d**, HB80.3 substitutions G12K, A35K, and S43K improve electrostatic interactions with HA.

Figure 1.3



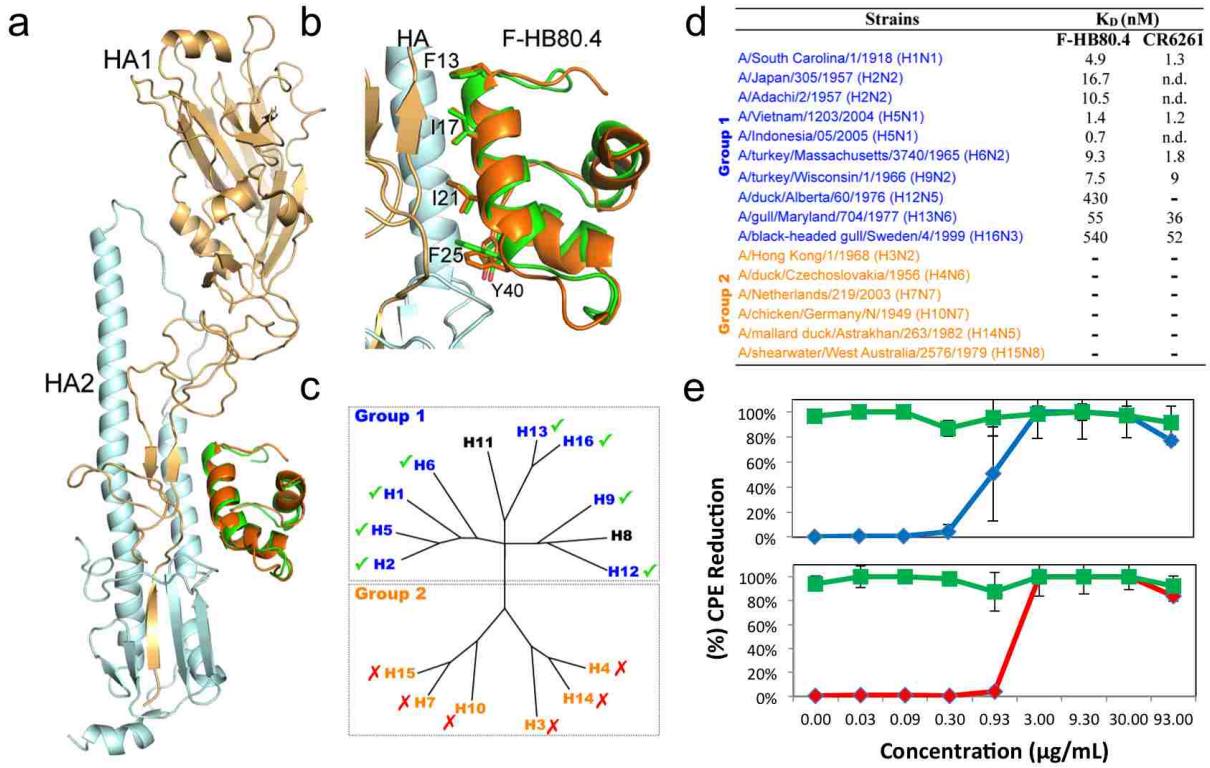
c

Construct	Mutations	K_d [nM]	k_{on} [$M^{-1}s^{-1}$]	k_{off} [s^{-1}]	log_2 enrich
F-HB80.3	--	15.2	1.7e6	.026	1.31
F-HB80.4	G12K, L17I, L21I, A35K, S42K	0.6	3.9e6	.0022	1.586
F-HB80.4.1	G12R, A35R,S42R	1.2	4.4e6	.0056	3.57
HB36.4	--	24.8	2.4e6	.0603	0.1
F-HB36.5	V33Y,E36A,A37E,Q40E,M56H,V60N,P66Q,P70A	0.9	1.6e6	.0015	1.1

Exploitation of sequence-function landscapes leads to subtype-specific HA binders and improved affinity variants. **a,** Scatter plot showing the positional HB36.4 enrichment values (in log_2 units) for medium stringency binding against H1 HA and H5 HA (for details of selections, see **Table 2.2**). Overall agreement between the two datasets is good as expected for epitopes that only differ by a few mutations. The mutation I58E is neutral in the medium stringency H1 population, but depleted in the medium stringency H5 population. The vertical and horizontal lines indicate enrichment for the starting sequence. The balance between selection stringency and information content is an important component of the overall experimental design. **b,** Yeast surface display titrations of HB36.4 (squares) and HB36.4 I58E (circles) against the H1 HA

subtype (dashed line/open symbols) or the H5 HA subtype (solid line/closed symbols) shows HB36.4 I58E selectively binds the H1 subtype. **c,** Binding affinity and kinetics of selected design variants as determined by surface plasmon resonance and corresponding enrichment data from next generation sequenced selections. Both selections and in vitro binding measurements were performed against SC1918/H1 HA.

Figure 1.4

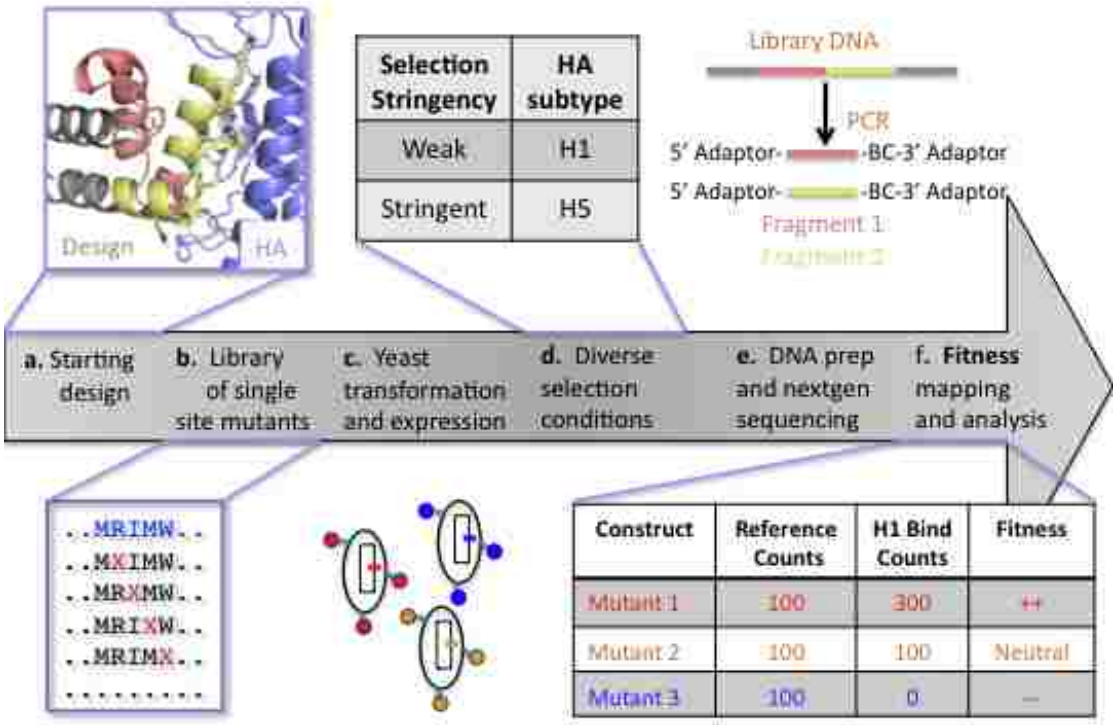


Structure and functional analysis of improved design variant F-HB80.4. **a**, Superposition of the crystal structure of F-HB80.4-SC1918/HA complex and the design model. The F-HB80.4 is represented in orange, SC1918 HA1 subunit in gold, HA2 subunit in cyan and the computational design in green. Superposition was performed using the HA2 subunits. For clarity,

only the HA from the crystal structure is depicted here (the HA used for superposition of the design, which is essentially identical to the crystal structure, was omitted) **b,** Close-up view of the F-HB80.4-SC1918/HA interface with the key HA-contacting residues labeled. The main contact helix on F-HB80.4 is well ordered and after refinement, electron density was apparent for most of the key contact residues on F-HB80.4, including Phe13, Ile17, Ile21, Phe25 and Tyr40. A total of 1460 $Å^2$ is buried at the interface with HA, similar to the surface area buried by CR6261. The coloring is the same and F-HB80.4 is oriented as in (**a**). **c**, Phylogenetic tree showing the relationships between the 16 HA subtypes that can be divided into two groups and a summary of F-HB80.4 binding. Green indicates positive binding by F-HB80.4 and red X no binding. Subtypes that have not been tested for binging are indicated in black. **d**, Affinity measurements (K_d) for binding of F-HB80.4 and CR6261 to representative members of most of the HAs subtypes. n.d. indicates binding was not determined for this experiment, and '–' indicates no binding to the specific HA subtype. **e**, Plot of cytopathic effect (CPE) reduction vs. F-HB80.4 concentration for seasonal flu virus A/H1N1/Hawaii/31/2007 (blue diamonds, top panel) and pandemic A/California/04/2009(H1N1) virus (red diamonds, bottom panel). Green squares are controls for cell viability at each F-HB80.4 concentration tested. Error bars represent a 95% confidence interval in the measurement. The calculated EC_{50} of F-HB80.4 for A/H1N1/Hawaii/31/2007 and pandemic A/California/04/2009(H1N1) viruses is 98 nM (0.9 μg/mL) and 170 nM (1.6 μg/mL), respectively.

Section 2: Supplementary Information for optimization of affinity, specificity, and function of designed Influenza inhibitors using next generation sequencing

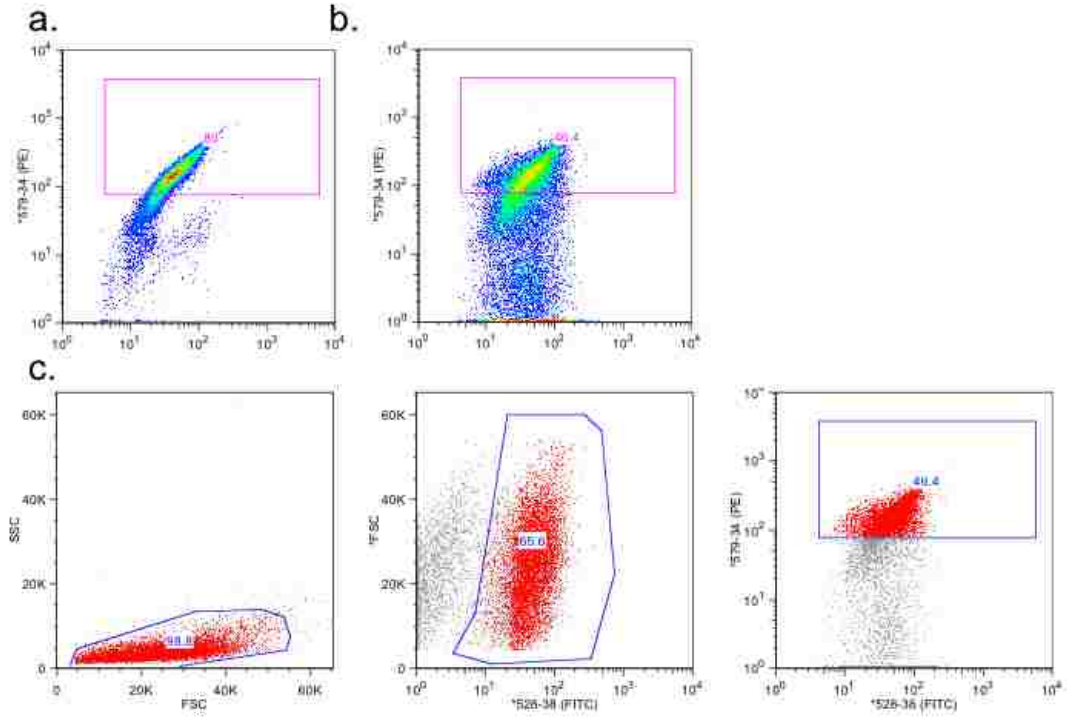
Figure 2.1



Overview of deep mutational scanning methodology as applied to profiling the fitness landscape of designed binders to the conserved stem region of H1 subtype hemagglutinin (H1 HA). **a,b,** Starting from a design variant with initial binding activity, a DNA library encoding single site saturation mutants (SSM) was created for nearly every position in the protein sequence. **c**, The library was transformed into yeast and protein induced to display on the cell surface. **d**, Diverse selections were applied including stringent or weak binding to H1 HA as well as towards differing HA subtypes using fluorescence activated cell sorting (FACS). These

selections are detailed in **Figure 2.2** and **Tables 2.2-3.** Surviving yeast harboring specific mutated designed proteins were collected and propagated. **e,** Plasmid DNA containing the mutant library was harvested and prepped for 76bp paired end Illumina DNA sequencing. Individual barcode (BC) sequences were added to partition sequencing reads to its appropriate selection condition within the same Illumina lane. The final libraries were sequenced in two segments read on separate Ianes which allowed the total variable region ~150bp interrogated to be extended beyond the 76bp limit. This fragmentation was possible due to the high fidelity in construction and use of an SSM library, where only a single mutation was likely to be present across the variable region. Indeed, only 7% of 76-bp reads passing the quality filters had more than one mutation. Reads with multiple mutations were discarded from the analysis. **f**, The fitness of each individual mutant in the library was determined by enumerating its frequency in the population and comparing it to a reference control. These are tabulated in the main text as enrichment ratios in $\log_2$ units.

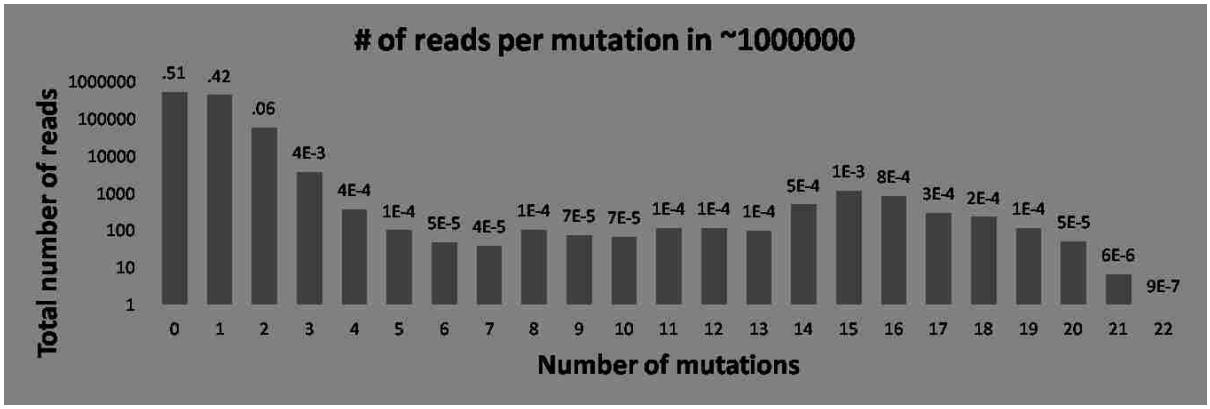
Figure 2.2



Principles of yeast cell surface selections using FACS. Comparison of **a,** a clonal population of the HB80.3 starting sequence or **b,** the HB80.3 SSM naïve library at a labeling concentration of 10 nM SC1918/H1 HA. The y-axis is the phycoerythrin (PE) fluorescent channel associated with binding HA, and the x-axis is the FITC fluorescent channel associated with surface display of the c-myc epitope tag (this tag is displayed on the c-terminus for all constructs). For clarification, only the c-myc displaying portions of the populations are shown in the scatter plots. **c,** Sample backtrace of the FACS gates used to sort the SSM populations. FACS gates are shown in blue lines, and the selected population colored in red. Samples were processed according to appropriate size using a gate on forward vs. side static light scattering (**left panel**), display of full-length protein (forward scatter vs. FITC fluorescent channel) using the c-myc epitope tag as a proxy (**middle panel**), and PE fluorescent channel associated with binding subtype-specific HA vs. FITC fluorescent channel (**right panel**). The stringency of the

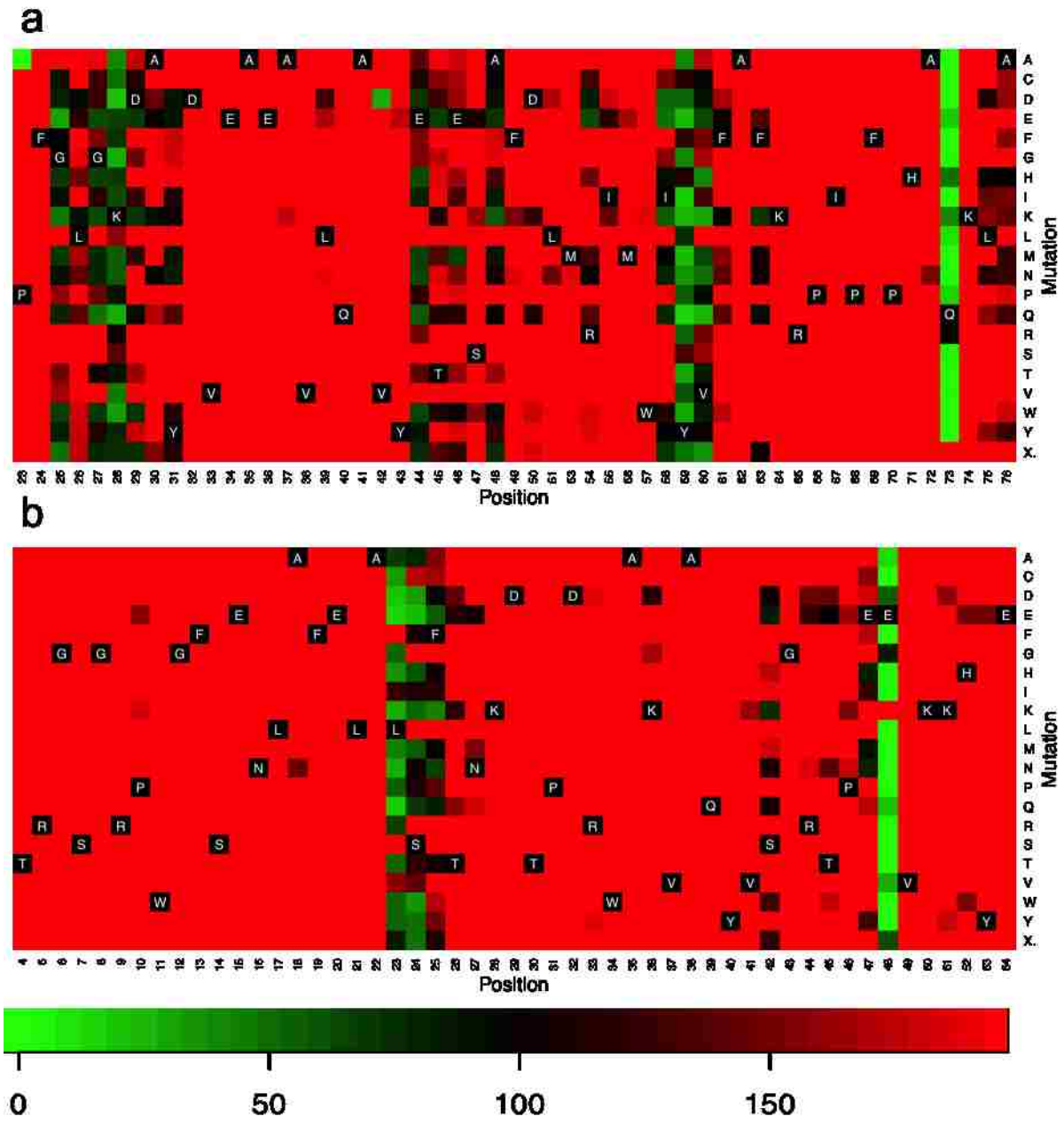
sort is a function of target labeling concentration and the height of the PE gate. All clones will express relatively the same PE fluorescence (normalized for surface display) at a labeling concentration high above the K_d of the starting sequence. Similarly, a gate height that allows most cells to be collected will not allow separation between variants of differing HA affinities. Thus for each experiment, we set labeling concentrations and PE gate heights to enable non-purifying selection of differing affinities. We did this by determining the apparent K_d of the starting variant against the specific HA subtype and setting the labeling conditions and PE gate stringency accordingly in **Tables 2.2-3**.

Figure 2.3



A histogram of the number of parsed reads (in log scale) vs. the number of mutations per sequence. The sample set is one million total DNA sequencing reads that have passed the Illumina quality filters. The data labels indicate the fractional proportion of those reads in the total population. These data represent one of the two fragments from the HB36.4 naïve library where 93% of the population is made up of either 1 or 0 mutations. As this fragment represents only half of the SSM variable region, the majority of the reads with no mutations represent sequences where the mutation (s) is found on the second fragment.

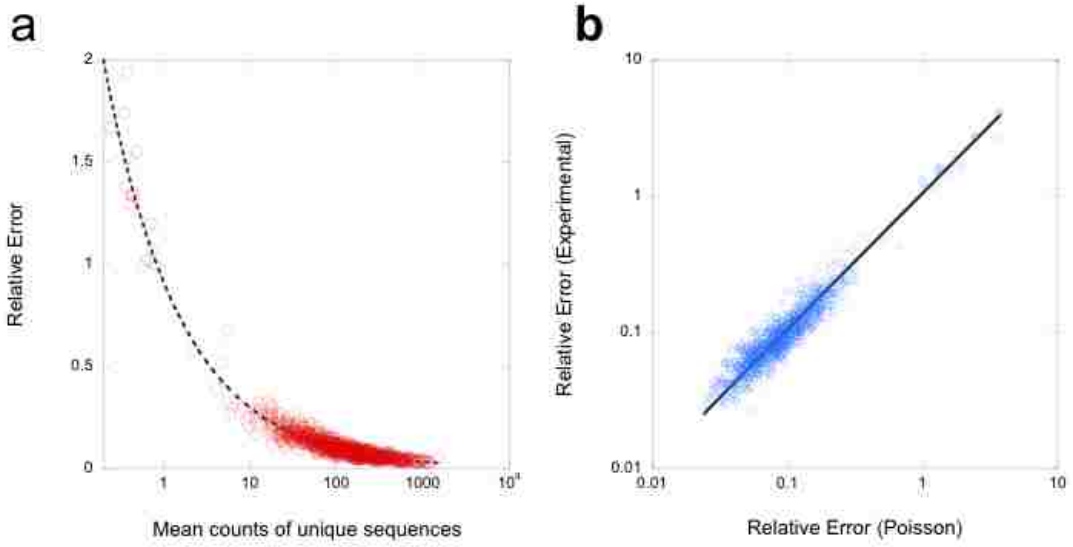
Figure 2.4



Heatmaps for **a**, HB36.4 and **b**, HB80.3 single-site saturation mutagenesis libraries showing positional non-synonymous mutation counts in the initial population (e.g. before any FACS was done). There is a marked heterogeneity to coverage per position in the input libraries,

highlighting the importance of ratiometric analysis in deep sequencing fitness landscape evaluations. Rows are the mutations listed in single letter amino-acid code, where X represents a stop codon.

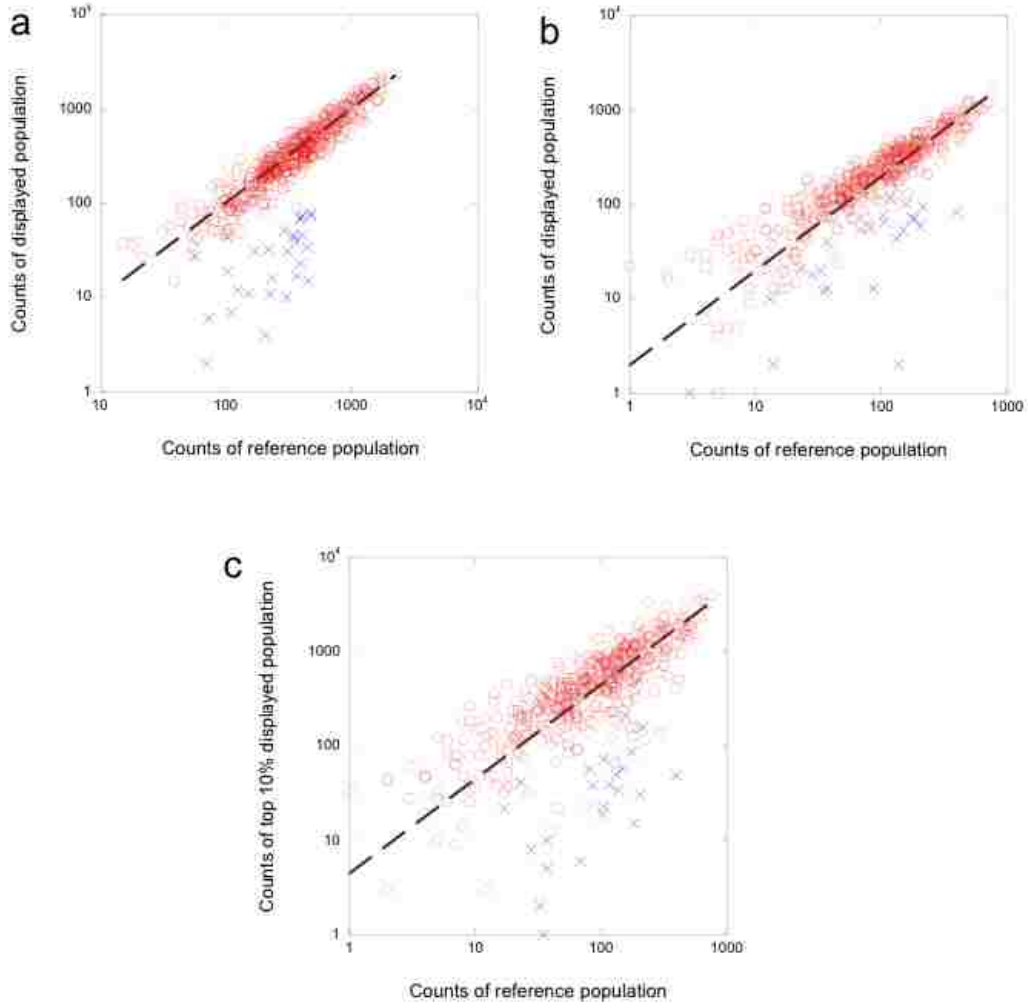
Figure 2.5



To determine sequencing error, 14 samples of the naïve HB36.4 single-site saturation mutagenesis (SSM) library and 14 samples of the naïve HB80.3 SSM library were processed separately with different barcodes and deep sequenced for a median of 220,894 reads (range 106,855-268,455). Variants in the population were enumerated, normalized to 200,000 reads, and compared across independent processing conditions. **a**, Plot of relative error vs. mean counts of unique sequences shows larger error at lower sequencing counts. Relative error is defined as standard deviation divided by the mean. The dashed line is a fit to guide the eye. **b**, Plot of relative error from sequencing vs. relative error expected from a Poisson model shows that the sequencing prep results in errors approaching the theoretical minimum. In a Poisson

model, the variance equals the mean. Thus, the relative error would decrease as the inverse square root of the mean. The solid line is a fit assuming that the relative errors are identical.

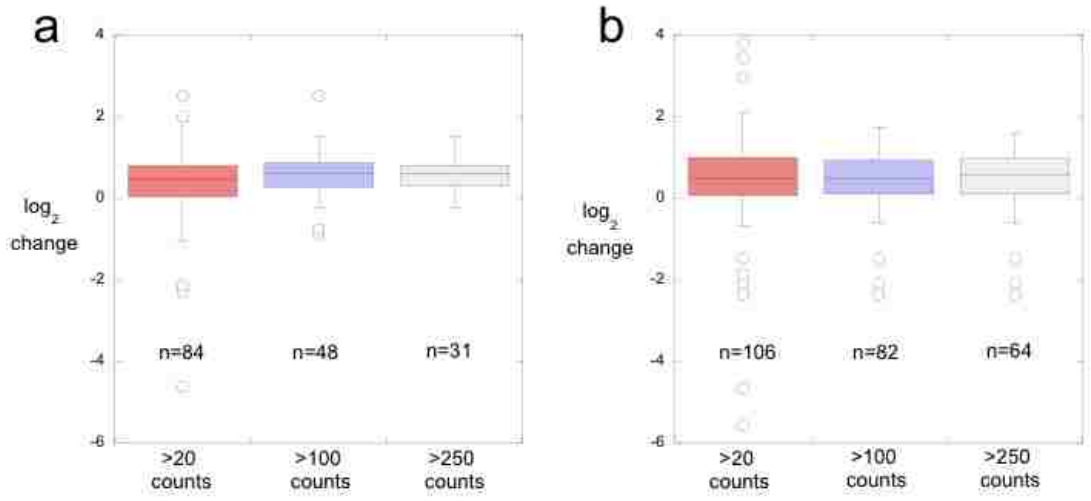
Figure 2.6



Population counts corresponding to selection for surface display of c-myc epitope tag vs. counts of unique variants in a passaged reference population for **a,** HB36.4 variants (positions 23-51, 578/580 possible non-synonymous mutations), and **b,** HB80.3 variants (positions 4-29

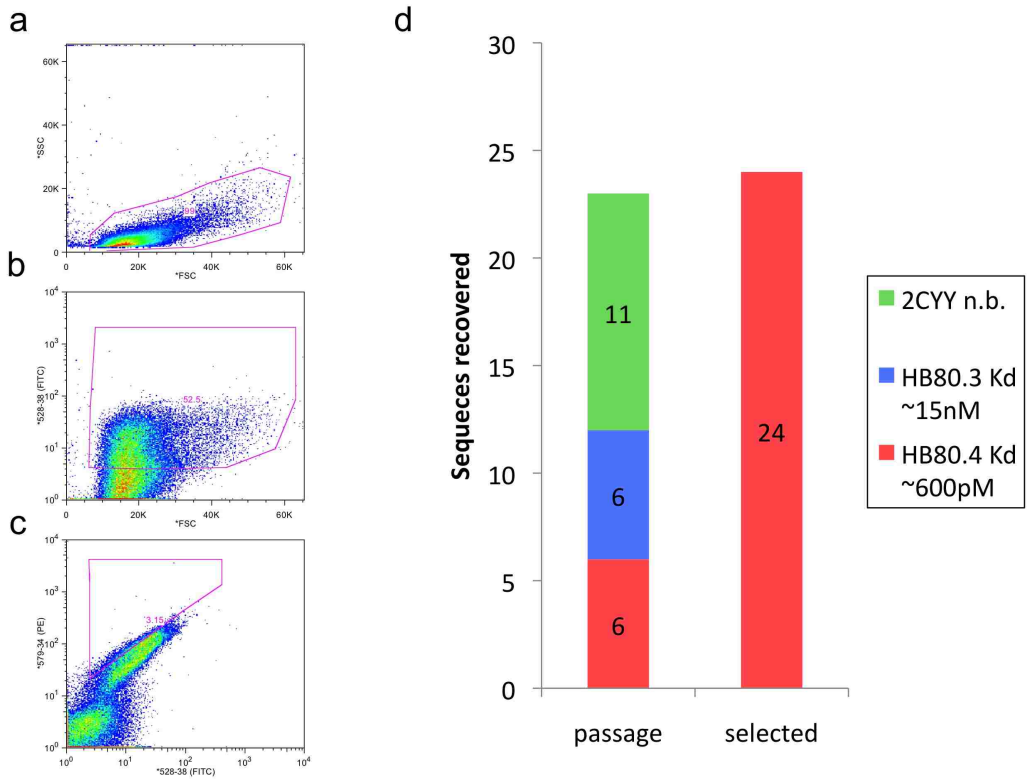
covering 492/520 possible non-synonymous mutations). Dashed black line is a fit to guide the eye through the non-synonymous mutations that result in display of the full-length construct (red open circles). Premature stop codons (blue crosses) are almost the only mutations significantly depleted in the population selected for display. **c,** A more stringent FACS gate for c-myc surface display for the HB80.3 variants (top 10% of displaying population) identifies several potential destabilizing mutations, most notably at Glu15, which is responsible for an intramolecular salt bridge with Arg44. However, as before the majority of the depleted substitutions were stop codons.

Figure 2.7



Box plots of stringent selections of binding to H1 HA relative to a reference population for **a,** HB36.4 and **b,** HB80.3 show agreement among synonymous mutations from the wild-type sequence. Number of synonymous sequences above the specified count threshold in the reference population is listed below the box plots.

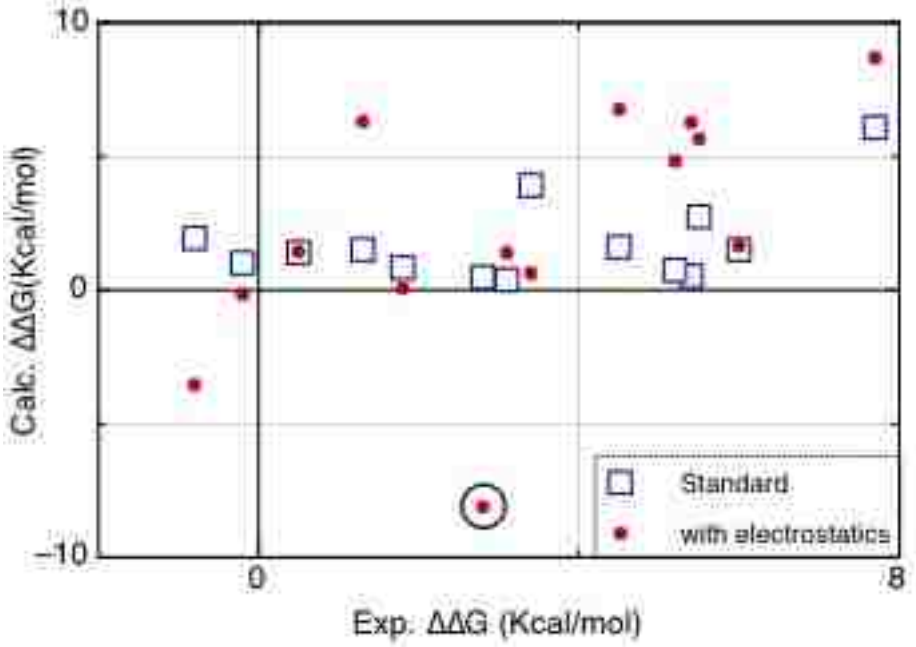
Figure 2.8



Results of selection to demonstrate yeast display enrichment correlates with *in vitro* binding affinities. The selection was performed on an equally mixed population made up of HB80.3, the optimized HB80.4 variant, and the starting scaffold (PDB ID: 2CYY) with a labeling concentration of 2nM SC1918/H1 HA. **a,** Sample backtrace of the FACS gates used to sort the mixed population. The upper left FACS gate shows the population selected for correctly sized cells **b,** The mixed population selected for cells displaying protein. **c,** The final FACS gate used to select the best binding cells. **d,** The results of Sanger sequencing a sample from the either the passage group (corresponding to only the gate in panel **a**) or the selected group (corresponding to the the combination of gates **a,b,** and **c**). HB80.4, the variant with the lowest

K_d, is enriched over both the HB80.3 and the 2CYY scaffold protein and is the only variant found in the sampled selected group.

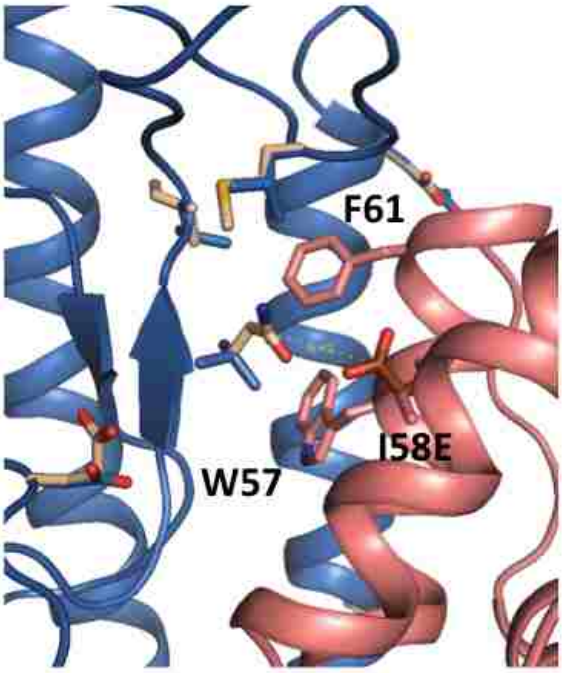
Figure 2.9



Scatter plot showing the calculated vs experimental ΔΔG[47,48] (binding) for various mutations on the highly electrostatic protein-protein complex Barnase-Barstar (PDB ID 1BRS[49]). The calculated ΔΔG's were determined using standard Rosetta force field and the modified force field with the additional electrostatic term as discussed in the **Methods**. The energy term is converted to the Rosetta score function term by 1 kT = 1 Rosetta energy unit (R.e.u.). With the additional electrostatic term, the agreement between the calculated and experimental ΔΔG is improved (correlation coefficient from 0.17 to 0.44; p-value for the correlations improve from 0.122 to 0.010). With one outlier excluded, the correlation coefficient between the calculated and the measured ΔΔG is R=0.66 (p-value=0.0004). The outlier (circled in black in the figure) originates from Glu73 which has been suggested not to interact favorably directly with Barstar;

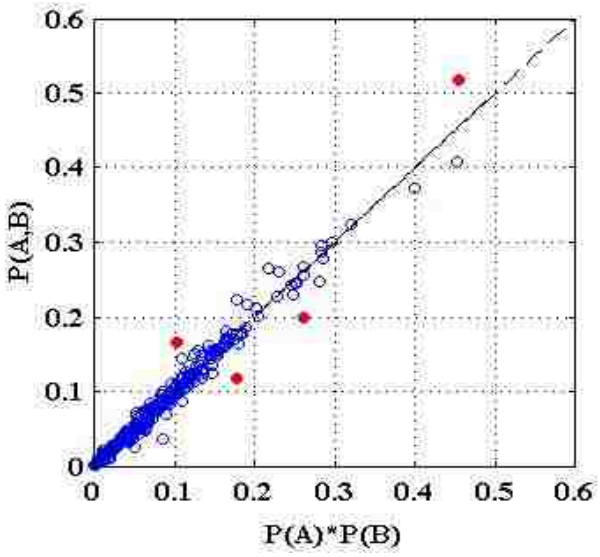
increased binding affinity with Glu73 is likely due to it pre-organizes positive charges on Barnase to interact with Asp 39 on Barstar[50].

Figure 2.10



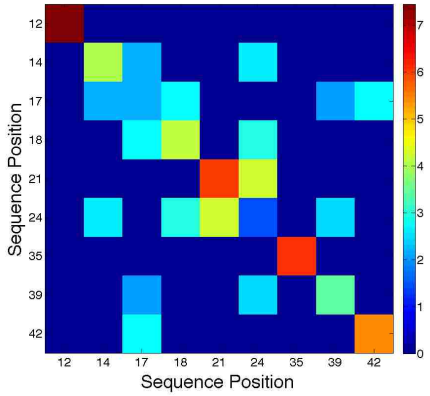
Model of H1 HA-specific HB36.4 variant bound to H1 HA based on the solved structure of HB36.3 bound to SC1918/H1 HA (PDB ID 3R2X)[1]. SC1918/H1 HA (blue cartoon) differs from the VN2004/H5 subtype in only a few conservative substitutions (H5 positions are shown as white sticks) at the main binding epitope for the HB36.4 design (pink cartoon). A single Ile58Glu mutation (Glu58 is shown as brown sticks) is sufficient to completely abrogate binding to the H5 HA subtype while maintaining binding to the H1 HA subtype. Trp57 and Phe61 are previously identified hot spots for HB36.4, and are shown as sticks. Reduced binding to H5 HA is likely due to the desolvation of the side-chain carboxylate on Glu58 by the Gln present on the VN2004/H5 HA.

Figure 2.11



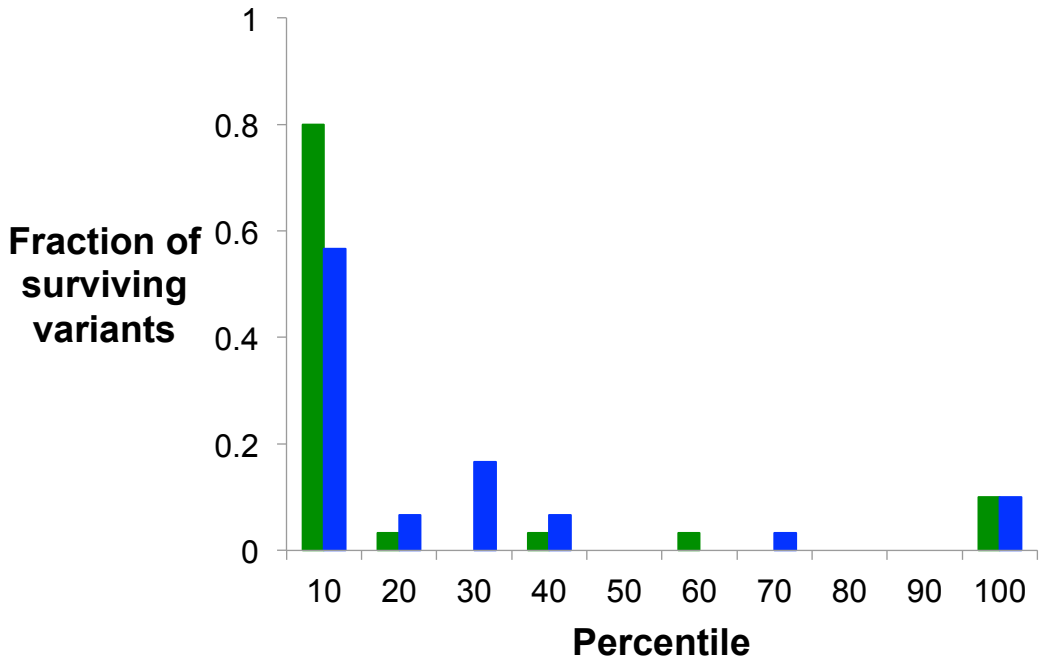
After one sort, enrichment probabilities of paired mutations in the second HB80.3 library (y-axis) are almost exactly predicted by a simple additive model of their individual likelihood (x-axis). A linear relationship implies little covariance between individual amino-acid substitutions. The dashed line is a guide for the eye. Red circles indicate the outlying pairs at positions 21/24 (four in total). Positions 21 and 24 attack neighboring regions of the HA epitope.

Figure 2.12



A graphical representation of the parameters of the model learnt from the sort 1 HB80.3 training set. The diagonal elements represent the extent of conservation at the corresponding position while the off-diagonal elements measure strength of sequence co-variation between the corresponding positions. The color represents the strength of the corresponding parameters in the model (norm of f_i vector for diagonal elements and of the f_{ij} matrix for off-diagonal elements – see **Methods**) and is in arbitrary statistical units (higher=>stronger conservation/co-variation). The strongest patterns are along the diagonal, in the one-body or position specific energies as expected. However, this analysis shows that there is some co-variation between positions with the strongest such co-variation being between positions 21 and 24.

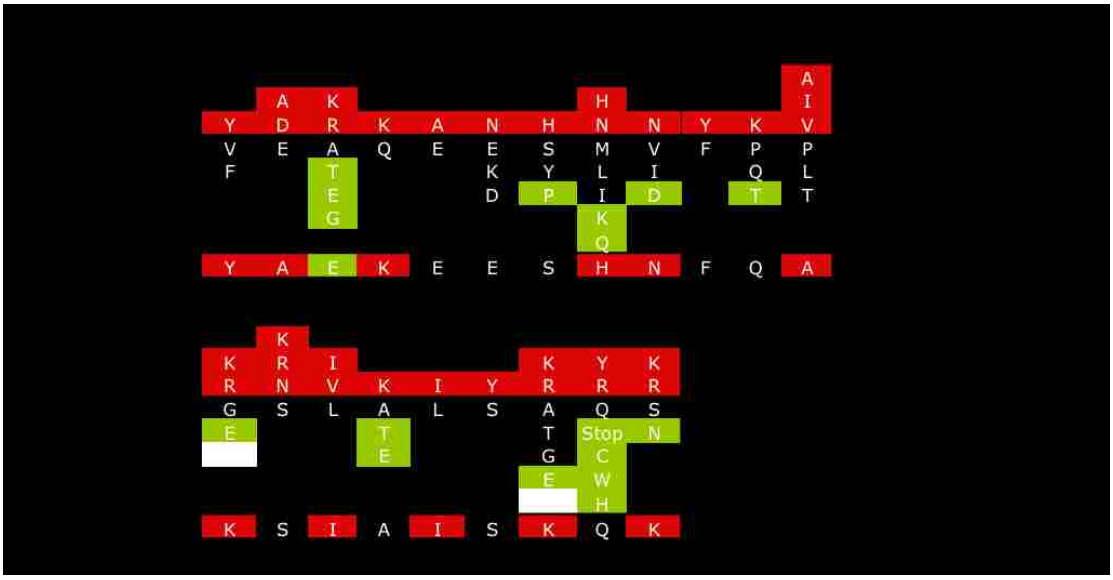
Figure 2.13



The second HB80.3 library was subjected to five yeast display sorts of increasing stringency. Individual clones were sequenced after the fourth and fifth sort (30 in total). This histogram shows the positions of the 30 sequenced surviving HB80.3 variants (after 4 or 5 sorts)

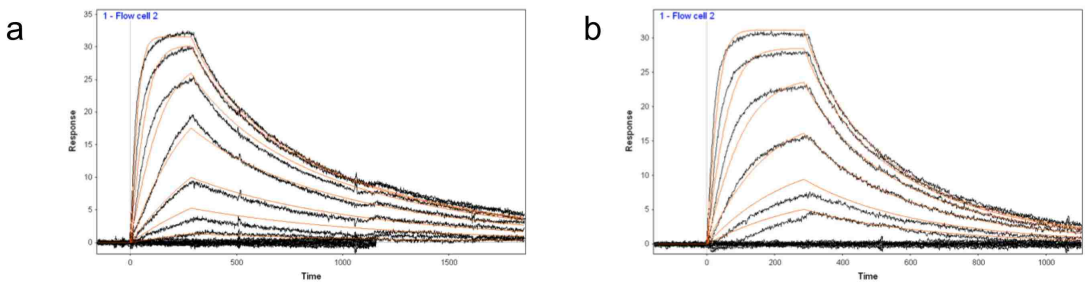
using 53,152 sequenced sort 1 variants as training data. To determine if a model utilizing amino-acid co-variance was a superior predictor of the likelihood of a sequence surviving the sorts, we constructed two statistical models of the sequences in the sort 1 pool. The first model captured only the positional sequence conservation patterns present in estimating the likelihood; the second model used the approach listed above to determine the likelihood of a sequence using the co-variation patterns identified in addition to the conservation patterns. The models were used to rank the likelihood of each of the 53,152 unique variants sequenced in sort 1. The model utilizing co-variation was a superior predictor of the variants surviving into sorts 4 and 5, placing 83% of them in the top 10% of the overall ranked sequences as opposed to 57% for a ranking using a model without covariance. The sort 1 variants were placed in rank order using a model based solely only positional enrichment data (blue bars) or one incorporating amino acid covariance (green bars). The positional enrichment data ranked each variant solely on the individual strength of each mutation. The sort 5 data was used as a positive control test set against the ranked sort 1.

Figure 2.14



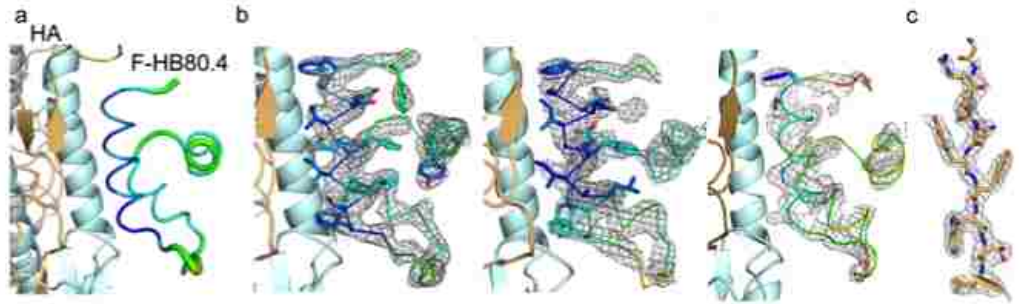
Tables showing the results of the HB80.3 and HB36.4 affinity maturation libraries. The initial library shows which residues were available at each position; red indicates residues predicted to be advantageous, black is neutral and starting sequence, and green deleterious according to the $\log_2$ enrichment maps. The deleterious residues are carried over from degenerate codons needed to access the advantageous residues. The best variants show the dominant sequences for each design variant following 5 sorts, differing at 5 positions from HB80.3 and 8 positions for HB36.4.

Figure 2.15



Sample SPR sensorgram for HB80.4 **a,** and HB80.4.1 **b,** binding to SC1918/H1 HA. H1 HA was immobilized at 500 response units (RU) on a SA chip and soluble binder was flowed over at 100μL/min at 8 different concentrations ranging from 0.1-12.8nM. Orange line represents best fit to the data using global fitting.

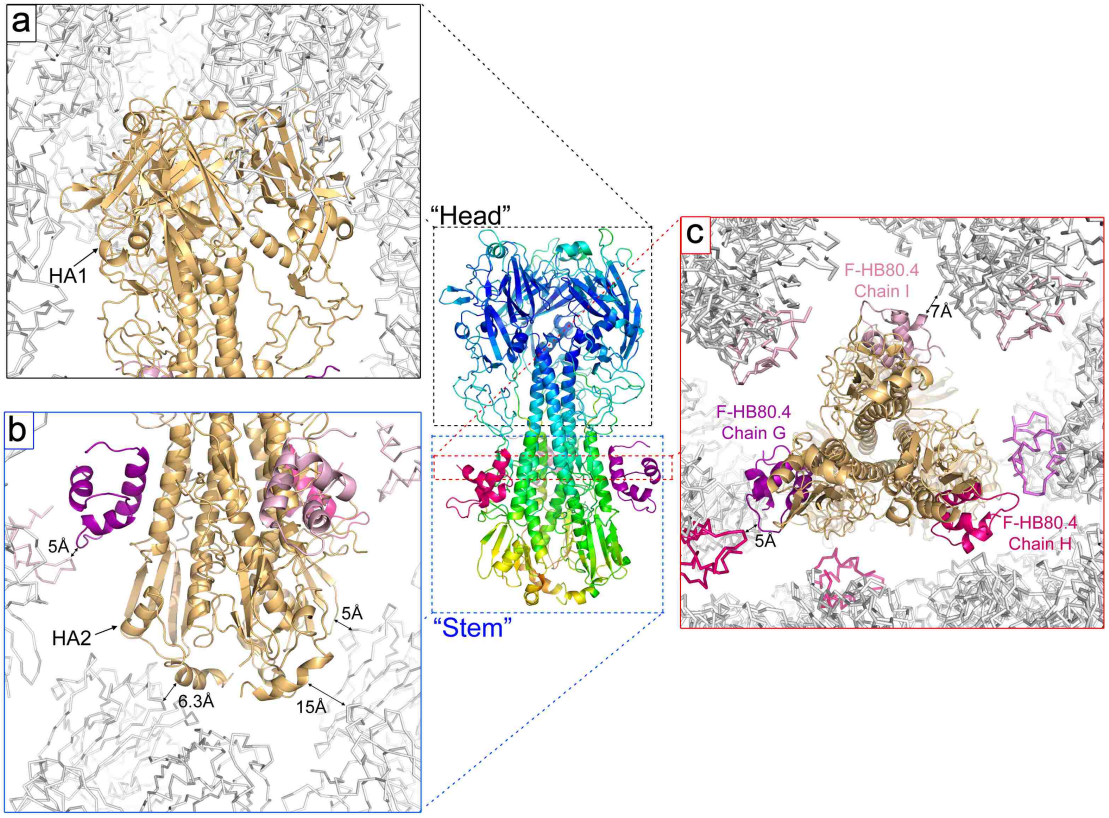
Figure 2.16



Plasticity in F-HB80.4. **(a)** B-value representation of the three F-HB80.4 molecules present in the asymmetric unit of the F-HB80.4-Sc1918/HA crystal structure. Warmer colors (green, yellow) and thicker ribbon indicate great conformational flexibility (range from 73-164 A^2). The Sc1918 HA1 subunit is represented in light yellow, and the HA2 subunit in light blue. **(b)** $2F_{obs}$-F_{calc} (gray mesh, contoured at 1σ) electron density map for the three F-HB80.4 molecules in the asymmetric unit. The overall backbone conformation of F-HB80.4 agrees well with the electron density maps, but atomic displacement parameters are elevated and few features, such as some side chains, are not apparent for residues distant from the F-HB80.4-HA interface. Chain G is represented on the left, chain H in the middle, and chain I on the right. **(c)** Representative $2F_{obs}$-F_{calc} electron density map (gray mesh, contoured at 2σ) for the HA1 head (section around Asn231, Tyr232, Tyr233, Try234 and Thr235 shown), which is well-ordered and reflects the resolution of the data and data quality (see Supplementary Table 12). The paucity of

crystal lattice contacts in the HA stem and especially around the inhibitor leads to some conformational heterogeneity or disorder as indicated by higher B-values.

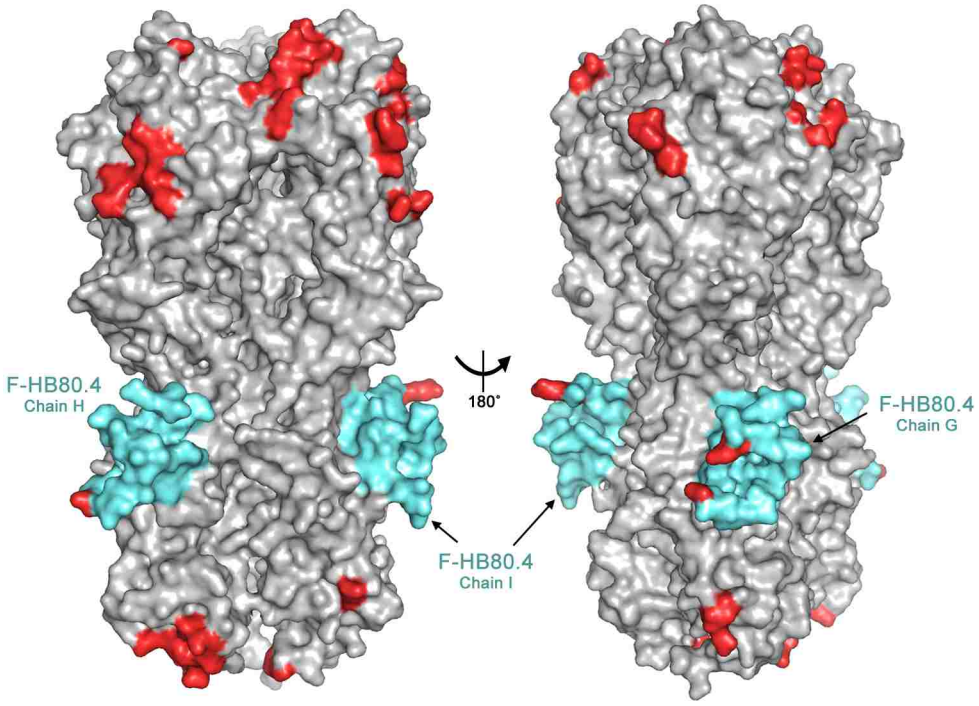
Figure 2.17



Lack of crystal lattice contacts contributes to high thermal disorder in HA2 and F-HB80.4. The B-value (20-190 A^2 from blue to orange) representation of the Sc1918/HA is shown in the middle. F-HB80.4 is colored in magenta, purple and pink for chain G, H and I, respectively (n.b. one F-HB80.4 is not very visible in the view as it on the back side of the trimer). The average B-values are 62 $Å^2$ for HA1, 106 $Å^2$ for HA2 and 133 $Å^2$ for F-HB80.4. **(a)** Packing around the head of the HA is extensive. The head of the trimeric HA is packed against 7 chains from the symmetry-related molecules (5 HA1 and 2 HA2). **(b)** Many fewer packing interactions are made around the HA stem resulting in more intrinsic disorder. The stem makes crystal contacts with the HA1 chains of only two symmetry-related molecules. **(c)** The view of

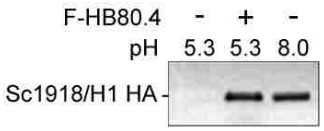
the crystal packing along the three-fold axis around the HA stem shows the paucity of the crystal contacts around the inhibitors. Crystallographic symmetry is represented in grey except for the F-HB80.4 design, which is colored in magenta, purple and pink. The shortest contacts between the backbone of the asymmetric unit and symmetry-related molecules are labeled.

Figure 2.18



Amino acids contacted by symmetry related molecules on F-HB80.4-Sc1918/HA. Crystal contacts made by amino acid within 5 Å are represented in red. HA is colored in grey and F-HB80.4 in cyan.

Figure 2.19



F-HB80.4 inhibits the pH-induced conformational changes that drive membrane fusion. Exposure to low pH converts 1918 H1 HA to a protease-susceptible state (lane 1), whereas HAs maintained at neutral pH are highly resistant to trypsin (lane 3). Preincubation of F-HB80.4 with H1 prevents pH-induced conformational changes and retains the HAs in the protease-resistant, prefusion state (lane 2).

Table 2.1.

DNA sequences of the single site saturation mutagenesis libraries. Base pairs shaded in light blue indicate start and end of design encoding sequence, while base pairs shaded in red indicate region of single site saturation mutagenesis.

>HB36.4

GACGATTGAAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTGGTGG
AGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTAGCCATATGCACATGT
CCAATGCTATGGATGGTCAACAATTGAACAGATTGTTATTGGAATGGATCGGTGCCTGGGA
CCCTTTTGGTTTGGGTAAAGATGCTTATGACGTCGAAGCCGAAGCTGTTTTACAAGCAGTA
TACGAAACTGAATCTGCATTTGATTTGGCCATGAGAATTATGTGGATCTATGTTTTTGCCTT
CAAGAGACCAATTCCTTTCCCACACGCTCAAAAATTGGCAAGAAGATTATTGGAATTGAAG
CAAGCTGCATCTTCACCTTTACCATTGGAACTCGAGGGGGGCGGATCCGAACAAAAGCTTA
TTTCTGAAGAGGACTTGTAATAGAGATCT

>HB80.3

GACGATTGAAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTGGTGG
AGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTAGCCATATGGCTTCTA
CTAGAGGTTCTGGTAGACCTTGGGGTTTTTCCGAAAATTTGGCCTTCGAATTGGCTTTAAGT
TTTACTAACAAAGATACACCAGACAGATGGGCTAAGGTTGCACAATATGTATCTGGTAGAA
CACCTGAAGAAGTTAAAAAGCATTACGAACTCGAGGGGGGCGGATCCGAACAAAAGCTTA
TTTCTGAAGAGGACTTGTAATAGAGATCT

Table 2.2

Summary of selection experiments undertaken in this study against the HB36.4 SSM library. ***Labeling Condition*** indicates how cells were labeled and prepared before fluorescence activated cell sorting (FACS). Many of the selections were off-rate. For off-rate selections, after labeling cells with indicated concentration of HA, cells were thoroughly washed and then incubated at 22°C in the presence of 1 mM of the designated soluble protein for the indicated time. ***PE Gate % Cells*** is the % of displaying cells that were collected by setting flow cytometry gates above a certain threshold in the PE fluorescence channel associated with HA binding events (for further details see **Supplementary Figure 2**). # ***Cells Collected*** is the total number of cells collected through FACS. # ***DNA reads*** is the number of DNA sequences that pass through the Illumina sequencing quality filters using a PHRED score of 30[3].

Selected Population	Sort	Labeling Condition	PE Gate % Cells	# Cells Collected	# DNA reads	# Possible non-synonymous substitutions
Naïve library						
	0	--	--	--	1.20E+06	1052/1060
Reference for sort 1						
	1	--	--	2.5E+05	1.50E+06	518/520
Selected for display of c-myc epitope tag only						
	1	--	100%	2.5E+05	1.37E+06	518/520
Reference for sort 2						
	2	--	--	2.5E+05	1.53E+06	1039/1060
Medium stringency binding to H1 HA						

1	18 nM H1	41%	2.5E+05	1.61E+06	1045/1060
2	3.5 nM H1	10%	1.6E+05	1.47E+06	914/1060
Medium stringency binding to H5 HA					
1	36 nM H5	33%	1.5E+05	1.58E+06	1026/1060
2	6 nM H5	6%	6.0E+04	1.51E+06	977/1060
Reference for sort 2					
2	--	--	1.5E+05	1.58E+06	1043/1060
High stringency binding to H1 HA					
1	4nM H1	19%	1.5E+05	1.51E+06	1003/1060
2	6nM H1, 120' off with HB80.3	3%	9.0E+04	1.72E+06	1003/1060

Table 2.3.

Summary of selection experiments undertaken in this study against the HB80.3 SSM library. ***Labeling condition*** indicates how cells were labeled and prepared before fluorescence activated cell sorting (FACS). Many of the selections were off-rate. For off-rate selections, after labeling cells with indicated concentration of HA, cells were thoroughly washed and then incubated at 22^{o}C in the presence of 1 mM of the designated soluble protein for the indicated time. ***PE Gate % Cells*** is the % of displaying cells that were collected by setting flow cytometry gates above a certain threshold in the PE fluorescence channel associated with HA binding events (for further details see **Figure 2.2**). ***# Cells Collected*** is the total number of cells collected through FACS. ***# DNA reads*** is the number of DNA sequences that pass through the Illumina sequencing quality filters using a PHRED score of 30[6].

Selected Population	Sort	Labeling Condition	PE Gate % Cells	# Cells Collected	# DNA reads	# Possible non-synonymous substitutions
Naïve library						
	0	--	--	--	1.93E+06	1012/1020
Reference for sort 2						
	2	--	--	1.5E+05	2.30E+06	1006/1020
High stringency binding to H1 HA						
	1	4nM H1	21%	1.5E+05	2.17E+06	1008/1020
	2	6nM H1, 40' off with HB80.3	2%	6.0E+04	2.58E+06	942/1020
Reference for sort 3 frag 1						
	1	--	--	5.0E+05	2.17E+06	492/520
Selected for display of c-myc epitope tag only						
	1	--	100%	5.0E+05	4.4E+06	492/520
Selected for better display of c-myc epitope tag only						
	1	--	9%*	5.0E+05	9.23E+06	490/520

* A gate on FITC fluorescent channel was drawn to sort this population

Table 2.4

Comparison of the mutations in this study to published results[5].

Design	Mutation	Deep Sequencing	Previous examples	Substitutions relative to starting design variant	Agreement

HB36.4	S47D	Deleterious	Deleterious	K64N,V60A	Yes
HB36.4	S47H	Neutral	Neutral	K64N,V60A	Yes
HB36.4	S47W	Deleterious	Deleterious	K64N,V60A	Yes
HB36.4	S47R	Deleterious	Deleterious	K64N,V60A	Yes
HB36.4	S47E	Deleterious	Deleterious	K64N,V60A	Yes
HB36.4	V60A	Slightly deleterious	Deleterious	K64N,S47D	Yes
HB36.4	K64N	Slightly deleterious	Slightly deleterious	--	Yes
HB36.4	F49A	Deleterious	Knockout	K64N	Yes
HB36.4	M53A	Deleterious	Deleterious	K64N	Yes
HB36.4	W57A	Deleterious	Knockout	K64N	Yes
HB80.3	T26M	Deleterious	Deleterious	A24S,G12D	Yes
HB80.3	K36N	Slightly deleterious	Deleterious	A24S,G12D	Yes
HB80.3	F13A	Deleterious	Knockout	A24S,G12D	Yes
HB80.3	F25A	Deleterious	Knockout	A24S,G12D	Yes
HB80.3	Y40A	Deleterious	Deleterious	A24S,G12D	Yes

Table 2.5.

3x3 contingency tables of the (**TOP)** HB36.4 and **(BOTTOM)** HB80.3 of comparisons between computational recapitulation and the experimental deep sequencing dataset for charged substitutions. Computational data was binned at >0.5 Rosetta energy units (R.E.U), ± 0.5 R.E.U., and <-0.5 R.E.U., while experimental data was binned at <-2 log2 enrichment, ± 2 log2 enrichment, and >2 log2 enrichment. The top black numbers represent the computational data using the original Rosettadesign energy function, while the lower red numbers represent bins using the energy function with the additional electrostatics term. Inclusion of the additional electrostatics term improves the statistical significance of the correlation between the experimental data and the computational recapitulation (two-tailed p-values decrease from 0.0131 to <0.0001 for HB36.4 and from 0.0165 to <0.0001 for HB80.3).

HB36.4

Computational Recapitulation	Experimental data		
	log_2enrich < -2	-2 < log2enrich < 2	log_2enrich > 2
> 0.5 R.E.U.	99 *103*	62 *51*	5 *0*
-0.5 > > 0.5 R.E.U.	11 *8*	14 *16*	4 *2*
< -0.5 R.E.U.	1 *0*	4 *13*	0 *7*

HB80.3

Computational Recapitulation	Experimental data		
	log_2enrich < -2	-2 < log2enrich < 2	log_2enrich > 2
> 0.5 R.E.U.	95 *85*	35 *35*	5 *1*
-0.5 > > 0.5 R.E.U.	15 *8*	18 *12*	4 *0*
< -0.5 R.E.U.	1 *18*	1 *7*	0 *8*

Table 2.6.

DNA sequences of the affinity maturation libraries constructed from the information

contained in the deep sequencing experiment. Codons bolded and in red indicate are degenerate and thus encode more than one amino acid.

>HB80.3_library

GACGATTGAAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTGGTGG
AGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTAGCCATATG

GCT TCT ACT AGA GGT TCT GGT AGA CCT TGG **RRG** TTT **ARS** GAA AAT **VTT RMG** TTC GAA **MTT** GCT TTA **TMT** TTT ACT AAC AAA GAT ACA CCA GAC AGA TGG **RVG** AAG GTT GCA **YDS** TAT GTA **ARS** GGT AGA ACA CCT GAA GAA GTT AAA AAG CAT TAC GAA

CTCGAGGGGGGCGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAG

AGATCT

>HB36.4_elibrary

GACGATTGAAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTGGTGG
AGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTAGCCATATGCACATGT
CCAATGCTATGGATGGTCAACAATTGAACAGATTGTTATTGGAATGGATCGGTGCCTGGGA
C

CCT TTT GGT TTG GGT AAA GAT GCT TAT **GMT KWT** GAA GCC GAA **RVA** GTT TTA **MAG** GCA GTA TAC **GMG** ACT **RAM YMT** GCA TTT GAT TTG GCC ATG AGA ATT **MWK** TGG ATC TAT **RWT** TTT GCC **TWT** AAG AGA **MMG** ATT CCT TTC **VYA** CAC GCT CAA AAA TTG GCA AGA

AGATTATTGGAATTGAAGCAAGCTGCATCTTCACCTTTACCATTGGAACTCGAGGGGGGCG
GATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGAGATCT

Table 2.7

Mutations found in individual sequenced variants from sort 4 and 5 of the designed HB36.4 library. Yellow positions denote wild-type positions. Variant E14 (bolded) was renamed FL-HB36.5 in the main text.

ID	SORT	33V	36E	37A	40Q	44E	46E	47S	56M	60V	63F	66P	70P
start	-	V	E	A	Q	E	E	S	M	V	F	P	P
E01	4	D	A	A	Q	E	N	H	H	V	F	K	L

E02	4	V	A	A	K	A	N	S	I	V	Y	K	A
E03	4	F	D	K	Q	E	N	S	N	V	F	P	V
E04	4	V	A	A	K	E	N	S	N	V	F	K	A
E05	4	V	D	K	Q	D	N	S	H	N	F	K	L
E06	4	D	D	R	Q	E	N	S	N	V	F	T	A
E07	4	Y	D	K	Q	E	N	S	H	I	F	P	V
E08	4	V	D	A	K	E	N	S	H	N	F	K	V
E09	4	D	D	K	Q	A	N	S	H	N	Y	T	V
E10	4	D	A	R	K	A	D	S	H	N	F	K	L
E11	4	V	D	K	Q	A	N	S	H	I	F	T	I
E12	4	Y	D	E	K	A	N	S	H	N	F	K	T
E13	4	V	D	K	Q	A	N	S	H	I	F	T	I
E14	**5**	**Y**	**A**	**E**	**K**	**E**	**E**	**S**	**H**	**N**	**F**	**Q**	**A**
E15	5	D	A	R	K	E	D	S	H	N	F	Q	L
E16	5	Y	D	E	K	?	N	S	H	N	F	Q	D
E17	5	V	A	R	Q	E	N	S		N	F	K	I
E18	5	V	D	K	Q	E	N	S	H	N	F	K	L
E19	5	Y	D	A	K	A	N	S	H	N	F	K	P
E20	5	F	A	E	K	E	E	H	H	N	F	P	I

Table 2.8.

Mutations found in individual sequenced variants from sort 4 and 5 of the HB80.3 designed library. Yellow positions denote wild-type positions. Variant E16 was renamed FL-HB80.4 and E31 is F-HB80.4.1 in the main text.

ID	Sort	12G	14S	17L	18A	21L	24S	35A	39Q	42S
start	**-**	**G**	**S**	**L**	**A**	**L**	**S**	**A**	**Q**	**S**

E01	4	G	S	I	A	L	Y	G	R	R
E02	4	G	R	I	A	I	Y	R	R	K
E03	4	K	S	V	A	I	S	A	R	R
E04	4	K	S	V	A	L	Y	G	Q	R
E05	4	K	S	V	A	L	Y	A	R	K
E06	4	K	S	V	A	L	Y	T	R	K
E07	4	K	S	V	A	L	Y	K	R	K
E08	4	K	S	I	A	L	Y	K	R	K
E09	4	R	S	V	A	I	S	K	R	K
E10	4	R	S	V	A	I	S	K	R	R
E11	4	R	S	V	A	I	S	R	R	R
E12	4	K	S	V	A	I	S	R	R	R
E13	4	K	S	V	A	I	S	R	Y	R
E14	4	R	S	V	A	I	S	R	R	R
E15	4	R	S	V	A	I	S	A	R	R
E16	**4**	**K**	**S**	**I**	**A**	**I**	**S**	**K**	**Q**	**K**
E17	4	K	S	V	A	I	S	R	R	R
E18	4	K	S	L	A	L	Y	A	Y	K
E19	4	K	K	L	E	I	S	K	Y	R
E20	5	K	S	V	A	I	S	K	R	R
E21	5	K	S	V	A	I	S	T	R	R
E22	5	K	S	V	A	I	S	A	Y	R
E23	5	K	S	I	A	I	S	K	S	K
E24	5	K	S	I	A	I	S	K	P	K
E25	5	K	S	I	A	I	S	K	Y	K
E26	5	R	S	V	A	I	S	R	R	R
E27	5	R	S	V	A	I	S	A	Y	R
E28	5	K	S	V	A	I	S	A	Y	R
E29	5	N	S	I	A	L	S	K	R	K

E30	5	K	S	I	A	L	Y	K	R	K
E31	**-**	**R**	**S**	**L**	**A**	**L**	**S**	**R**	**Q**	**R**

Table 2.9.

Table showing dissociation constants (K_d) as determined by surface plasmon resonance for variants with corresponding enrichment data from next-gen sequenced selections both against SC/1918 H1 HA[6]. Bold indicates best variants and asterisk low counts in the sequenced pool.

Design Variant	Mutations	Kd [nM]	log2 enrich
HB80.3	0	10	1.31
F-HB80.4.1	3	1.3	1.586
F-HB80.4	**5**	**0.6**	**3.57**
HB36.4	0	26.1	0.1
HB36.3	1	29	-4.449
F-HB36.5	**8**	**0.9**	**1.1***
F-HB36.E13	10	2.8	4.28
F-HB36.E20	8	2.8	2.81

Table 2.10.

Data collection and refinement statistics for the F-HB80.4-SC1918/H1 HA complex.

Data collection	F-HB80.4-SC1918/H1 Complex
Beamline	APS ID23D
Wavelength (Å)	1.03326
Space group	$P2_12_12_1$
Unit cell parameters	a = 72.33 Å b = 126.15 Å c = 243.29 Å
	$\alpha = \beta = \gamma = 90.0^{\circ}$

Resolution (Å)	50 - 2.7 (2.75 – 2.70)[a]
Observations	244,858 (10,131)
Unique reflections	58,197 (2,469) [a]
Redundancy	4.2 (4.0) [a]
Completeness (%)	94.1 (81.2) [a]
$\langle I/\sigma_I \rangle$	22.8 (3.5) [a]
R_{sym} [b]	0.05 (0.43) [a, b]
Z_a [c]	3
Refinement statistics	
Resolution (Å)	50 - 2.7
Reflections (work)	58,120
Reflections (test)	2,939
R_{cryst} (%)[d]	22.8 [d]
R_{free}(%)[e]	28.5 [e]
Average B-value (Å^2)	
HA	77
HA1	62
HA2	106
Inhibitor	133
Wilson B-value (Å^2)	60
Protein atoms	12,846
Carbohydrate atoms	116
Waters	139
RMSD from ideal geometry	
Bond length (Å)	0.009
Bond angles (°)	1.37
Ramachandran statistics (%)[f]	
Favored	96.3
Outliers	0.06

PDB ID	4EEF[g]

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

[b] $R_{sym} = \Sigma_{hkl} \Sigma_i | I_{hkl,i} - \langle I_{hkl} \rangle | / \Sigma_{hkl} \Sigma_i I_{hkl,I}$ and $R_{pim} = \Sigma_{hkl} (1/(n-1))^{1/2} \Sigma_i | I_{hkl,i} - \langle I_{hkl} \rangle | / \Sigma_{hkl} \Sigma_i I_{hkl,I}$, where $I_{hkl,i}$ is the scaled intensity of the i^{th} measurement of relection h, k, l, $\langle I_{hkl} \rangle$ is the average intensity for that reflection, and n is the redundancy[51].

[c] Z_a is the number of HA monomer-Fab complexes per crystallographic asymmetric unit.

[d] $R_{cryst} = \Sigma_{hkl} | F_o - F_c | / \Sigma_{hkl} | F_o |$ x 100

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

[f] Calculated using Molprobity[43].

[g] Coordinates and structure factors will be deposited in the PDB prior to publication and be available immediately on publication.

Table 2.11.

List of primers used for this study. Use denotes the way in which the primer was used: HB36.4 & HB80.3 ssm primers were used to amplify the genes encoding the protein binders from plasmid DNA. Each population had a unique 8 bp barcode (bolded) that was appended to the reverse ('rev' in the primer name) primer – this allowed quantification of separate populations on the same Illumina flowcell. **Next-gen** primers were primers used during Illumina sequencing: the DNA encoding the protein libraries was sequenced using two sets of sequencing primers ('f1', 'f2') on two separate flowcells. The 'index' primer was used to sequence the barcode. **Universal** primers were upstream and downstream of the entire protein-encoding insert in the yeast display pETCON plasmid. **Elibrary** primers were used for the construction of the libraries shown in **Table 2.5**.

Primer Name	Sequence	Use
PCR77_fwd	AATGATACGGCGACCACCGAGATCTACACcggctagccatatggcttct	HB80.3ssm
PCR77_rev_BC1	CAAGCAGAAGACGGCATACGAGAT**CAAGGTCA**gatccgcccccctcgag	HB80.3ssm

PCR77_rev_BC10	CAAGCAGAAGACGGCATACGAGAT**ACGTACTC**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC11	CAAGCAGAAGACGGCATACGAGAT**CTTCTAAG**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC12	CAAGCAGAAGACGGCATACGAGAT**ACTATGAC**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC13	CAAGCAGAAGACGGCATACGAGAT**GACGTTAA**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC14	CAAGCAGAAGACGGCATACGAGAT**ACAAGATA**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC15	CAAGCAGAAGACGGCATACGAGAT**GACTAAGA**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC16	CAAGCAGAAGACGGCATACGAGAT**GTGTCTAC**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC17	CAAGCAGAAGACGGCATACGAGAT**TTCACTAG**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC18	CAAGCAGAAGACGGCATACGAGAT**AATCGGAT**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC19	CAAGCAGAAGACGGCATACGAGAT**AGTACCGA**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC2	CAAGCAGAAGACGGCATACGAGAT**GCATAACT**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC3	CAAGCAGAAGACGGCATACGAGAT**CTCTGATT**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC30	CAAGCAGAAGACGGCATACGAGAT**GTAGCAGT**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC31	CAAGCAGAAGACGGCATACGAGAT**GGATCATC**gatccgcccccctcgag	HB80.3ssm
PCR77_rev_BC32	CAAGCAGAAGACGGCATACGAGAT**GTGAACGT**gatccgcccccctcgag	HB80.3ssm
HA77_f1_fwd	cggctagccatatggcttct	Next-gen
HA77_f1_rev	gtgcaaccttagcccatctgtctggtg	Next-gen
HA77_f2_fwd	ggccttcgaattggctttaagttttactaacaaagat	Next-gen
HA77_f2_rev	gatccgcccccctcgag	Next-gen
HA77_index	ctcgaggggggcggatc	Next-gen
PCR35_fwd	AATGATACGGCGACCACCGAGATCTACACgatcggtgcctgggac	HB36.4ssm
PCR35_rev_BC20	CAAGCAGAAGACGGCATACGAGAT**TTGCCTCA**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC21	CAAGCAGAAGACGGCATACGAGAT**TCGTTAGC**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC22	CAAGCAGAAGACGGCATACGAGAT**TATAGTTC**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC23	CAAGCAGAAGACGGCATACGAGAT**TGGCGTAT**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC24	CAAGCAGAAGACGGCATACGAGAT**TGGACATG**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC25	CAAGCAGAAGACGGCATACGAGAT**AGGTTGCT**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC26	CAAGCAGAAGACGGCATACGAGAT**ATATGCTG**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC27	CAAGCAGAAGACGGCATACGAGAT**GTACAGTG**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC40	CAAGCAGAAGACGGCATACGAGAT**AATCCTGC**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC41	CAAGCAGAAGACGGCATACGAGAT**GTTATATC**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC42	CAAGCAGAAGACGGCATACGAGATACACACGTcagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC43	CAAGCAGAAGACGGCATACGAGAT**ATACGACT**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC44	CAAGCAGAAGACGGCATACGAGAT**ATCTTCGT**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC45	CAAGCAGAAGACGGCATACGAGAT**ACATGTAT**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC46	CAAGCAGAAGACGGCATACGAGAT**TCCACAGT**cagcttgcttcaattccaataatc	HB36.4ssm
PCR35_rev_BC47	CAAGCAGAAGACGGCATACGAGAT**CAGTCTGT**cagcttgcttcaattccaataatc	HB36.4ssm
HA35_f1_fwd	gatcggtgcctgggac	Next-gen
HA35_f1_rev	tcttgaaggcaaaaacatagatccacataattctcatgg	Next-gen
HA35_f2_fwd	acaagcagtatacgaaactgaatctgcatttgatttgg	Next-gen
HA35_f2_rev	cagcttgcttcaattccaataatc	Next-gen
HA35_index	gattattggaattgaagcaagct	Next-gen
Up-GS-pCons	ggacaatagctcgacgattgaaggtagatacccata	Universal
Down_Cmyc	caagtcctcttcagaaataagcttttgttc	Universal
HB80_front_rev	tggtctaccggaacctctggtggatgc	Elibrary

HB80_back_fwd	actcctgaagaagtcaaaaagcattacgaa	Elibrary
HB80_klenow	ttcgtaatgctttttgacttcttc	Elibrary
E80 ultramer	Gcatccaccagaggttccggtagaccatggrrgttcarsgaaaacvttrmgtttgaamt	Elibrary
	Tgctttgtmttttacgaataaggacacaccagatagatggrvgaaggttgcayrstatgt	construction
HB36_front_rev	gtcataggcatctttacccaaacc	Elibrary
HB36_back_fwd	catgcccaaaagttggctaga	Elibrary
HB36_klenow	tctagccaacttttgggcatgt	Elibrary
E36 ultramer	Cctttggtttgggtaaagatgcctatgackwtgaagccgmtrvagttttamaggcagta	Elibrary
	Tacgmgactramymtgcttttgacttggcaatgagaattmwktggatctatrwttttgcct	construction

Table 2.12.

Summary of selection experiments undertaken in this study against the HB36.4 and HB80.3 designed libraries. ***Labeling condition*** indicates how cells were labeled and prepared before fluorescence activated cell sorting (FACS). Many of the selections were off-rate. For off-rate selections, after labeling cells with indicated concentration of HA, cells were thoroughly washed and then incubated at 22^{o}C in the presence of 1 mM of the designated soluble protein for the indicated time. ***PE Gate % Cells*** is the % of displaying cells that were collected by setting flow cytometry gates above a certain threshold in the PE fluorescence channel associated with HA binding events (for further details see **Figure 2.2**). ***# Cells Collected*** is the total number of cells collected through FACS. ***# DNA reads*** is the number of DNA sequences that pass through the Illumina sequencing quality filters using a PHRED score of 30[6].

Selected Population HB36.4	Sort	Labeling Condition	PE Gate % Cells	# Cells Collected	# DNA reads
Naïve library					
	0	--	--	--	3.60E+05
Stringent Sort 1					
	2	8 nM	3%	7.00E+05	7.34E+05
Stringent Sort 2					
	1	0.5nM	2%	1.70E+05	--
Stringent Sort 3					
	1	1nM	4%	1.50E+05	--
		240' off with HB36.4			
Strignet Sort 4					
	1	1nM	1%	9.00E+04	--
		120' off with HB36.4 at 22°C			
		60' off with HB36.4 at 37°C			
Stringent Sort 5					
	1	1nM	2%*	7.00E+04	--
		120' off with FL HB80.4 at 22°C			
		60' off with FL HB80.4 at 37°C			

Selected Population HB80.3	Sort	Labeling Condition	PE Gate % Cells	# Cells Collected	# DNA reads
Naïve library					
	0	--	--	--	1.14E+06
Stringent Sort 1					
	2	8 nM	3%	7.00E+05	2.09E+06
Stringent Sort 2					
	1	0.5nM	2%	1.70E+05	--
Stringent Sort 3					
	1	1nM	4%	1.50E+05	--
		120' off with HB36.4			
Strignet Sort 4					
	1	1nM	1%	9.00E+04	--
		120' off with HB36.4 at 22°C			
		30' off with HB36.4 at 37°C			
Stringent Sort 5					
	1	1nM	2%*	7.00E+04	--
		120' off with FL HB80.4 at 22°C			
		30' off with FL HB80.4 at 37°C			

Table 2.13.

Sequences of HA proteins used in binding studies. The sequences listed below represent the full-length ORF as cloned in the baculovirus transfer vector. Most of the N-terminal signal peptide (MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFA) is presumably removed during secretion, leaving four non-native residues (ADPG) at the N-terminus of HA1. The C-terminal biotinylation site, trimerization domain, and His-tag are retained on all proteins.

>A/South Carolina/1/1918 (H1N1)
MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLED SHNGKLCKLKGIAPLQLGKCNIAGWLLGNPECDLLLTASSWSYIVETSNSENGTCYPGDFIDYEELREQLSSVSSFE KFEIFPKTSSWPNHETTKGVTAACSYAGASSFYRNLLWLTKKGSSYPKLSKSYVNNKGKEVLVLWGVHHPPTGTDQQ SLYQNADAYVSVGSSKYNRRFTPEIAARPKVRDQAGRMNYYWTLLEPGDTITFEATGNLIAPWYAFALNRGSGSGII TSDAPVHDCNTKCQTPHGAINSSLPFQNIHPVTIGECPKYVRSTKLRMATGLRNIPSIQSRGLFGAIAGFIEGGWTG MIDGWYGYHHQNEQGSGYAADQKSTQNAIDGITNKVNSVIEKMNTQFTAVGKEFNNLERRIENLNKKVDDGFLDIWT YNAELLVLLENERTLDFHDSNVRNLYEKVKSQLKNNAKEIGNGCFEFYHKCDDACMESVRNGTYDYPKYSEESKLNR EEIDGVSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH
>A/Japan/305/1957 (H2N2)
MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDQICIGYHANNSTEKVDTILERNVTVTHAKDILEK THNGKLCKLNGIPPLELGDCSIAGWLLGNPECDRLLSVPEWSYIMEKENPRDGLCYPGSFNDYEELKHLLSSVKHFE KVKILPKDRWTQHTTTGGSRACAVSGNPSFFRNMVWLTEKGSNYPVAKGSYNNTSGEQMLIIWGVHHPNDETEQRTL YQNVGTYVSVGTSTLNKRSTPEIATRPKVNGQGGRMEFSWTLLDMWDTINFESTGNLIAPEYGFKISKRGSSGIMKT EGTLENCETKCQTPLGAINTTLPFHNVHPLTIGECPKYVKSEKLVLATGLRNVPQIESRGLFGAIAGFIEGGWQGMV DGWYGYHHSNDQGSGYAADKESTQKAFDGITNKVNSVIEKMNTQFEAVGKEFSNLERRLENLNKKMEDGFLDVWTYN AELLVLMENERTLDFHDSNVKNLYDKVRMQLRDNVKELGNGCFEFYHKCDDECMNSVKNGTYDYPKYEEESKLNRNE IKSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/Adachi/2/1957 (H2N2)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDQICIGYHANNSTEKVDTILERNVTVTHAKDILEK
THNGKLCKLNGIPPLELGDCSIAGWLLGNPECDRLLSVPEWSYIMEKENPRNGLCYPGSFNDYEELKHLLSSVKHFE
KVKILPKDRWTQHTTTGGSQACAVSGNPSFFRNMVWLTKKGSDYPVAKGSYNNTSGEQMLIIWGVHHPIDETEQRTL
YQNVGTYVSVGTSTLNKRSTPEIATRPKVNGLGSRMEFSWTLLDMWDTINFESTGNLIAPEYGFKISKRGSSGIMKT
EGTLENCETKCQTPLGAINTTLPFHNVHPLTIGECPKYVKSEKLVLATGLRNVPQIESRGLFGAIAGFIEGGWQGMV
DGWYGYHHSNDQGSGYAADKESTQKAFDGITNKVNSVIEKMNTQFEAVGKEFGNLERRLENLNKKMEDGFLDVWTYN
AELLVLMENERTLDFHDSNVKNLYDKVRMQLRDNVKELGNGCFEFYHKCDDECMNSVKNGTYDYPKYEEESKLNRNE
IKSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/Vietnam/1203/2004 (H5N1)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEK
KHNGKLCDLDGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKANPVNDLCYPGDFNDYEELKHLLSRINHFE
KIQIIPKSSWSSHEASLGVSSACPYQGKSSFFRNVVWLIKKNSTYPTIKRSYNNTNQEDLLVLWGIHHPNDAAEQTK
LYQNPTTYISVGTSTLNQRLVPRIATRSKVNGQSGRMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSTIMK
SELEYGNCNTKCQTPMGAINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNSPQRERRRKKRGLFGAIAGFIEGG
WQGMVDGWYGYHHSNEQGSGYAADKESTQKAIDGVTNKVNSIIDKMNTQFEAVGREFNNLERRIENLNKKMEDGFLD
VWTYNAELLVLMENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNGCFEFYHKCDNECMESVRNGTYDYPQYSEEAR
LKREEISSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

> A/Indonesia/05/2005 (H5N1)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEK
THNGKLCDLDGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKANPTNDLCYPGSFNDYEELKHLLSRINHFE
KIQIIPKSSWSDHEASSGVSSACPYLGSPSFFRNVVWLIKKNSTYPTIKKSYNNTNQEDLLVLWGIHHPNDAAEQTR
LYQNPTTYISIGTSTLNQRLVPKIATRSKVNGQSGRMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSAIMK
SELEYGNCNTKCQTPMGAINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNSPQRESRRKKRGLFGAIAGFIEGG
WQGMVDGWYGYHHSNEQGSGYAADKESTQKAIDGVTNKVNSIIDKMNTQFEAVGREFNNLERRIENLNKKMEDGFLD
VWTYNAELLVLMENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNGCFEFYHKCDNECMESIRNGTYNYPQYSEEAR
LKREEISSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/turkey/Massachusetts/3740/1965 (H6N2)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDKICIGYHANNSTTQVDTILEKNVTVTHSVELLES
QKEERFCRVLNKTPLDLKGCTIEGWILGNPQCDILLGDQSWSYIVERPGAQNGICYPGVLNEVEELKAFIGSGEKVQ
RFEMFPKSTWTGVDTNSGVTRACPYTTSGSSFYRNLLWIIKTRSAAYPVIKGTYNNTGSQPILYFWGVHHPPNTDEQ
NTLYGSGDRYVRMGTESMNFAKSPEIAARPAVNGQRGRIDYYWSVLKPGETLNVESNGNLIAPWYAYKFTSSNNKGA
IFKSNLPIENCDAVCQTVAGALKTNKTFQNVSPLWIGECPKYVKSESLRLATGLRNVPQAETRGLFGAIAGFIEGGW
TGMIDGWYGYHHENSQGSGYAADKESTQKAIDGITNKVNSIIDKMNTQFEAVEHEFSNLERRIDNLNKRMEDGFLDV
WTYNAELLVLLENERTLDLHDANVKNLYEKVKSQLRDNAKDLGNGCFEFWHKCDDECINSVKNGTYDYPKYQDESKL
NRQEIDSVSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/turkey/Wisconsin/1/1966 (H9N2)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDKICIGYQSTNSTETVDTLTESNVPVTHTKELLHT
EHNGMLCATDLGHPLILDTCTIEGLIYGNPSCDILLGGKEWSYIVERSSAVNGMCYPGNVENLEELRSLFSSAKSYK
RIQIFPDKTWNVTYSGTSRACSNSFYRSMRWLTHKSNSYPFQNAHYTNNERENILFMWGIHHPPTDTEQTDLYKNAD
TTTSVTTEDINRTFKPVIGPRPLVNGQQGRIDYYWSVLKPGQTLRIRSNGNLIAPWYGHVLTGESHGRILKTDLNNG
NCVVQCQTEKGGLNTTLPFHNISKYAFGNCPKYVGVKSLKLAVGLRNVPAVSSRGLFGAIAGFIEGGWPGLVAGWYG
FQHSNDQGVGMAADKGSTQKAIDKITSKVNNIIDKMNKQYEVIDHEFNELEARLNMINNKIDDQIQDIWAYNAELLV
LLENQKTLDEHDANVNNLYNKVKRALGSNAVEDGNGCFELYHKCDDQCMETIRNGTYDRQKYQEESRLERQKIEGVS
GGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/duck/Alberta/60/1976 (H12N5)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDTICVGYHANNSTDTVDTVLEKNVTVTHSVNLLED
SHNGKLCSLNGIAPLQLGKCNVAGWLLGNPECDLLLTANSWSYIIETSNSENGTCYPGEFIDYEELREQLSSISSFE
KFEIFPKASSWPNHETTKGVTAACSYSGASSFYRNLLWITKKGTSYPKLSKSYTNNKGKEVLVLWGVHHPPSVSEQQ
SLYQNADAYVSVGSSKYNRRFAPEIAARPEVRGQAGRMNYYWTLLDQGDTITFEATGNLIAPWYAFALNKGSDSGII
TSDAPVHNCDTRCQTPHGALNSSLPFQNVHPITIGECPKYVKSTKLRMATGLRNVPSIQSRGLFGAIAGFIEGGWTG
MIDGWYGYHHQNEQGSGYAADQKSTQNAIDGITNKVNSVIEKMNTQFTAVGKEFNNLERRIENLNKKVDDGFLDVWT
YNAELLVLLENERTLDFHDSNVRNLYEKVKSQLRNNAKEIGNGCFEFYHKCDDECMESVKNGTYDYPKYSEESKLNR
EEIDSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/gull/Maryland/704/1977 (H13N6)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDRICVGYLSTNSSERVDTLLENGVPVTSSIDLIET
NHTGTYCSLNGVSPVHLGDCSFEGWIVGNPACTSNFGIREWSYLIEDPAAPHGLCYPGELNNNGELRHLFSGIRSFS
RTELIPPTSWGEVLDGTTSACRDNTGTNSFYRNLVWFIKKNNRYPVISKTYNNTTGRDVLVLWGIHHPVSVDETKTL
YVNSDPYTLVSTKSWSEKYKLETGVRPGYNGQRSWMKIYWSLIHPGEMITFESNGGFLAPRYGYIIEEYGKGRIFQS
RIRMSRCNTKCQTSVGGINTNRTFQNIDKNALGDCPKYIKSGQLKLATGLRNVPAISNRGLFGAIAGFIEGGWPGLI
NGWYGFQHQNEQGTGIAADKESTQKAIDQITTKINNIIDKMNGNYDSIRGEFNQVEKRINMLADRIDDAVTDIWSYN
AKLLVLLENDKTLDMHDANVKNLHEQVRRELKDNAIDEGNGCFELLHKCNDSCMETIRNGTYDHTEYAEESKLKRQE
IDGISGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/black-headed gull/Sweden/4/99 (H16N3)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDKICIGYLSNNSTDTVDTLTENGVPVTSSIDLVET
NHTGTYCSLNGVSPIHLGDCSFEGWIVGNPSCASNINIREWSYLIEDPNAPHKLCFPGEVDNNGELRHLFSGVNSFS
RTELIPPSKWGDILEGTTASCQNRGANSFYRNLIWLVNKLNKYPVVKGEYNNTTGRDVLVLWGIHHPDTEATANKLY
VNKNPYTLVSTKEWSRRYELEIGTRIGDGQRSWMKIYWHLMHPGERITFESSGGLLAPRYGYIIEKYGTGRIFQSGV
RLAKCNTKCQTSMGGINTNKTFQNIERNALGDCPKYIKSGQLKLATGLRNVPSIVERGLFGAIAGFIEGGWPGLING
WYGFQHQNEQGTGIAADKTSTQKAINEITTKINNIIEKMNGNYDSIRGEFNQVEKRINMIADRVDDAVTDIWSYNAK
LLVLIENDRTLDLHDANVRNLHEQIKRALKDNAIDEGDGCFSILHKCNDSCMETIRNGTYNHEDYKEESQLKRQEIE
GISGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/Hong Kong/1/1968 (H3N2)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQS
SSTGKICNNPHRILDGIDCTLIDALLGDPHCDVFQNETWDLFVERSKAFSNCYPYDVPDYASLRSLVASSGTLEFIT
EGFTWTGVTQNGGSNACKRGPGSGFFSRLNWLTKSGSTYPVLNVTMPNNDNFDKLYIWGVHHPSTNQEQTSLYVQAS
GRVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPGDVLVINSNGNLIAPRGYFKMRTGKSSIMRSDAPIDT
CISECITPNGSIPNDKPFQNVNKITYGACPKYVKQNTLKLATGMRNVPEKQTRGLFGAIAGFIENGWEGMIDGWYGF
RHQNSEGTGQAADLKSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDLWSYNAELLVA
LENQHTIDLTDSEMNKLFEKTGRQLRENAEDMGNGCFKIYHKCDNACIESIRNGTYDHDVYRDEALNNRFQIKGVSG
GGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/duck/Czechoslovakia/1956 (H4N6)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGPVICMGHHAVANGTMVKTLADDQVEVVTAQELVES
QNLPELCPSPLRLVDGQTCDIINGALGSPGCDHLNGAEWDVFIERPNAVDTCYPFDVPEYQSLRSILANNGKFEFIA
EEFQWNTVKQNGKSGACKRANVNDFFNRLNWLVKSDGNAYPLQNLTKINNGDYARLYIWGVHHPSTDTEQTNLYKNN
PGRVTVSTKTSQTSVVPNIGSRPLVRGQSGRVSFYWTIVEPGDLIVFNTIGNLIAPRGHYKLNNQKKSTILNTAIPI
GSCVSKCHTDKGSLSTTKPFQNISRIAVGDCPRYVKQGSLKLATGMRNIPEKASRGLFGAIAGFIENGWQGLIDGWY
GFRHQNAEGTGTAADLKSTQAAIDQINGKLNRLIEKTNDKYHQIEKEFEQVEGRIQDLEKYVEDTKIDLWSYNAELL
VALENQHTIDVTDSEMNKLFERVRRQLRENAEDKGNGCFEIFHKCDNNCIESIRNGTYDHDIYRDEAINNRFQIQGV
SGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/Netherlands/219/2003 (H7N7)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDKICLGHHAVSNGTKVNTLTERGVEVVNATETVER
TNVPRICSKGKRTVDLGQCGLLGTITGPPQCDQFLEFSADLIIERREGSDVCYPGKFVNEEALRQILRESGGIDKET
MGFTYSGIRTNGTTSACRRSGSSFYAEMKWLLSNTDNAAFPQMTKSYKNTRKDPALIIWGIHHSGSTTEQTKLYGSG
NKLITVGSSNYQQSFVPSPGARPQVNGQSGRIDFHWLILNPNDTVTFSFNGAFIALDRASFLRGKSMGIQSEVQVDA
NCEGDCYHSGGTIISNLPFQNINSRAVGKCPRYVKQESLLLATGMKNVPEIPKRRRRGLFGAIAGFIENGWEGLIDG
WYGFRHQNAQGEGTAADYKSTQSAIDQITGKLNRLIEKTNQQFELIDNEFTEVERQIGNVINWTRDSMTEVWSYNAE
LLVAMENQHTIDLADSEMNKLYERVKRQLRENAEEDGTGCFEIFHKCDDDCMASIRNNTYDHSKYREEAIQNRIQID
PVSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/chicken/Germany/n/1949 (H10N7)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDRICLGHHAVANGTIVKTLTNEQEEVTNATETVES
TNLNKLCMKGRSYKDLGNCHPVGMLIGTPVCDPHLTGTWDTLIERENAIAHCYPGATINEEALRQKIMESGGISKMS
TGFTYGSSINSAGTTKACMRNGGDSFYAELKWLVSKTKGQNFPQTTNTYRNTDTAEHLIIWGIHHPSSTQEKNDLYG
TQSLSISVESSTYQNNFVPVVGARPQVNGQSGRIDFHWTLVQPGDNITFSHNGGLIAPSRVSKLTGRGLGIQSEALI
DNSCESKCFWRGGSINTKLPFQNLSPRTVGQCPKYVNQRSLLLATGMRNVPEVVQGRGLFGAIAGFIENGWEGMVDG
WYGFRHQNAQGTGQAADYKSTQAAIDQITGKLNRLIEKTNTEFESIESEFSETEHQIGNVINWTKDSITDIWTYQAE
LLVAMENQHTIDMADSEMLNLYERVRKQLRQNAEEDGKGCFEIYHTCDDSCMESIRNNTYDHSQYREEALLNRLNIN
SVSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/mallard/Astrakhan/263/1982 (H14N5)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGPIICLGHHAVENGTSVKTLTDNHVEVVSAKELVET
NHTDELCPSPLKLVDGQDCDLINGALGSPGCDRLQDTTWDVFIERPTAVDTCYPFDVPDYQSLRSILASSGSLEFIA
EQFTWNGVKVDGSSSACLRGGRNSFFSRLNWLTKATNGNYGPINVTKENTGSYVRLYLWGVHHPSSDNEQTDLYKVA
TGRVTVSTRSDQISIVPNIGSRPRVRNQSGRISIYWTLVNPGDSIIFNSIGNLIAPRGHYKISKSTKSTVLKSDKRI
GSCTSPCLTDKGSIQSDKPFQNVSRIAIGNCPKYVKQGSLMLATGMRNIPGKQAKGLFGAIAGFIENGWQGLIDGWY
GFRHQNAEGTGTAADLKSTQAAIDQINGKLNRLIEKTNEKYHQIEKEFEQVEGRIQDLEKYVEDTKIDLWSYNAELL
VALENQHTIDVTDSEMNKLFERVRRQLRENAEDQGNGCFEIFHQCDNNCIESIRNGTYDHNIYRDEAINNRIKINPV
SGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

>A/shearwater/Western Australia/2576/1979 (H15N9)

MVLVNQSHQGFNKEHTSKMVSAIVLYVLLAAAAHSAFAADPGDKICLGHHAVANGTKVNTLTERGVEVVNATETVEI
TGIDKVCTKGKKAVDLGSCGILGTIIGPPQCDLHLEFKADLIIERRNSSDICYPGRFTNEEALRQIIRESGGIDKES
MGFRYSGIRTDGATSACKRTVSSFYSEMKWLSSSMNNQVFPQLNQTYRNTRKEPALIVWGVHHSSSLDEQNKLYGTG
NKLITVGSSKYQQSFSPSPGARPKVNGQAGRIDFHWMLLDPGDTVTFTFNGAFIAPDRATFLRSNAPSGIEYNGKSL
GIQSDAQIDESCEGECFYSGGTINSPLPFQNIDSRAVGKCPRYVKQSSLPLALGMKNVPEKIRTRGLFGAIAGFIEN
GWEGLIDGWYGFRHQNAQGQGTAADYKSTQAAIDQITGKLNRLIEKTNKQFELIDNEFTEVEQQIGNVINWTRDSLT
EIWSYNAELLVAMENQHTIDLADSEMNKLYERVRRQLRENAEEDGTGCFEIFHRCDDQCMESIRNNTYNHTEYRQEA
LQNRIMINPVSGGGGLNDIFEAQKIEWHERLVPRGSPGSGYIPEAPRDGQAYVRKDGEWVLLSTFLGHHHHHH

Protocol 2.1 - Stability.xml

There are three Rosettascripts used in the document. The first is a script run to assess ***Stability*** of the protein monomer upon complexation. The second is a script to assess change in ***binding free energy*** of the complex using the standard energy function. The third is a script to assess change in ***binding free energy*** of the complex using the standard energy function appended with the modified electrostatics scoring term.

(1) Rosettascript for stability calculation

<dock_design>

<SCOREFXNS>

<local_score weights=score12_full

patch="add_hack_elec.wts_patch"/> # hack_elec = 2.0

<local_score_soft weights=soft_rep patch="add_hack_elec.wts_patch"/>

</SCOREFXNS>

<TASKOPERATIONS>

<InitializeFromCommandline name=init/>

<ProteinInterfaceDesign name=pido allow_all_aas=1 design_all_aas=1 interface_distance_cutoff=999.0/>

<ProteinInterfaceDesign name=nodesign design_chain1=0 design_chain2=0 interface_distance_cutoff=999.0/>

<RestrictAbsentCanonicalAAS name=all keep_aas="ACDEFGHIKLMNPQRSTVWY"/>

</TASKOPERATIONS>

<MOVERS>

<AtomTree name=docking_tree docking_ft=1/> connect chains by their geometric centres. Good for minimization

<MinMover name=min_sc_soft scorefxn=local_score_soft bb=0 chi=1 jump=0> minimize sc, rb

<MoveMap>

<Chain number=2 bb=0 chi=1/>

</MoveMap>

</MinMover>

<MinMover name=min_all_soft scorefxn=local_score_soft chi=1 jump=0> minimize sc, rb, and bb (of chain 2 only)

<MoveMap>

<Chain number=2 bb=1 chi=1/>

</MoveMap>

</MinMover>

<MinMover name=min_sc scorefxn=local_score bb=0 chi=1 jump=0> minimize sc, rb

<MoveMap>

<Chain number=2 bb=0 chi=1/>

</MoveMap>

</MinMover>

<MinMover name=min_all scorefxn=local_score chi=1 jump=1> minimize sc, rb, and bb (of chain 2 only)

<MoveMap>

<Chain number=2 bb=1 chi=1/>

</MoveMap>

</MinMover>

<PackRotamersMover name=pack scorefxn=local_score task_operations=init,nodesign/>

<PackRotamersMover name=pack_soft scorefxn=local_score_soft task_operations=init,nodesign/>

<ParsedProtocol name=relax_before_baseline>

<Add mover=docking_tree/>

<Add mover=pack_soft/>

<Add mover=min_sc_soft/>

<Add mover=pack_soft/>

<Add mover=min_all_soft/>

<Add mover=pack_soft/>

<Add mover=min_sc/>

<Add mover=pack_soft/>

<Add mover=min_all/>

<Add mover=pack/>

```
            <Add mover=min_sc/>
            <Add mover=pack/>
            <Add mover=min_all/>

    </ParsedProtocol>

    </MOVERS>

    <FILTERS>

        <ScoreType name=total_score scorefxn=local_score
score_type=total_score threshold=0.0/>
        <Delta name=delta_score filter=total_score upper=1 lower=0 range=0.2
unbound=1 jump=1 relax_mover=relax_before_baseline/>
        <FilterScan name=scan_stabilizing scorefxn=local_score
relax_mover=relax_before_baseline task_operations=pido,init,all filter=delta_score
resfile_name="stability.resfile"/>
        <Time
name=scan_stabilizing_timer/>

    </FILTERS>

    <PROTOCOLS>

        <Add mover=docking_tree/>
        <Add filter=scan_stabilizing_timer/>
        <Add filter=scan_stabilizing/> fastest goes first (no repacks
here)
        <Add
filter=scan_stabilizing_timer/>

    </PROTOCOLS>
```

</dock_design>

Protocol 2.2 - Binding_energy.xml

(2) Rosettascript for change in binding free energy for the complex using the standard scoring function

<dock_design>

<SCOREFXNS>

<local_score weights=score12_full/>

<local_score_soft weights=soft_rep/>

</SCOREFXNS>

<TASKOPERATIONS>

<InitializeFromCommandline name=init/>

<ProteinInterfaceDesign name=pido allow_all_aas=1 design_all_aas=1 interface_distance_cutoff=999./>

<ProteinInterfaceDesign name=repack_interface design_chain1=0 design_chain2=0/>

<RestrictAbsentCanonicalAAS name=all keep_aas="ACDEFGHIKLMNPQRSTVWY"/>

</TASKOPERATIONS>

<MOVERS>

<AtomTree name=docking_tree docking_ft=1/>

<MinMover name=min_sc_soft scorefxn=local_score_soft bb=0 chi=1 jump=0> minimize sc, rb

<MoveMap>

<Chain number=1 bb=0 chi=1/>

<Chain number=2 bb=0 chi=1/>

</MoveMap>

</MinMover>

<MinMover name=min_all_soft scorefxn=local_score_soft chi=1 jump=0> minimize sc, rb, and bb (of chain 2 only)

<MoveMap>

<Chain number=1 bb=1 chi=1/>

<Chain number=2 bb=1 chi=1/>

</MoveMap>

</MinMover>

<MinMover name=min_sc scorefxn=local_score bb=0 chi=1 jump=0> minimize sc, rb

<MoveMap>

<Chain number=1 bb=0 chi=1/>

<Chain number=2 bb=0 chi=1/>

</MoveMap>

</MinMover>

<MinMover name=min_all scorefxn=local_score chi=1 jump=1> minimize sc, rb, and bb (of chain 2 only)

<MoveMap>

<Chain number=1 bb=1 chi=1/>

<Chain number=2 bb=1 chi=1/>

</MoveMap>

</MinMover>

<PackRotamersMover name=pack_interface scorefxn=local_score task_operations=init,repack_interface />

<PackRotamersMover name=pack_interface_soft scorefxn=local_score_soft task_operations=init,repack_interface/>

<ParsedProtocol name=relax_before_baseline> before running the baseline calculation in filter scan, you need to relax the ppk'ed structure

<Add mover=docking_tree/>

<Add mover=pack_interface_soft/>

<Add mover=min_sc_soft/>

<Add mover=pack_interface_soft/>

<Add mover=min_all_soft/>

<Add mover=pack_interface_soft/>

<Add mover=min_sc/>

<Add mover=pack_interface_soft/>

<Add mover=min_all/>

<Add mover=pack_interface/>

<Add mover=min_sc/>

<Add mover=pack_interface/>

<Add mover=min_all/>

</ParsedProtocol>
</MOVERS>

<FILTERS>

<Ddg name=ddg scorefxn=local_score confidence=0.0/>

<Delta name=delta_ddg filter=ddg upper=1 lower=0 range=-0.5 relax_mover=relax_before_baseline/>

<FilterScan name=scan_binding scorefxn=local_score relax_mover=relax_before_baseline task_operations=pido,init,all filter=delta_ddg triage_filter=delta_ddg resfile_name="binding.resfile"/>

<Time name=scan_binding_timer/>

```
        </FILTERS>

        <PROTOCOLS>

                <Add mover=docking_tree/>
                <Add filter=scan_binding_timer/>
                <Add filter=scan_binding/>
                <Add
filter=scan_binding_timer/>

        </PROTOCOLS>

</dock_design>
```

(3) Rosettascript for change in binding free energy for the complex using the scoring function with the modified electrostatics scoring term.

Protocol 2.3 - Score_electrostatics.xml

```
<dock_design>

        <SCOREFXNS>

                <local_score weights=score12_full patch="add_PB_elec.wts_patch"/>
                <local_score_soft weights=soft_rep patch="add_PB_elec.wts_patch"/>

        </SCOREFXNS>

        <TASKOPERATIONS>

                <InitializeFromCommandline name=init/>
```

<ProteinInterfaceDesign name=pido allow_all_aas=1 design_all_aas=1 interface_distance_cutoff=999./>

<ProteinInterfaceDesign name=repack_interface design_chain1=0 design_chain2=0/>

<RestrictAbsentCanonicalAAS name=all keep_aas="ACDEFGHIKLMNPQRSTVWY"/>

</TASKOPERATIONS>

<MOVERS>

<AtomTree name=docking_tree docking_ft=1/>

<MinMover name=min_sc scorefxn=local_score bb=0 chi=1 jump=0> minimize sc, rb

<MoveMap>

<Chain number=1 bb=0 chi=1/>

<Chain number=2 bb=0 chi=1/>

</MoveMap>

</MinMover>

<PackRotamersMover name=pack_interface scorefxn=local_score task_operations=init,repack_interface />

<PackRotamersMover name=pack_interface_soft scorefxn=local_score_soft task_operations=init,repack_interface/>

<ParsedProtocol name=relax_before_baseline> before running the baseline calculation in filter scan, you need to relax the ppk'ed structure

<Add mover=docking_tree/>

<Add mover=pack_interface/>

<Add mover= min_sc/>

```
        </ParsedProtocol>

        </MOVERS>

        <FILTERS>

                <Ddg name=ddg scorefxn=local_score confidence=0.0/>

                <Delta name=delta_ddg filter=ddg upper=1 lower=0 range=-0.5
relax_mover=relax_before_baseline/>

                <FilterScan name=scan_binding scorefxn=local_score
relax_mover=relax_before_baseline task_operations=pido,init,all filter=delta_ddg
triage_filter=delta_ddg resfile_name="binding.resfile"/>

                <Time
name=scan_binding_timer/>

        </FILTERS>

        <PROTOCOLS>

                <Add mover=docking_tree/>

                <Add filter=scan_binding_timer/>

                <Add filter=scan_binding/>

                <Add
filter=scan_binding_timer/>

        </PROTOCOLS>

</dock_design>
```

Section 3: High-throughput design and testing of *de novo* disulfided influenza binders

Abstract

We show a high-throughput computational and experimental methodology capable of designing and testing tens of thousands of *de novo* protein binders. We used this approach to generate 12,383 disulfided mini-protein binders against a neutralizing epitope found on the stem region of Influenza A Hemagglutinin. Deep mutational scanning was used to structurally validate 17 of these binders based on their computational design models. We also used the sequence function information generated to affinity mature 10 these binders towards either Group I (H1N1) or Group II (H7N9) Influenza subtypes. The best of these variants is a 37 residue triple disulfided miniprotein that binds H1N1 with a K_d of less than 47nM and T_m of > 95°C. Such *de novo* hyperstable small binders may be ideal candidates for diagnostics and therapeutics.

Background and Motivation

In the United States Influenza, combined with primary viral pneumonia, is the 6th leading cause of death[52]. Due to the cumulative morbidity and mortality caused by seasonal influenza in the United States and to the high infection rate approximately 30,000 deaths are attributed to it annually[53]. Furthermore there have been four wide spread pandemics occurring in the 20th century; in 1918(H1N1), 1957(H2N2), 1968(H3N2), and 2009(H1N1). Although the 1918 outbreak was the most deadly with death estimates ranging from 50-100 million[54], the most recent 2009 H1N1 pandemic had a death toll of approximately 300,000 world wide[55,56]. Though vaccination and anti-viral therapeutics such as Oseltamivir[57] greatly reduce the probability of

pandemic losses equaling those seen in 1918, the 2009 pandemic demonstrates that influenza remains serious public health concern.

The most effective countermeasure remains vaccination, however differences between circulating strains and the isolates included in the vaccine is a potential problem area. Antigenic drift is strong enough that in as short as 3 to 5 years predominant strains of Influenza A are replaced by variants that have undergone sufficient mutation to evade existing antibody responses[58]. Isolates included in vaccines must be constantly updated based on predictions from World Health Organization (WHO) monitoring and surveillance. The weakness in this approach is that strains must be selected up to 9 months in advance prior to influenza season to allow time for manufacturing[59]. Interspecies transmission of influenza is also particularly hard to predict as the jump relies on many environmental factors[60].

The current last line of defense is small molecule antiviral therapeutics. Four such drugs are currently licensed for sale in the United States and are either M2 (Amantadine/Rimantadine) or neuroamidase (Zanamivir/Oseltamivir) inhibitors. However, according the Center for Disease Control (CDC), 100% of H3N2 and 2009 pandemic flu have shown resistance to both of the M2 inhibitors and neither are recommended for current treatment within the United States[61]. Of the neuroamidase inhibitors recently reported H7N9 strains in eastern China have been found to be resistant to both[62]. The emergence of antiviral resistance[63] to these small molecules drugs indicate that the continued development of treatment options as combination or standalone therapies will be critical to the maintaining effective therapy in the future.

Lastly, therapeutic neutralizing antibodies have begun to be developed and enter clinical trials, such as CR8020[64] and CR6261[65]. These new class of antibodies generally bind the conserved stem region of HA, though some bind and block the sialic acid binding site[66], and

because of the high conservation in the stem epitope,[67] have the ability bind to broad spectrums of subtypes. Binding to this stem region epitope inhibits the pH induced conformational change the HA uses to fuse the virus and host membrane upon endocytosis. These monoclonal antibodies (mAb) have been shown to have both prophylactic and therapeutic efficacy and were the initial motivation for computational design of *de novo* small protein binders.

mAB influenza therapies are generally intended for already hospitalized elderly, young, or immune-compromised patients with existing intravenous (IV) lines[68]. Infusion treatments are performed in hospital settings, as opposed to small molecule treatments which are available over the counter, making them significantly more expensive. Their production is also expensive as mABs are 150kDa multidomain proteins that contain numerous post-translational modifications requiring complex eukaryotic machinery to produce[69]. Furthermore, mABs can be difficult to concentrate which leads to a high-volume of fluid (~2L) during infusion[70]. Infusion also distributes the mABs throughout the body instead of concentrating them in the lung epithelial tissue where the infection is located, ultimately requiring a higher effective dose[71]. Finally, mABs are either stored refrigerated or frozen and can be easily heat denatured making long term storage and shipping expensive[72]. However, mABs remain a viable therapeutic option as they generally have excellent PK characteristics, can signal a secondary immune response, are degraded by well understood catabolic pathways[73], and when humanized generally show low immunogenicity[74].

We have previously reported on the generation of two small *de novo* binders against several Group I HA subtypes[10], specifically SC1918/H1(H1) and VN2004/H5 (H5). The two binders, known as HB80 and HB36 were generated using a hot-spot based design methodology and after affinity maturation using SSM libraries combined with deep mutational scanning bound

with low nanomolar affinity (~1nM Kd)[2]. The first of these binders, HB36, is a four helix bundle that was originally a transcription factor from the thermophilic bacteria *Bacillus stearothermophilus* (**PDB 1U84**[2]) with a MW of ~10 kDa. The second binder, HB80.4, is a truncated three helix bundle that was also originally transcription factor from *Antirrhinum Majus (***PDB 4EEF**[1]) with a MW of ~8 kDa. While a variant of the first of these binders, HB36.6, is beginning to show therapeutic efficacy in animals (see **Introduction**) neither binder has shown binding to Group II influenza virus strains *in vitro.* Furthermore, both binders express at moderate levels and offer relatively low Tm as determined by CD (see **Figure 4.2 and 4.3**) as well as being protease susceptible. Though generally considered small, neither protein is small enough to be easily synthesizable (<40AA).

***De Novo* Scaffold Proteins**

The ideal designed protein therapeutic should combine aspects of small molecules drugs; chemically synthesizable, long term storage without refrigeration, and hyperstable with monoclonal antibodies; high affinity, well understood degradation pathways, easily modified and designable. Until recently[75] the design of new protein binders relied on the re-engineering of existing proteins, whose structures were solved and published in the protein data bank. These naturally occurring scaffolds are the end product of evolutionary pressure and as such are optimized only for their particular functional niche, not for the constraints of protein therapeutics. However, there exists *de novo* designed proteins[76], created wholly within the framework of Rosetta[77], which can be engineered from the start with a particular function in mind. We aim to create the next generation of designed protein binders utilizing small *de novo* designed proteins as starting scaffolds to overcome some of the limitations associated with natural scaffolds.

Cystine-rich miniproteins, often known as knottins, are small (<40AA) structurally similar proteins that exhibit hyperthermal and proteolytic stability as result of multiple intramolecular disulfides[78]. A common structural trait is an embedded disulfide ring penetrated by another disulfide bond, creating what is known as "cysteine knot." These recently discovered motifs are found in a large variety of organisms but are most commonly in the form of toxins and growth factors. They are known to be hyperstable with the ability to resist degradation by serum and membrane bound proteases[79]. Their stability is a result of multiple disulfides creating a overlapping and interconnected covalent bond network in the core of the protein. This stability translates into a long serum half life and low immunogenicity[80]. Furthermore, some knottins can pass the blood brain barrier and gut mucosal[81]. They are also beginning to be engineered for molecular recognition as peptide based alternatives to mAB[82,83]. Yeast display libraries of knottins have resulted in high affinity binders to cancer targets and used as the basis of molecular imaging agents[82]. They are also well within the chemical synthesis range and have been shown to be readily manufacturable. A naturally occurring knottin from a tropical cone snail was FDA approved in 2004 for pain reduction medication[84]. All of these characteristics make them excellent scaffolds for drug discovery. One of the only disadvantages is the relative scarcity of knottins in the PDB that are appropriate for use in computational pipelines. However, their small size, defined folds[85], and known connectivity make them amenable to *de novo* design in the framework of Rosetta.

Computational protein engineering allows for the design of any buildable protein topology, which in the context of the Rosetta Computational Suite means capable of being assembled from existing 3 and 9-mer protein fragments mined from known protein structures[86]. Topology is defined as a specific ordering of secondary structure elements. The nomenclature

EHEE defines a particularly protein where the secondary structure elements are N to C terminal edge(E-sheet) – helix (H) –edge(E)-edge(E). Topological variant includes the length of the secondary structure elements in the definition. For EHEE an example of a unique 35 residue topological variant is; 5E.3L.11H.3L.5E.3L.5E which defines all the edge components as being 5 residues in length, the connecting loops (implied before) all being 3 residues in length, and single 11 residue helix. Our initial disulfided miniprotein scaffold generation scheme began with a single topological variant based on that of *Orthochirus scrobiculosus* (scorpion) toxin[87], as we knew *a priori* it was a buildable topology. Only the secondary structure lengths and connectivity were used from the original NMR solved structure (**PDB ID: 1SCO**)[88], no backbone or side chains constraints were passed to Rosetta. This topological variant was chosen as it was formed almost entirely from defined secondary structure with minimal loop connectivity. The topology can also accommodate a 13 residue helix, which was a our pre-defined core interface residue to be docked/grafted. A further explanation of scaffold topologies can be found in **Figure 4.1**.

Oligo pools

Our experimental protein screening/selection technology primarily uses yeast display and flow cytometery to pan for binding against specific antigens. Potential protein binders coding sequences open reading frame (ORF) are reverse translated and manufactured using column-based oligo synthesis methods. The ORF gene inserts are combined into plasmids either through *in vitro* assembly[89] or homologous recombination[90] using yeast. Using this method single proteins are synthesized as complete genes separately and tested individually. This paradigm ultimately limits the ability to test larger numbers of computational hypotheses experimentally. We are limited to a small number of design models that can be tested (<100) as both the cost of synthesis and work required to test designs is not easily scalable. As demonstrated in **Section 1**;

our ability to generate protein libraries and measure outcomes has increased in scale owing to NGS. Applying this type of massively parallel screening and scanning to initial binder screening, and not just affinity maturation, requires the ordering of tens of thousands of explicit genes which until recently has not been possible. A number of recent advances in DNA synthesis, specifically the advent of microarray based oligo pools, now makes ordering thousands of explicit genes possible.

Microarray based oligo pools are *de novo* DNA synthesis methods that utilize spatially located polymer synthesis on functionalized surfaces originally developed for microarray applications[91]. In oligo pools, thousands of short oligos are synthesized in parallel as a microarray then cleaved from the surface and collected into a single tube[92]. Oligos manufactured in this way are 2-4 orders of magnitude cheaper than traditional column-based methods[93]. However, even though microarray based oligo pools are significantly less expensive, there are several complications in using them for gene synthesis of protein binders. First, oligo pools have significantly higher error rates than column synthesized DNA. The error rates are in the range of ~1-1.5% which translates into a predicted final yield of 23-10% for a 150bp synthesis (**Figure 4.7)**. Errors are most commonly deletions and truncations (85%) which reduces the yields of the long-length oligos and limiting the overall length to less than 200bp[94]. Second, the chemistry used in the cleavage step, where the oligos are released from the microarray surface, leaves a 3' terminal phosphate, which most polymerases are not capable of extending from. This means oligo pools are often difficult to use directly as primers or in gene assembly reactions without enzymatic (phosphatase) removal of the phosphate[95]. Finally, as oligo pools are cleaved and provided in parallel they represent highly heterogeneous mixture of oligos in sequence, concentration, and length. These complex mixtures are difficult to PCR amplify uniformly

without errors due cross hybridizations and PCR bias[96]. Common sequences within the pools exacerbate problems where parallel gene assembly from oligos requires orthogonal sequences between genes. Concentration differences between likely annealing partners, found to be as high as 1E3, combined with the large numbers of oligos produced make the probability of cognate partner interactions low, making PCR difficult[97]. It should be noted that oligo pools are relatively new and only a smaller number of manufactures exist than can supply them, most prominent of those are CustomArray (CustomArray, Bothell, WA) and Agilent (Agilent Technologies, Santa Clara, CA). However, we will show that by using each oligo in the pool as a standalone gene template many of the problems with gene assembly can be avoided. This combined with PCR purification and NGS will allow us track and ultimately normalize for the error inherent in oligo pool synthesis.

Computational Design Strategy

One of the primaries goals was the creation of a highly automated design protocol that did not depend on *a priori* hotspot placements mined from known antibody/antigen structures or manual pruning of designs. Instead we choose to focus on the general area targeted by the stem binding neutralizing antibodies and identify helix orientations that contained the most available residue-residue interactions in an automated fashion. This was accomplished by adapting the recently published MotifDock protocol[98,99] for symmetric design to asymmetric (one-sided) design. A disembodied helix with no side chain information was docked against the target epitope, defined only as a general area and kept fixed, and multiple rigid body orientations scored and ranked by the designability of that orientation. Our strategy thus centered on the design of interfaces built around core helical interface fragments. A helix was chosen as we considered it a feature common to successful Rosetta designed protein complexes. It differs

from “hot spot” based design as no side chain information was used from antibody interactions, only that the general epitope was suitable.

Following helix placement, amino acid side chains were added in energetically favorable conformations based on residue suggestions from the MotifDock protocol in combination with Rosetta. Residues were further packed and minimized to optimize for shape complementarity (SC), delta delta G of binding ($\Delta\Delta G$) and total solvent-exposed surface area (SASA). The MotifDock and add side chains protocol were iterated over to create large numbers unique helical interface fragments suitable for scaffold matching. One advantage of docking and building off of the same initial helical backbones was that comparisons between unique build trajectories were much easier to perform. Also, as only the position and sequence of the helix was different in each fragment, and not the backbone, scaffolds could be easily built to accommodate all the fragments universally.

Once the interface fragments were generated a protocol was established to identify the most suitable scaffold topologies to graft the fragments into. Scaffolds were super-positioned into place, clash checked, and peripheral interface designed. The number of completed designs was dependent on the total number of scaffolds built. As this is an easily scalable computational process, simply generating additional scaffolds of a defined topology could easily generate more binders. Our aim by keeping interface fragment generation and scaffold building separate was to design more modularity into the design process. As more advanced scaffold building methods are introduced they can be incorporated into the overall protein protein interface design (PPID) pipeline in a streamlined fashion.

Design of HA-Binding Proteins

We choose to target the stem region of multiple (H1/H3/H7) HAs targeted by neutralizing antibodies consisting of a hydrophobic groove and a component of the fusion peptide loop. We defined the center of the docking region as a conserved Trp21[100] that is present in all three crystal structures used (**H1 PDB ID: 4EEF[2], H3: 3ZTJ[101], H7: 4FQV[102]**), the radius of the docking circle extends ~20Å from this Trp. Nearly all solved crystal structures of stem region neutralizing antibodies contain an energetically favorable aromatic hotspot interaction with this Trp21, as does our two previously published binders.

An idealized 13 residue poly alanine helix was docked into the defined stem region epitope and sampled by systemically varying the translational and rotational degrees of freedom in increments of 1Å and 1 degree, respectively. Each orientation was clash checked and those that passed were scored based on a measure of designability. The "designability" metric was the same as previously described[98,99] and incorporates the number of unique residue-residue interactions as a primary component. We carried forward the top 100 designable helical orientations from each of the 3 targets to side chain adding.

In the second step, 100 attempts were made to add amino acid side chains to each of the 300 helical orientations from step one, in context of their respective targets. Each of these 100 trajectories began with a random fine grained perturbation of the rigid body orientation followed by three rounds of residue design, repacking, and minimization using RosettaDesign (**Protocol 4.3**). Once the final sequence of helical fragment had been designed, interface metrics (ΔΔG/SASA/SC) were used to score the designed interface. It should be noted that a defining trait of the H3 and H7 stem epitopes is a glycosylation at position Asn23 that is absent in all Group I HAs. The glyosylation, which was initially removed before the MotifDock step, was

modeled back in during the add side chain process. This unique Group II feature greatly reduces the number of models that passed out of the H3/H7 trajectories versus H1 as it often clashed with the backbone interface helix (see **Figure 4.4**). Furthermore it restricted the real estate available at the epitope and as such lowered the number of diverse helical orientations possible. Of the 10,000 design trajectories for each target; 1000 interface fragments passed filters for H1, 300 for H3, and 250 H7. A combination of root-mean-square deviation (RMSD) clustering and manual inspection resulted in 29 H1, 31 H3, and 29 H7 helical interface fragments.

In the third step, *de novo* disulfide miniprotein scaffold sets were built. We defined an initial EHEE topology based on the NMR structure of a 37 residue known scorpion toxin as described earlier. This topology was used to generate multiple independent poly-valine backbones using Rosetta folding simulations (**Protocol 4.1**). Though the population of resulting backbones was all of the same topological variant there was structural diversity in the angles between the secondary structure elements and loop regions. Side chains were introduced in a Monte-Carlo fashion using fixed-backbone Rosetta design followed by relaxation using Rosetta all-atom energy function. Multiple disulfides were added using RosettaRemodel[103] fast design protocol (**Protocol 4.2**). Sequence design and refinement were iterated and resulting scaffolds scored based on a number of metrics including total Rosetta energy (REU) and packing (RosettaHoles). For the initial EHEE topology variant 142 scaffolds were generated. We chose to further refine and diversify our scaffold topologies by varying the secondary structure lengths and overall topology (EHEE and EEHE). We attempted to build 80 topological variants and computationally screened them for interface fragment compatibility. Ultimately, 5 unique topological variants were used to create a heterogeneous scaffold set with a population of 192 (see **Figure 4.5**). A third homogeneous scaffold set was also built using different filtering

metrics. An advantage of the large number of designs experimentally tested was it allowed for the comparison different scaffold generation protocols.

The final step was merging the generated helical interface fragments with the multiple scaffold sets to design the complete binding proteins. This was accomplished using MotifGraft, an application in Rosetta (**Protocol 4.4**). MotifGraft allows the superposition of the scaffold (target) onto the helical interface fragment (motif) in the context of the HA target (context) similar to previously published EpiGraft[104]. The combined scaffold sets consisted of 640 unique *de novo* disulfided miniproteins and was matched against the 105 *de novo* interface fragments. Once the scaffold was placed and clash checked, residues within 15Å of the interface on the scaffold were designed, repacked, and minimized a final time. The final design trajectories were repeated in a Monte-Carlo fashion such that multiple sequence unique designs models were generated from each fragment scaffold pairing. Also included was an interface fragment positive control, which was the HB80.4 binding helix in context of H1, which was grafted against the combined scaffold set. Final designs were filtered and ranked and 12,383 designs were ordered; 2743 H1 designs, 6521 H3 designs, and 3109 H7 designs.

Experimental Screening of 12,383 Designs

12,472 oligos were ordered from CustomArray as an oligo pool. The oligos ranged in size from 147bp to 153bp and represented the 12,383 total designs (see **Online Repository**) with some randomized repeats. Each oligo contained the ORF of the binder flanked by 18bp priming regions that doubled as homologous recombination sites designed for compatibility with linearized yeast display vector pETCON (see **Gene Synthesis** in **Methods**). The designs underwent and initial round of PCR, were agarose gel extracted, followed by a second round of amplification. The method was optimized after multiple polymerases, PCR conditions, and gel

purifications conditions were tested (**Protocol for Amplification from CustomArray Pools** in **Methods**). The pool of PCR purified designs and linearized pETCON was transformed into EBY100 yeast using a high efficiency method for homologous recombination[26]. After transformation cells were grown overnight in SDCAA media in 30ml cultures and at 30°C, passaged once, and combined 1:1 with 40% glycerol in 1e8 aliquots and stored at -80°C

Cell aliquots were thawed on ice and 1e7 cells were used to inoculate 10ml cultures of SDCAA media and grown for 24hrs at 30°C. The cultures were passaged into 10ml of SGCAA and induced for 24hrs at 30°C . The designed proteins displayed on the cell surface between the yeast surface protein Aga2p and a C-terminal c-Myc tag as previously described[105]. The displaying cells were incubated with 1uM of either biotinylated SC1918/H1 [A/South Carolina/1/1918 (H1N1)], HK68/H3 [A/Hong Kong/1/1968(H3N2)], or Netherlands/H7 [A/Netherlands/219/2003(H7N7)] with avidity provided by phycoerythrin-conjugated streptavidin (Invitrogen), washed and labeled with fluorescein-conjugated antibody against c-Myc (Miltenyi Biotech). Using fluorescence-activated cell sorting, we collected the population of cells that bound to their respective targets (see **Figure 4.8).** We performed a second round of sorting after recovery under the same conditions as the first. We also included a second round competition sort where 10uM soluble FI6v3 scFv was pre-incubated with the biontinylated HAs. The competition sort was a control for non-specific binding, as the scFv should obstruct the stem epitope on the HA, so any cells collected were most likely not binding in the designed mode.

From each selected population, plasmid DNA was extracted and genes PCR amplified followed by sequencing using Illumina Miseq V2 PE-300 (Illumina, San Diego, CA) paired-end deep sequencing as described in **Section 1**.

Analysis of the unselected rd0 yeast pool showed near complete sequence coverage; 11,910 of the 12,383 designs (~96%) ordered were visible with at least 1 count and 0 DNA errors. Each selected population (rd1-rd2, comp rd2) was sampled with a median depth of coverage of 30 reads per variant with little sequencing error. The median number of DNA reads in the selected populations was 45 and minimum was 1. However, only 3.2% of the unselected rd0 pool contained full length designed DNA sequence with no errors (design mismatches). The percentage of correct full length reads increased dramatically in the selected populations indicating that we were enriching for full length protein as opposed to fragments (see **Table 4.1**).

The enrichment ratio of the frequencies of each design in the selected versus unselected populations was used to determine the overall fitness of each design (see **Table 4.2)**. We searched for designs that were enriched in both rd1 and rd2, with a minimum of 100 counts, and were competed away by the FI6v3 antibody scFv. We identified 18 H1 designs that met these criteria, as well as 15 H7 and 2 H3 designs. A subset of these designs were cloned individually and tested monoclonally for binding and scFv competition. Approximately ~85% of these designs identified through deep sequencing showed strong binding at 1uM and competition. Unfortunately, of the 2 H3 designs tested neither bound strongly to H3HA monoclonally. All successfully H1 and H7 binding designs consisted were of a sheet helix topology with a helix at the core of the interface (see **Figure 3.3**).

SSM Affinity Maturation and Structural Validation

In order to quickly validate the binding mode and find beneficial position specific mutations we chose to create and test site saturation mutagenesis (SSM) libraries of 17 designs in parallel (see **Table 4.3**). The designs were tested based on their overall enrichment score combined with visual inspection to select diverse designs. We investigated the contributions to

binding of all single amino acid substitutions at all positions of 5 H1 binding proteins, 2 H3 binding proteins (though they didn't show binding monoclonally), and 10 H7 binding proteins (see **Table 4.4**). The complete set of amino acid variants, 12,045 sequences in total including parent sequences, was ordered as a CustomArray oligo pool with the same parameters as the initial screen (see **Online Repository**).

The SSM pool was amplified and transformed as previously described. We choose to perform 3 independent selection trajectories of increasing stringency for each of the 3 target HAs beginning from the original pool. A "loose" stringency first round selection using 1uM biontinylated HA was followed by increasing stringency 160nM rd2 and 40nM rd3 rounds. Unfortunately, no binding was seen during the H3 sorts so no cells were collected. Plasmid DNA from each round of selection were extracted, amplified, and sequenced per above.

Analysis of the unselected rd0 yeast pool showed 85% coverage of designs ordered. Interestingly the 15% of absent sequences were not evenly distributed throughout but consisted primarily of two designs with very low counts. Each selected population (rd1-rd3) was sampled with a median depth of coverage 35 reads per variant with little sequencing error. The log base 2 ratio of frequencies of a single substitution variant in the selected population versus unselected (rd0) was calculated and used to score the variants. As all of the SSM variants were tested in parallel individual variants with the highest affinities took over the population by rd3. We then used this sequence-function information to validate our structural design models.

Maps of the enrichment values for H1/H7 of the 17 SSM libraries, each containing ~740 single amino-acid substitutions, were generated (see **Online Repository**). As the enrichment values for rd1 were based on selections where antigen concentrations were well above (10X) the expected K_d of the binders, only mutations that were highly deleterious to binding are shown as

depleted in the data. Positions that showed enrichment indicated only that those positions could accommodate variation without abrogating binding completely. To validate our design models (see **Figure 3.4 A-B**) we examined whether key interface and core residues could tolerate sequence variation by mapping structurally this enrichment or depletion. We were able to validate 10 of our design models (5-H1/5-H7) as having highly conserved (>90% conservation) interface residues. Interestingly all of these validated models also showed high conservation of their cysteine's, indicating that their disulfides were most likely formed and contributing substantially to the binding of the protein.

We used the later rounds of selection to identify combinations of substitutions that would yield the highest affinity binders. As the later rounds of selections had antigen concentrations below the expected K_d of the binders, only mutations that were highly advantageous are shown as enriched. The pattern of this data is quite different from rd1 data; the beneficial mutations are scattered throughout the protein but often located at the ends of secondary structure components (see **Figure 3.4 C-D**).

Beneficial rd3 mutations were enumeratively combined and brute force scored (log2 enrichment product) and ranked. The top ~10% of ranked mutation combinations for each validated binder (1300 total unique variants) were ordered as a subset of a CustomArray oligo pool with the same parameters as the initial screen (see **Online Repository**). These were amplified and transformed as described above. This final high affinity combination library underwent 3 rounds (100/10/1nM) of increasing stringency selection against both H1 and H7. Sequencing of these variants is currently pending.

Expression and *in vitro* Binding

Two of the H1HA binding disulfided miniproteins were expressed in *Escherichia Coli* with an N-terminal SUMO tag and purified by affinity chromatography[106–108]. The binding affinities for hemagglutinin of two of the variants were determined using biolayer interferometry, as previously described[2]. The highest affinity variant tested HB1.1, which had not yet been optimized via SSM or affinity maturation, binds SC1918/H1 Hemagglutitin with a binding dissociation constant of K_d of 47nM (**Figure 4.10**) and a T_m of >95°C (**Figure 4.9**) as determined by biolayer interferometry and circular dichroism respectively. This best variant differed from its 37 residue EHEE *de novo* parent scaffold by 8 mutations.

Discussion

We show an approach for high-throughput design and testing of computationally designed binders on an unprecedented scale. Our approach of using *de novo* disulfided miniproteins as scaffolds allows an entire hyperstable binding protein to be encoded on a relatively short piece of DNA (<150 bp). The small encoding region permits the use of oligo pools, which allowed tens of thousands of designs to be ordered and tested in parallel. Furthermore, we applied *de novo* interface methods in an automated fashion, rather than using antibody hotspots, and generated designs with minimal human inspection. This combination of both experimental and technical advances greatly increases our ability to both generate hypotheses and measure outcomes. Specifically, the near-exclusion of human "manual refinement" from the design pipeline should allow for better mapping of computational parameters to functional outputs in the future.

Our SSM and deep sequencing results show the potential for utilizing sequence-function information in structural validation. Though deep sequencing of populations has been used to

determine the fitness landscapes of known binders[109], it has not been utilized to accept or reject computational design models for which structures have not yet been determined. The landscapes generated are also the first comprehensive view of the stability and binding of disulfided miniproteins in sequence space. The high conservation of cysteines was surprising in that successful binding was dependent on all predicted disulfides forming and could be disrupted when even a single partner failed to form a covalent linkage. Furthermore, by testing 17 SSMs in parallel we show how multiplexing was feasible with NGS and can vastly increase the quantity of information gathered from a single yeast display experiment.

We were also able to rapidly obtain large increases in binding affinity, which has been shown before, but not with computationally designed proteins of this size. By not using degenerate codon based libraries, but by explicitly specifying variants and ordered in oligo pools, we greatly reduced the total number of designs that needed to be screened. The specific combination of mutations contained within the best variants would be very difficult to happen upon by conventional affinity maturation approaches such as PCR mutagenesis. The results of multiple low affinity(<100nM) binders are a further illustration of how a combination of multiple low affinity peripheral substitutions can have large effects on affinity.

Though the creation of a broadly cross-reactive small protein therapeutic remains the ultimate goal, binding within the Group II clade of Influenza A is an important step in that direction. We show here multiple helical interface based solutions to the stem region epitope of the H7 subtype. Binding and potential neutralization of H7 HA is a significant milestone as human cross-over from highly pathogenic Oseltamivir resistant H7N9[63,110] circulating in avian reservoirs remains a real threat. By targeting the same epitope as Group II neutralizing antibodies FI6v3 and CR6261 we believe the 5 H7 binding proteins described here are strong

candidates for therapeutic applications. We have previously shown with Group I binders that helical based high-affinity interactions, as opposed to antibody CDR loops, are capable of providing high levels of HA neutralization.

Methods

Gene synthesis

The amino acid sequence for each design was reverse translated and codon optimized for s. *cerevisiae* using the DNAworks2.0 software suite (Helix Systems)**.** A 5' and 3' adaptor was added (listed below 5' to 3'), which was kept as short as possible to reduce the overall synthesis of the oligo. These adaptors are only suitable for pETCON3 transformation. pETCON3 is a variant of the normally used pETCON2 with modifications to homologous recombination overlap regions for optimized primer annealing.

front adaptor – gene – rev adaptor

front_adaptor

>GGGTCGGCTTCGCATATG

rev_adaptor

>CTCGAGGGTGGAGGTTCC

Protocol for Amplification from CustomArray oligopools

In summary this protocol starts with a 1st round qPCR to determine raw pool amplification conditions, then 1st round amplification PCR, agarose gel extraction, followed by a 2nd round qPCR on extracted products, and then finally 2nd large scale amplification. Over PCRing was a significant problem with these oligo pools as they are heterogeneous in nature and too many PCR cycles results in a large portion of the amplified DNA being ssDNA.

	20uL KAPA HIFI reaction			Cycling conditions	
	HF	10X		time	temp
5X buffer	4	40	step 1	3:00	95°C
10mM DNTPS	0.4	4	step 2	:20	98°C
Fwd Primer (10uM)	1	10	step 3	:15	68°C
Rev Primer (10uM)	1	10	step 4	:15	72°C
template DNA	1	10	step 5	GOTO 2 33X	
10X SYBRsafe dye*	2	20	step 6	5:00	72°C
KAPA Pol**	0.4	4			
H20	10.2	102			

NGS_fwd_extend_adaptor (Fwd Primer)

>AGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTTCGCATATG

NGS_rev_extend _adaptor (Rev Primer)

>GATCTCTATTACAAGTCCTCTTCAGAAATAAGCTTTTGTTCGGAACCTCCACCCTCGAG

*the primers will add ~84bp to overall length of gene. After PCR the length of PCR products was expected to be 231bp as the ordered oligos had a complete length of 147bp.

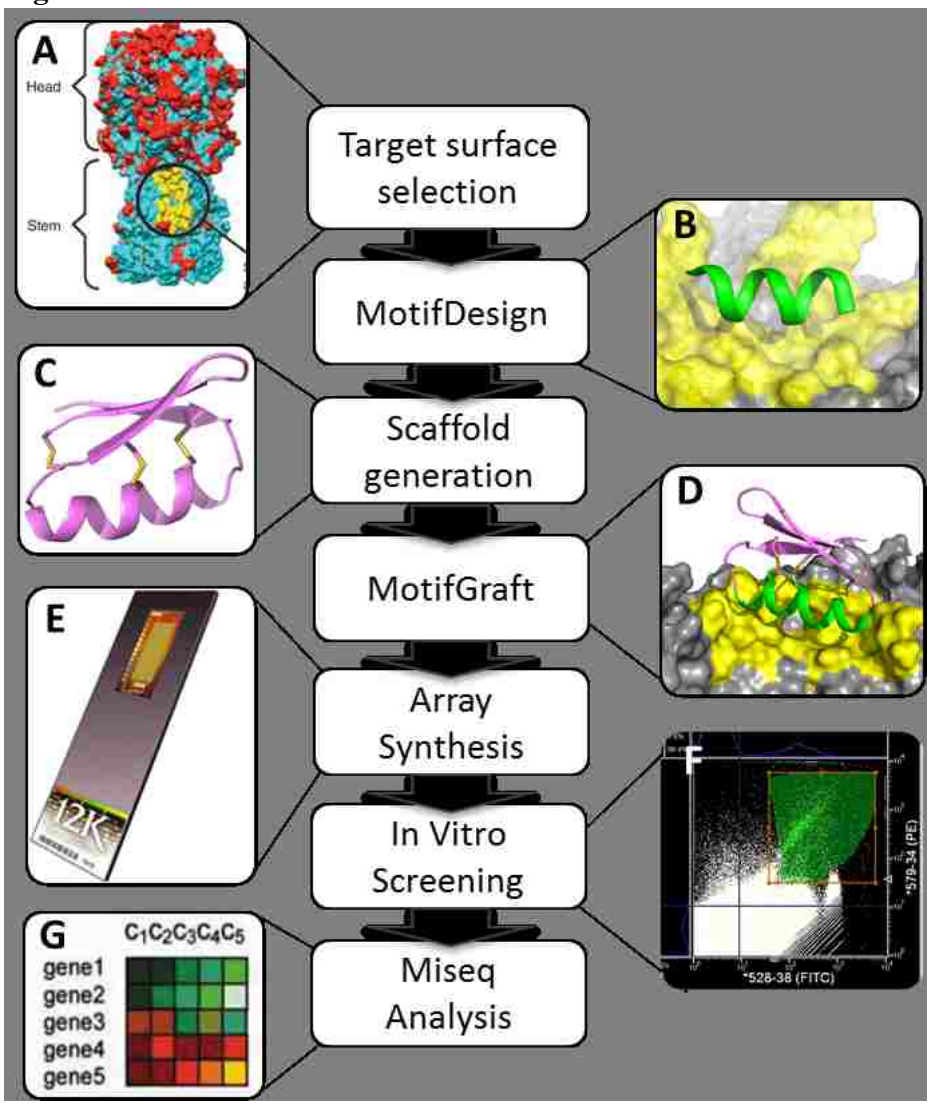
Initial test amplification was 33 cycles of qPCR using BioRad CFX96 thermocycler to determine optimal amplification cycles. The qPCR data was analyzed to setup the 1st round amplification PCR that was run for 22 cycles (before saturation) using the same conditions as above. All of the PCR reaction was run on a large 2.5% agarose gel for imaging and gel extraction. 1µl of the gel extracted product was run in the qPCR and cycled under the same conditions first qPCR. A final large scale PCR reaction was performed (10 X 20ul) using conditions determined second qPCR. The PCRs were purified using two columns from a Qiagen PCR cleanup kit resulting in ~200ng/µl in 80 µl or 16000ng of insert. 1µl of PCR product was run on a 2.5% agarose gel for quality control.

Circular Dichroism

An Aviv 62A DS spectrometer was used to collect all circular dichroism data. Far-ultraviolet circular dichroism spectra of designed proteins were measured at temperatures of 25 and 95 °C from 260 to 200 nm for 10–25 μM protein samples in HBS buffer (pH 7.4) in a 1-mm-path-length cuvette. The protein concentrations were determined either from the absorbance at 280nM[111] using an ultraviolet spectrophotometer (NanoDrop, Thermo Scientfic) or by fluorometric quantitation (Qubit, Thermo Fisher)[112]. T_m is the temperature where the number of folded proteins equals he number of unfolded proteins.

Figures

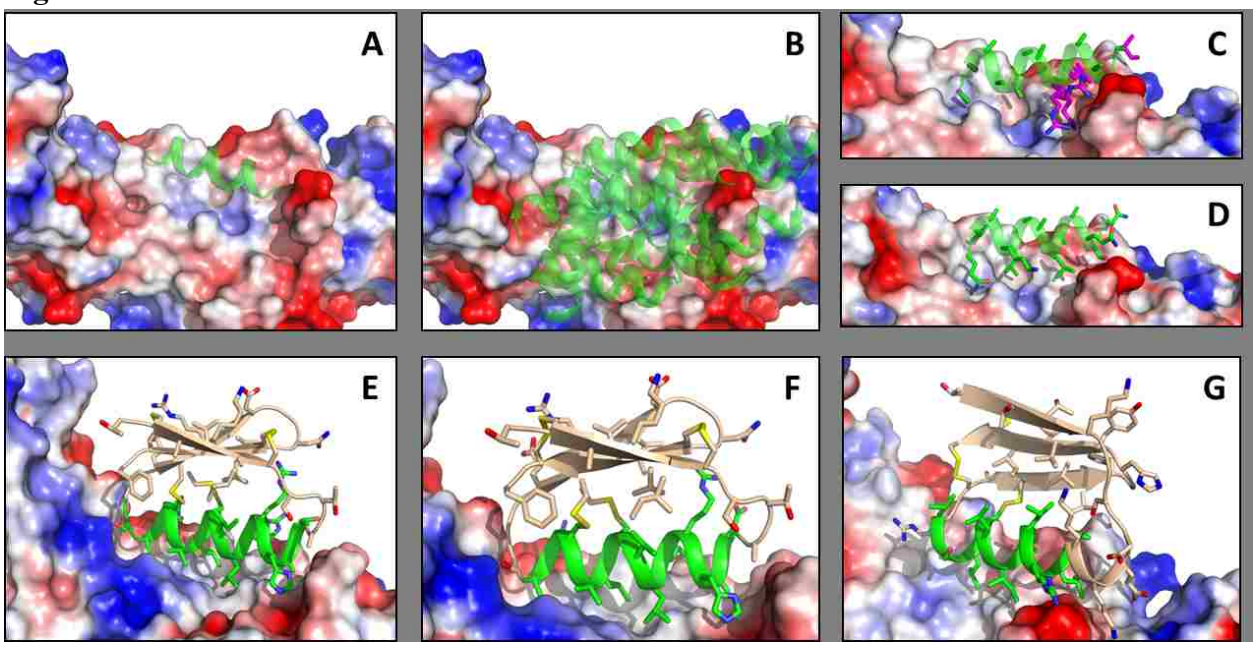
Figure 3.1



Flow chart illustrating the key steps in the high-throughput design of *de novo* binding proteins. The first step (A) was defining the binding epitope as a general physical area to be targeted during docking; this could potentially be any target with structural information. The MotifDesign protocol (B) was then used to dock an ideal poly-alanine helix into the defined target area and score different rigid body orientations on their potential residue-residue interactions. Hundreds of *de novo* disulfided miniproteins (C) were built *in silico* that can

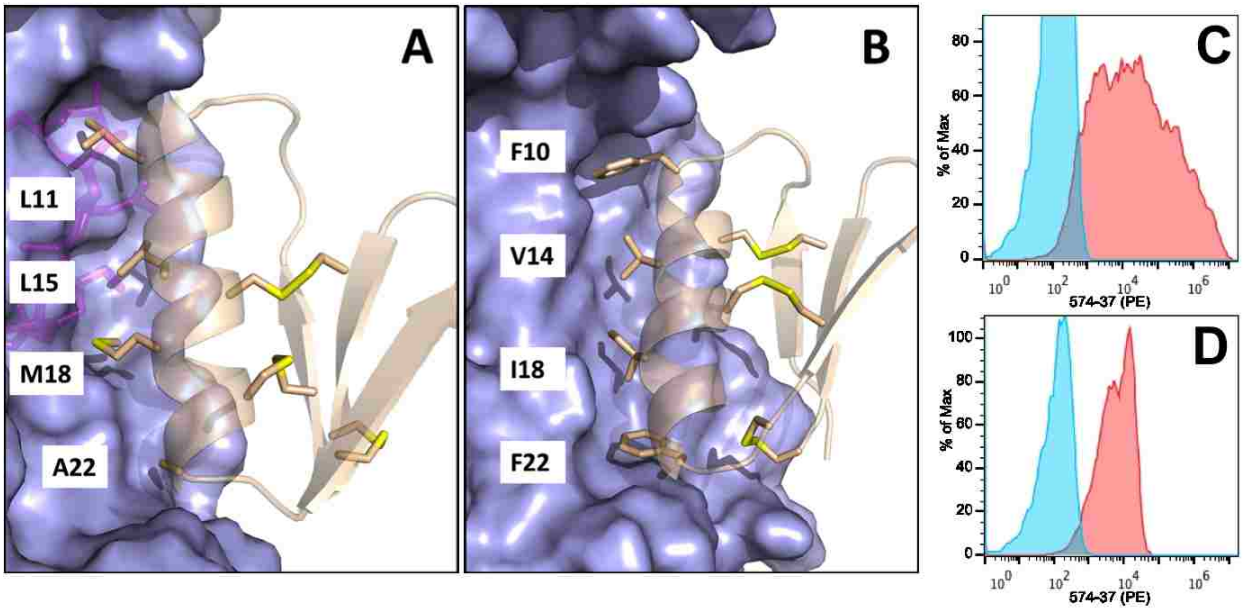
accommodate a helix of the correct length. The docked helices, with side chain interactions added, were grafted (D) onto the *de novo* scaffolds and further designed in context of the target structure. 12,383 unique HA binder genes were manufactured as DNA using array synthesis (E) oligo pool technology. The oligos were amplified, purified, and combined into yeast display vectors followed by multiple rounds of selection (F) to enrich for proteins that bound to various HAs. Finally, plasmid DNA was extracted from pools of yeast from each round of selection and analyzed using deep-sequencing (G) to identify the sequences of the enriched binders.

Figure 3.2



Panels showing specifics of the computational design pipeline. Step 1 (A,B) was the rigid body docking of a poly alanine helix (green) against the stem epitope region of HA using MotifDesign, colored by charge positive blue/negative red. Step 2 (C,D) was adding side chains to the docked helixes using amino acid suggestions (magenta) from MotifDesign that were based on existing residue-residue interactions found in the PDB. The final step (E-G) was using MotifGraft, a RosettaScripts[11] version of Epigraft[104], to superposition the helices onto various *de novo* scaffolds (beige), clash check, and design the peripheral interface.

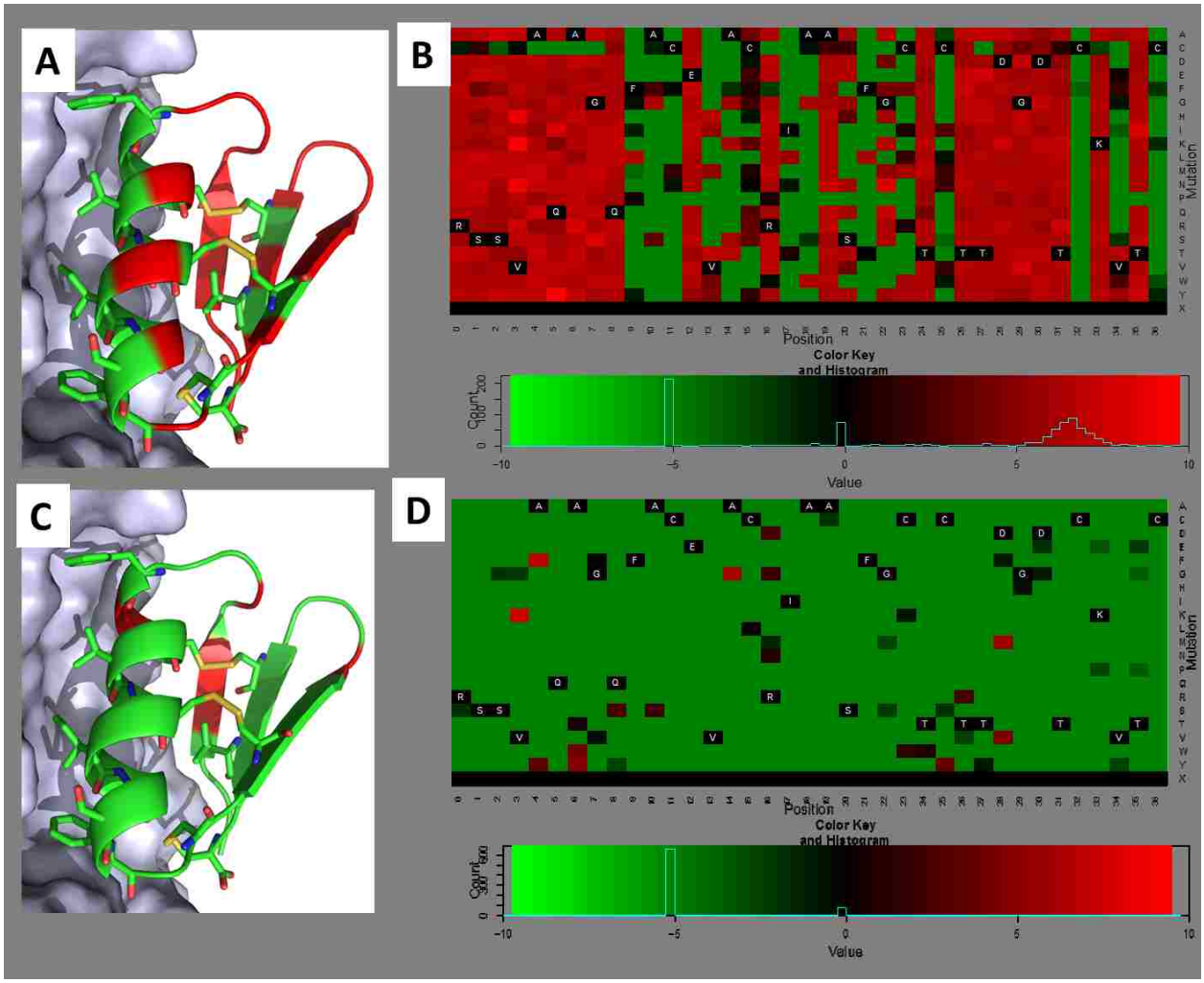
Figure 3.3



Panels showing two binders identified during the initial screen of 12,383 designs. Key interface residues and disulfides are shown as sticks. The HA (light blue) is oriented with the head region towards the top. Panel A shows the design model of 37 residue binder HB7.11 bound to the stem region of H7HA (A/Netherlands/219/2003). Ala22 makes interactions with HA target residue Trp21 (described in main text). This Trp residue is conserved in all Influenza HAs but the Group IIs use a rotamer that creates a shallow hydrophobic surface cavity resulting in the use of a smaller aliphatic opposite of this position. The glycosylation position on H7 (Asn28) near the epitope is shown in purple. Panel B shows the design model of 37 residue binder HB1.1 bound to the stem region of H1HA (A/South Carolina/1/1918). This design places Phe22 across from the conserved Trp, which because of the rotamer used can accommodate a larger aromatic residue due to the deeper pocket. Both designs use hydrophobic residues to build the remainder of the interface. It should be noted that the composition and placement of these key interface residues were not mined from bound antibody complexes, though they strongly overlap hotspots found on stem binding antibodies, but discovered through a computational *de*

novo search of the surface. Initial yeast binding data for HB7.11 (C), before affinity maturation, shows an increase in fluorescent phycoerythirin (PE) signal when incubated with 1280nM H7HA (red with HA, blue without). HB1.1 also shows (D) increased signal when incubated with 1280nM H1HA.

Figure 3.4

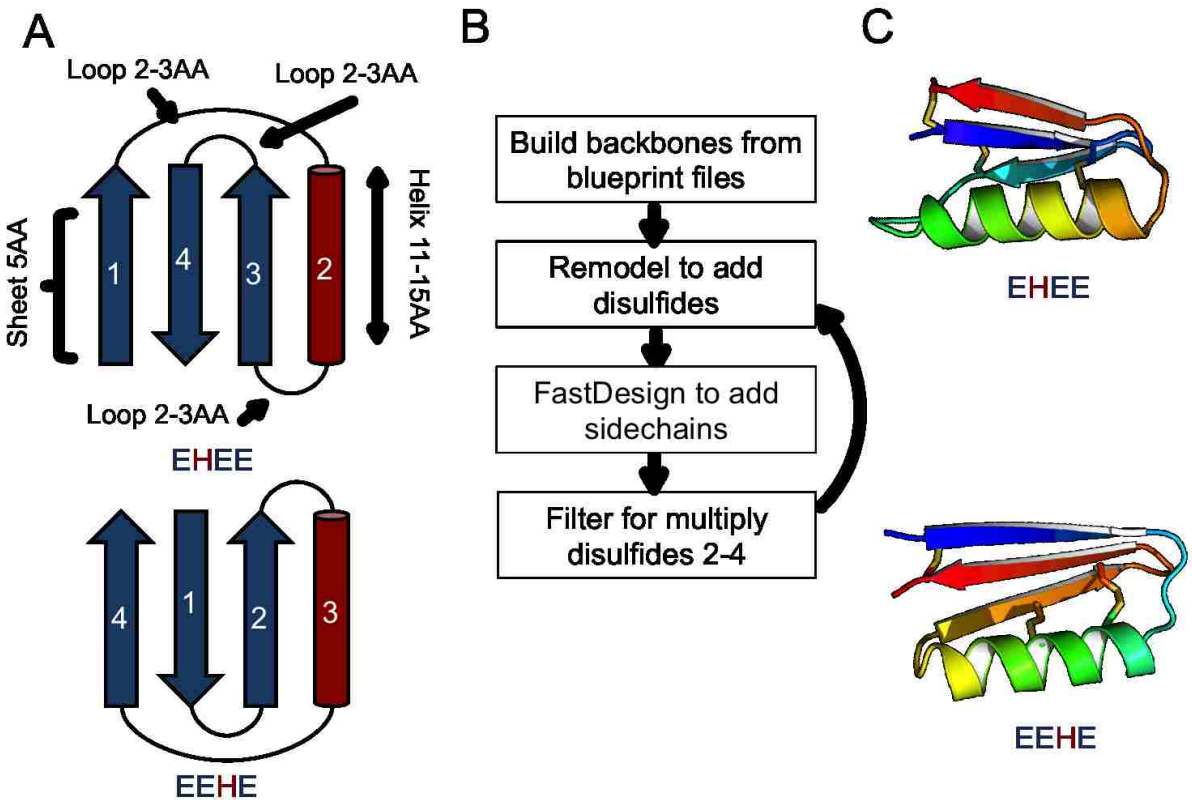


Sequence-function landscapes of *de novo* disulfided influenza binding mini-protein HB1.1. Red indicates enrichment during the selection and green indicates depletion. Panels A and B show the results of the rd1 selection using 1μM H1HA. In this loose selection, antigen concentration 10X higher than predicted K_d of starting variant, only positions that abrogate

binding completely are shown as depleted. Panel A shows the overall conservation mapped onto the design model docked to H1HA (light blue). Interface and core positions are higly conserved (green) while solvent exposed positions can accomidate mutatations (red). Panel B depicts a heatmap of the log2 binding enrichment values. Starting residue identities are shown in white font. The central binding helix (position 9-21) is easily idenetified by its conservation (rows entirely green). Interestingly, mutations to any of the cysteins abrogate binding suggesting that they are forming disulfides and critical for correct folding. Panels C and D show the results of the rd3 selection using 40nM H1HA. In this stringent selection, antigen concentration was equal to prediced K_d of starting variant, only positions that improve binding and shown as enriched. Panal C shows the enriching mutatations mapped onto the design model. The marjoity of beneficial mutations were found at the ends of secondary structure elements. Panel D depicts a heat map of the stringent selection. Only a small number of mutations show enrichment at 40nM H1HA. Positions with enrichment greater than two fold were included in the subsequent high affinity combined libraries.

Section 4: Supplementary information for high-throughput design and testing of de novo disulfided influenza binders

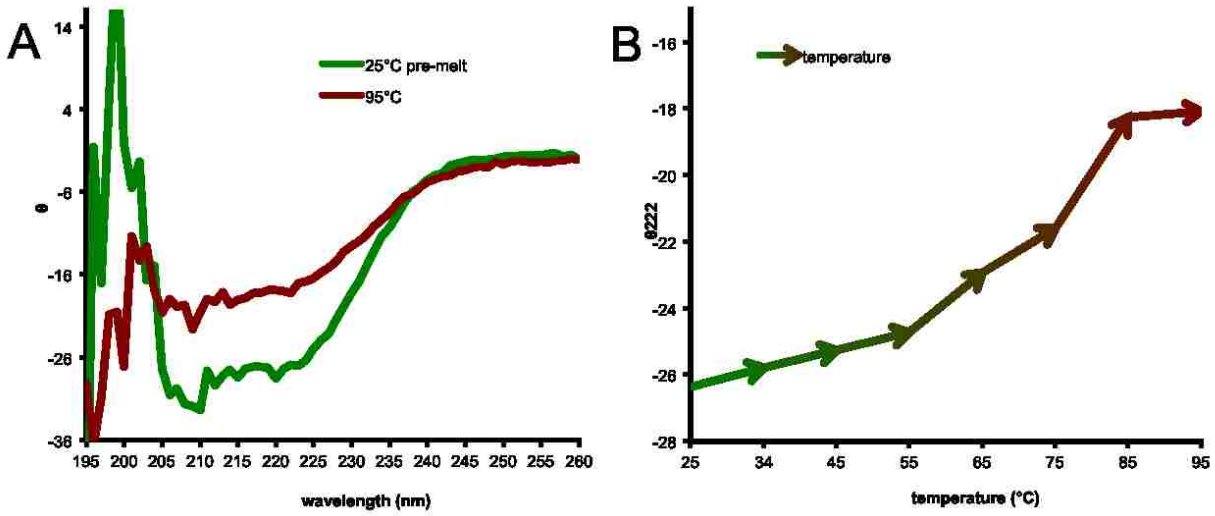
Figure 4.1



Two disulfided mini-protein topologies were built using **Protocol 4.1** and **4.2** each with 40 topological variants. The variants have different secondary structure and loop lengths as described in panel A. The topological variants are enumerated in files called "blueprints" that define the lengths and order of the elements. Panel B describes the building process. An initial tertiary protein backbone was assembled from Rosetta folding simulations to match the two-dimensional representations listed in the blueprint files. Multiple disulfides were added using RosettaRemodel. Finally, side chain information was introduced using RosettaDesign

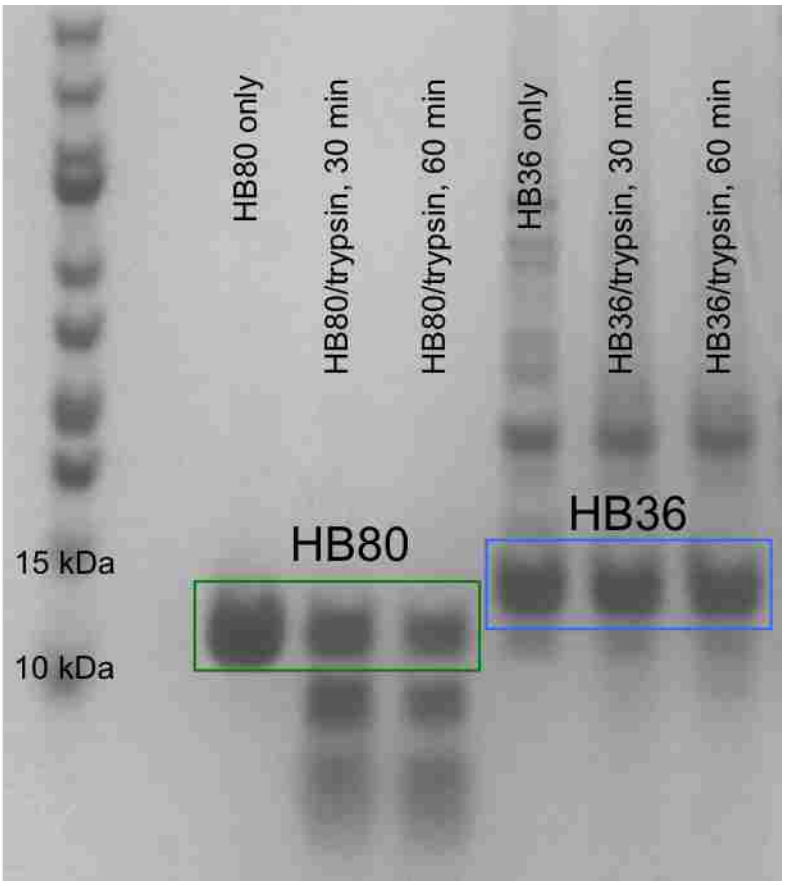
calculations followed by backbone relaxation using the Rosetta all-atom energy function. For designs that only contain a single disulfide the last two steps were iterated over to attempt to accommodate the maximum number of disulfides. Panel C shows example design models of both topologies that were built.

Figure 4.2



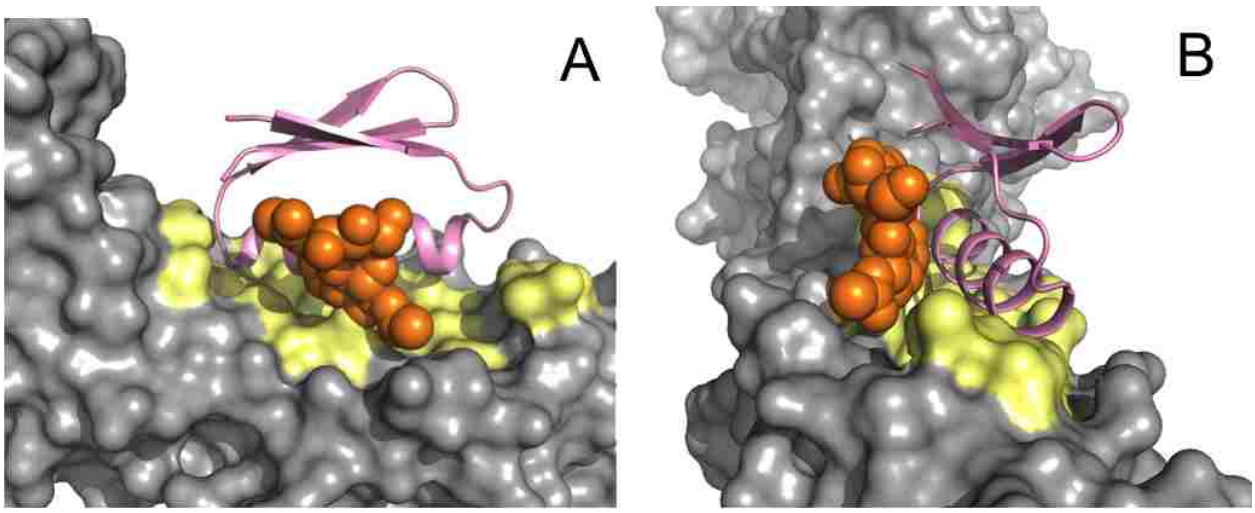
Circular dichroism data for HB36.5. Panel A shows a spectral scan at various wavelengths at two different temperatures. At 95°C HB36.5 shows poor ellipticity in the range of 222nM, indicating the protein is unfolded. As the protein is entirely alpha helix panel B shows the thermal melt curve with ellipticity at 222nM on the y-axis. The T_m was calculated from this data to be 60.4°C. Furthermore, the protein did not reversibly fold so post-melt data was not shown.

Figure 4.3



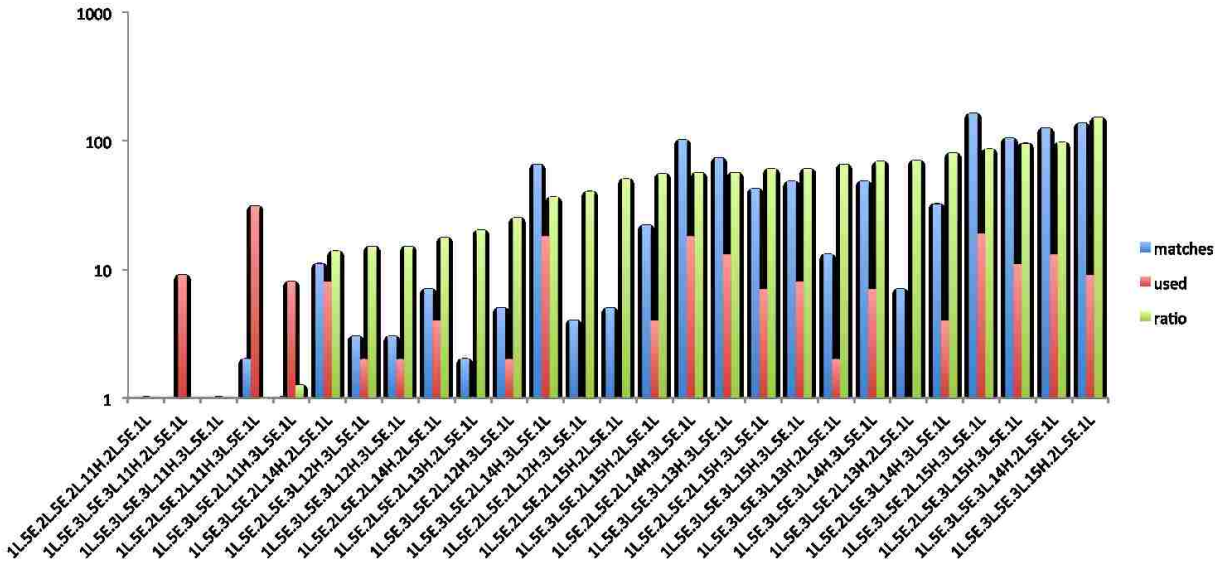
SDS-PAGE gel analysis of HB80.4 and HB36.5 binders incubated with trypsin. Trypsin was added to all samples except controls, at a final ratio of 1:25 for both binder reactions. Samples were digested overnight at 37 °C for various time points then quenched by addition of non-reducing SDS buffer and boiled for ~2min. HB80.4 shows strong degradation after 30min while HB36.4 remains mostly intact for 60min.

Figure 4.4



Panels showing the glycosylation on the stem region of H7HA (A/Netherlands/219/2003). The HA (grey) is oriented with the head region towards the left in panel A and to the top in panel B. The design model of the 37 residue binder HB7.11 (pink) is shown bound to the stem region epitope (yellow). The glycosylation, shown in orange spheres, was modeled within 2Å of the design HB7.11 backbone. This glycosylation is absent[113,114] in all Group I HAs but conserved most Group II HAs, specifically H3 and H7, making binding to the stem region epitope of these subtypes more challenging.

Figure 4.5



Shows a subset of the 80 topological variants that were built *in silico* from blueprint files. Topological variants were benchmarked by allocating each one CPU hour of build time. After one hour the number of design models generated for each variant was an indication of how "buildable" each variant was. 55 topological variants generated no models, indicating that those secondary structure combinations were unlikely to be compatible with each other (not shown here). Of the 25 topological variants built, the most buildable was 1L.5E.2L.5E.2L.11H.3L.5E.1L, which generated 31 models. The number built is shown above in red. These design models were then used as scaffolds in a trial MotifGrafting run. This was to further benchmark which of the built scaffolds were match and graft compatible with the helical interface fragments. The number of matches is shown above in blue. Some scaffolds, though buildable like 1L.5E.2L.5E.2L.11H.3L.5E.1L, were incompatible with the helical fragment sets most likely due to the amino acid length differences between the interface (13) and scaffold (11) helices. The ratio of the number of matches to designs is shown above in green. We selected the 5 topological variants with the highest ratio to include in the ACMIX scaffold

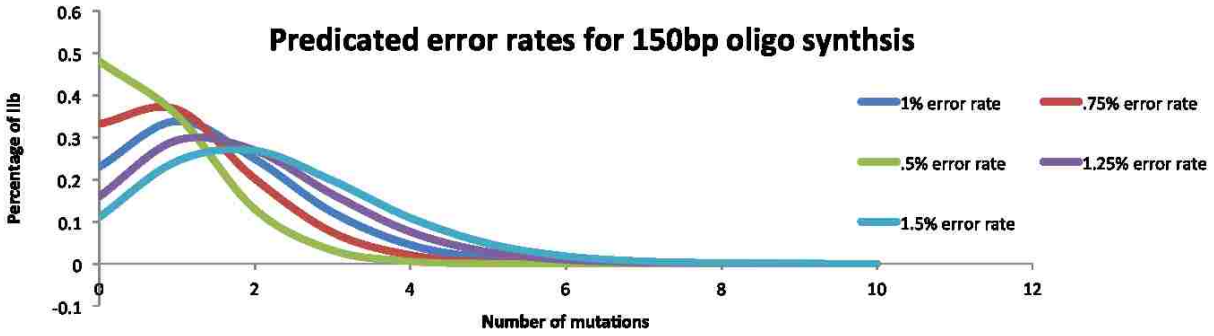
set, as they were shown to be both buildable and highly compatible with the designed interface fragments. The final mixed scaffold set (ACMIX) was comprised of 193 design models distributed across these 5 topological variants.

Figure 4.6

Designs ordered	EHEE (205)	ACMIX (193)	BC3.0 (54)	ProtG (20)
H1 de novo (23)*	1047	1005	221	15
H3 de novo (34)	2339	3450	724	8
H7 de novo (32)	1050	1695	303	20
HB80_helix (1)**	145	239	79	0
H7_helix (1)***	9	31	0	1
H1 Positive control	2	0	0	0

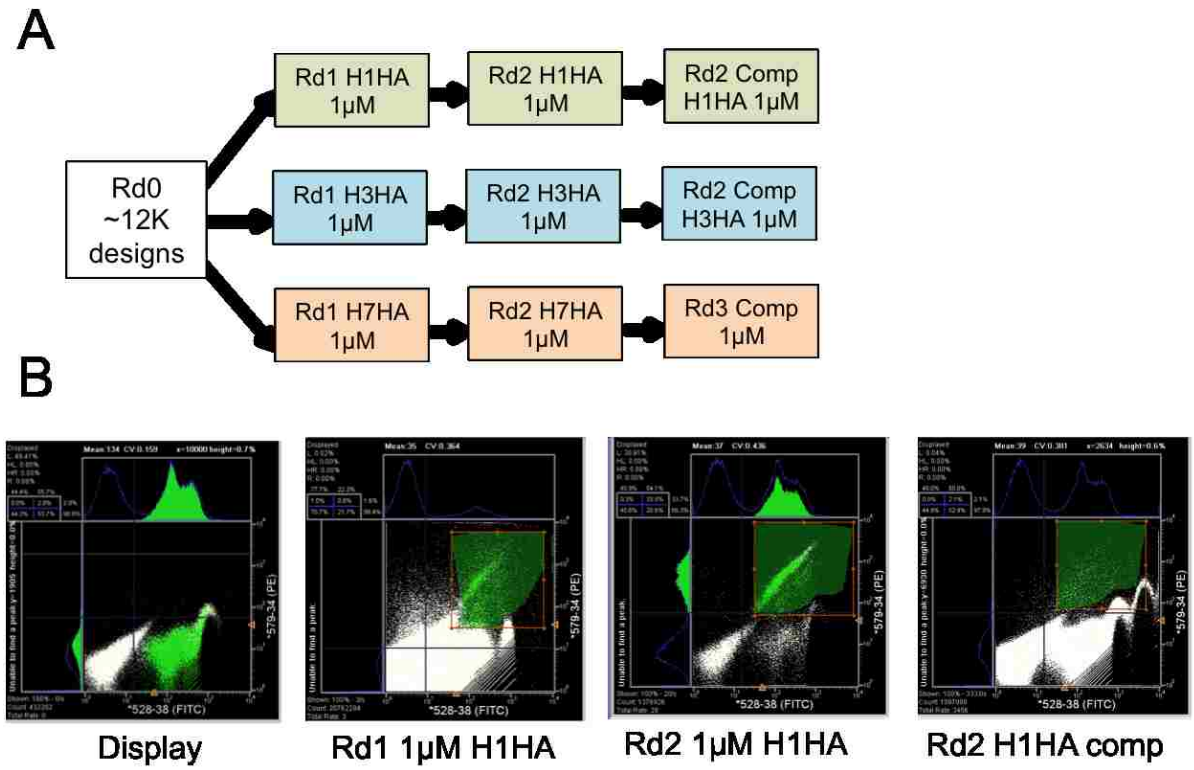
Shows the breakdown of the 12,383 designs ordered in the initial oligo pool. The headers of the columns indicate the scaffold set used with the number of scaffolds in parenthesis. The row labels indicate the interface fragment set used with number of interface fragments in parenthesis. For example 1047 designs were ordered that were designs made from MotifGrafting 23 *de novo* H1 helices into a scaffold set of 205 EHEE scaffolds. H1/H3/H7 *de novo* indicate interface fragments generated using the MotifDesign protocol. HB80_helix was a single helix taken from HB80.4, a previously published H1 binder. This was included as a control for the fragment generation protocol. The H7_helix was a single helix taken from an unpublished variant of HB80.4 that bound to H7HA. H1 positive control were complete designs that were known to bind H1 generated from the EHEE scaffold set. The average Rosetta calculated $\Delta\Delta G$ was -21.783, SC .69, and SASA 1630$Å^2$ across the entire order.

Figure 4.7



Graph modeling the predicted distribution of the number of mutations for various synthesis error rates. With a high error rate of 1.5% base misincorporation; only 12% of the 150bp oligos were predicted to have no errors. However, the array synthesis method creates many molecules per unique sequence, so in a 1ng pool made of 150bp oligos there exists 6E9 molecules representing 12,383 input sequences. A 1.5% error still results in 6E8 correct molecules available in the pool for amplification and transformation. Using high-efficiency transformation and NGS this small fraction of correct sequence was easily sorted out.

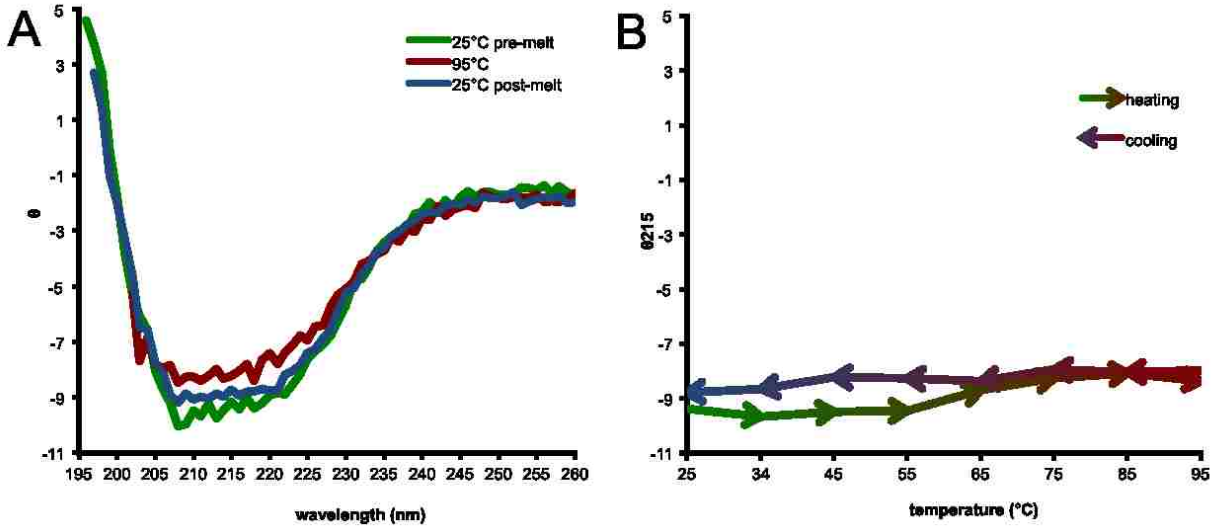
Figure 4.8



FACS screening strategy for binders. Panel A shows an outline of the two rounds of sorting against the different HA antigens. Both rounds of sorting were at high antigen concentrations (1μM) and were separated by two days of growth. A competition sort was performed during rd2 where the HA antigen was pre-incubated with 10μM known stem binding antibody FI6v3[101]. Antibody binding should block access to the stem epitope, so any displayed protein binding to the HA under these conditions would be considered a false positive (non-specific binder). Panel B shows FACS plots during each round sorting against H1HA. The X-axis shows the FITC channel (protein display) and Y-axis PE channel (binding). The green highlighted areas were area where cells were collected. Enrichment from rd1 to rd2 was demonstrated by the increase in density of binding cells. The rd2 competition sort plot shows

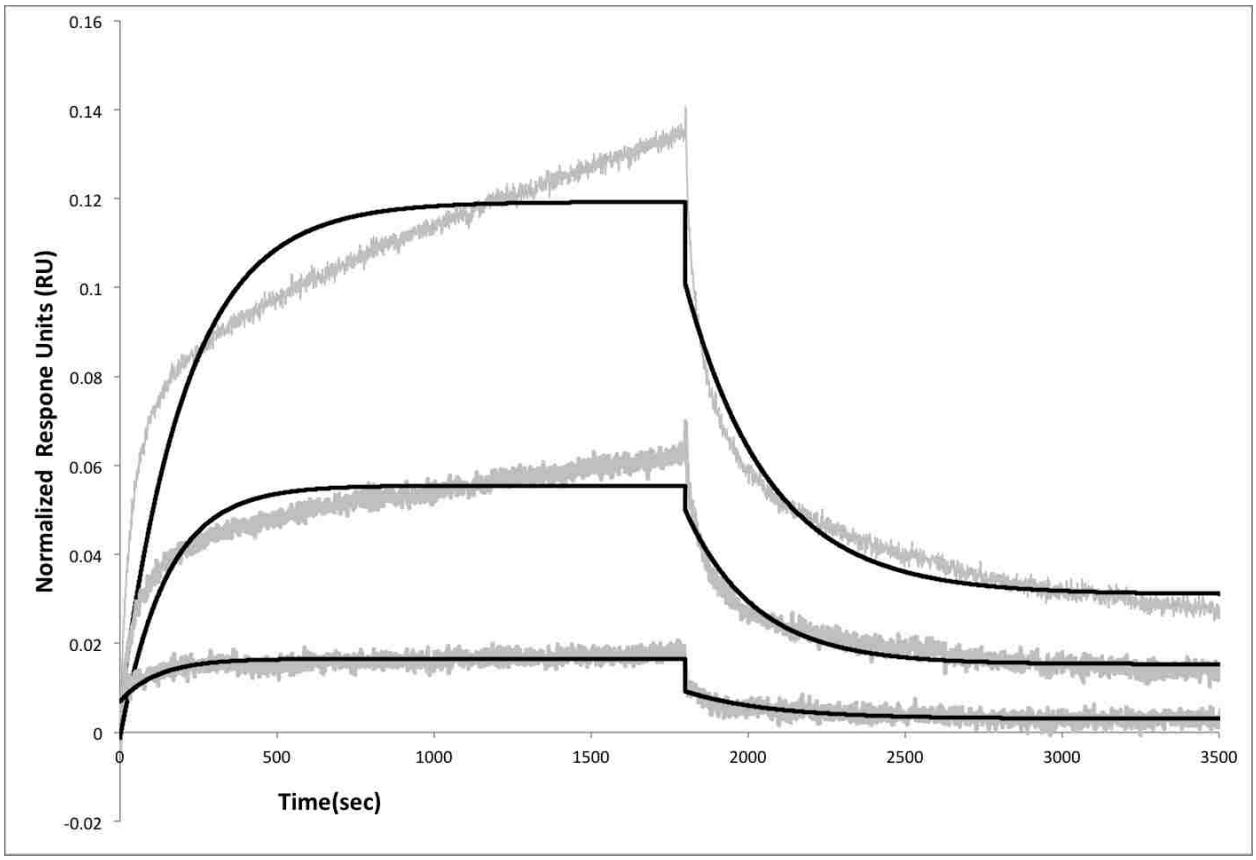
that when the binding site was blocked the majority of displayed protein binding decreased dramatically.

Figure 4.9



Circular dichroism data for HB1.1. Panel A shows a spectral scan at various wavelengths at two different temperatures including pre and post melt. At 95°C HB1.1 shows good ellipticity in the range of 215nM, indicating the protein remained folded at this high temperature. As the protein is mixed alpha beta panel B shows the thermal melt curve with ellipticity at 215nM on the y-axis. The T_m was difficult to calculate as very little structure was lost at high temperatures. Furthermore, the protein did reversibly fold, and returned to its pre-melt spectral signature post-melt.

Figure 4.10



Biolayer inferometry using an Octet Red (ForteBio, Menlo Park, CA) was used to determine subtype-specific binding for HB1.1 against biotinylated A/South Carolina/1/1918. Briefly the HA at ~10-50 µg/mL in 1x kinetics buffer (1x PBS, pH 7.4, 0.01% BSA, and 0.002% Tween 20) were loaded onto streptavidin coated biosensors and incubated with varying concentrations of HB1.1 in solution. All binding data was collected at 30°C. The experiments comprised 5 steps: 1. Baseline acquisition (60 s); 2. HA loading onto sensor (300 s); 3. Second baseline acquisition (180 s); 4. Association of HB1.1 for the measurement of k_{on} (1800 s); and 5. Dissociation of HB1.1 for the measurement of k_{off} (1700 s). Three concentrations of HB1.1 were used; 100, 50 and 10nM. Baseline and dissociation steps were carried out in buffer only. Binding kinetics were determined using a 1:1 Langmuir binding model in kinetics data analysis

mode using the ForteBio data processing software. The K_D (M) was determined to be 4.28E-08 with an error of 9.80E-10 and K_{on} of 2.17E04.

Table 4.1

	collected	viewed	%
unsorted library	0	0	100.00%
display rd1	2443934	12017390	20.34%
bind H1 rd1	3468	20762293	0.02%
bind H3 rd1	8185	22031996	0.04%
bind H7 rd1	3080	21201796	0.01%
display H1 rd2	202568	432202	46.87%
display H3 rd2	406005	765343	53.05%
display H7 rd2	225000	500000	45.00%
bind H1 rd2	405155	1376926	29.42%
bind H3 rd2	12060	3360024	0.36%
bind H7 rd2	6152	1766249	0.35%
comp H1 rd2	404	1096000	0.04%
comp H3 rd2	2395	1041406	0.23%
comp H7 rd2	2043	1056784	0.19%

Shows the number of cells collected and total number of cells viewed on the FACS sorter during each round of selection. Each population of collected cells was grown to saturation (24-48hrs) and prepped for NGS. Both the display population and binding population increased between rd1 and rd2 indicating the selection was enriching for binders.

Table 4.2

seq_ID	naïve	rd1 display	rd1 H1 bind	rd1 H3 bind	rd1 H7 bind	rd2 display H1	rd2 display H3	rd2 display H7	rd2 bind H1	rd2 bind H3	rd2 bind H7	comp H1	comp H3	comp H7
HB1.1	64	1	113	0	0	67	0	0	85	1	0	2	0	0
HB1.2	143	22	7544	36	10	6169	4	10	11952	14	7	9	10	10
HB1.3	146	24	12207	88	9	12533	18	25	27082	25	8	914	22	22
HB1.4	4	1	800	2	0	288	0	0	385	0	1	2	0	0
HB1.5	16	1	2189	3	2	1124	2	1	1707	3	2	4	2	2
HB3.1	2	1	0	30	0	0	16	0	0	161	0	0	0	0
HB3.2	3	1	0	117	0	0	59	0	1	188	0	2	0	0
HB7.10	166	18	32	17	561	10	71	597	5	26	12280	35	19	19
HB7.11	28	3	4	6	571	1	4	292	2	2	7654	3	5	5
HB7.13	2174	368	249	279	2489	191	725	2764	160	450	293564	553	415	415
HB7.14	581	108	16	27	3433	11	100	5255	18	47	8984	100	40	40
HB7.3	353	85	18	3	791	3	13	856	3	8	3955	8	4	4
HB7.4	148	9	0	0	221	0	0	8	1	1	68	0	0	0
HB7.5	20	6	1	1	78	5	30	29	0	2	251	18	2	2
HB7.7	8	0	0	0	102	0	0	37	0	1	160	1	0	0
HB7.7	149	7	1	1	228	0	2	167	0	6	2698	1	2	2
HB7.8	58	22	2	6	731	4	20	572	7	16	9901	30	16	16
HB7.9	11	8	1	4	124	1	20	68	0	2	1215	20	3	3
totals	272344	34488	314347	21275	53672	312963	18067	54327	623737	2533	353699	340633	2617	2617

Shows a subset of raw deep sequencing count data from 14 yeast pools collected during the selection, each pool individually barcoded and sequenced. The column headers describe which pool the counts originated from. The rows represent individual designs, with the last row the total number of counts found in that pool. Each count represents a paired-end Miseq read that perfectly matches a designed ordered sequence. Enrichment data was calculated directly from the raw count data.

Table 4.3

	rd1 Enrich H1	rd1 Enrich H3	rd1 Enrich H7	rd2 Enrich H1	rd2 Enrich H3	rd2 Enrich H7	Comp H1	Comp H3	Comp H7
HB1.1	161.88	0.00	0.00	1.75	#DIV/0!	--	0.03	--	--
HB1.2	491.23	1.62	0.57	2.68	3.14	1.40	0.00	1.88	1.34
HB1.3	728.62	3.63	0.47	2.99	1.25	0.64	0.08	0.92	1.18
HB1.4	1146.02	1.98	0.00	1.85	--	--	0.01	--	--
HB1.5	3135.80	2.97	2.51	2.10	1.35	3.99	0.00		2.68
HB3.1	0.00	29.67	0.00	--	9.03	--	--	0.00	--
HB3.2	0.00	115.70	0.00	--	2.86	--	--	0.00	--
HB7.10	2.55	0.93	39.12	0.69	0.33	41.07	3.62	0.20	0.04
HB7.11	1.91	1.98		2.76	0.45	52.34	3.10	0.94	0.02
HB7.13	0.97	0.75	8.49	1.16	0.56	212.07	3.00	0.43	0.20
HB7.14	0.21	0.25	39.90	2.26	0.42	3.41	9.40	0.30	0.01
HB7.3	0.30	0.03	11.68	1.38	0.55	9.23	2.76	0.23	0.01
HB7.4	0.00	0.00	30.83	--	--	16.97	--	--	0.00
HB7.5	0.24	0.16	16.32	0.00	0.06	17.28	3.72	0.05	0.09
HB7.7	--	--	--	--	--	8.63	--	--	0.00
HB7.7	0.20	0.14	40.89	--	2.69	32.26	--	0.75	0.02
HB7.8	0.13	0.27	41.71	2.42	0.72	34.56	7.76	0.60	0.04
HB7.9	0.18	0.49	19.46	0.00	0.09	35.68	20.69	0.11	0.06

Shows a subset of the enrichment data from 14 yeast pools collected during the selection. The enrichment score was calculated by dividing the population fraction of the variant in the binding pool by its population fraction in the corresponding display pool. We selected the highest enriching variants (enriched in both rd1 and rd2), shown here, for analysis by deep sequencing. Enrichment scores were analyzed in combination with total raw counts to avoid selecting high enriched low counts variants (noise) for structural validation. We were unable to find any variant that was strongly enriched for multiple subtypes. Also, two H3 binders were selected for (HB3.1 and HB3.2) for SSM structural validation even though they had relatively low raw counts and mediocre rd2 enrichments.

Table 4.4

Seq ID	ordered DNA	length	Aasequence	length	variants
HB1.1	CGCTCTTCTGTTGCTCAAGCTGGTCAATTTGCTTGTGAAGTGGCATGTCGTATCGCTGCTTCATTCGGTTGCACTTGTACTACAGATGGTGACACATGTAAAGTTACTTGT	111	RGGVAQAGQFACEVACRIAASFGCTCTTDGDTCKVTC	37	703
HB1.2	CCATGTGCTAGAATTGATTCAAACACTTTCGCTGCACAGATAGCTTGTGAAATTTGCAAGGATTTTGGTGCTGAATGTAGGGATGACGGTAATGTTGTTGAATGCTGTTTG	111	PCARIDSNTFAAQIACEICKDFGAECRDDGNVVECCL	37	703
HB1.3	AGGGAAAAAACTCACGACACAAGGAGATTCGCATGCTCAGTTGTAAAAGAGATTGCTAGGTCTTTCGGATGCAGGGTTGAAGAGGATGGTTCTAGATGTGTTGTTAAATGT	111	RKKTHDTRRFACSVVKEIARSFGCRVEEDGSNCVVKC	37	703
HB1.4	GAAACTAGAAGATGCAAAGGTTTCGCAGAGTGCATAAGGTGTGAGATAGAGGCAGAGTTCAAGAAGGGATGCACTTCAAAGAGACATGGTGAATACTGCGAAGTTTTTTGT	111	ETRRCKGFAECIRCEIAAEFKKGCTSKRNGNVCEVFC	37	703
HB1.5	TTGTGTTCTACTTGTGACTGTAACAACTTCGCTTGTAAGATTGCAGCAGAAATTATGAGAGAATTTGGTGGAACTTCAACTTGCTCTGGTGGTTTGTTGACTTATTGTCAA	111	LCSTCDCNNFACEIAAEIMREFGGTSTCGGGLLTYCQ	37	703
HB1.6	CCATGTACTCAATTGGATACCACCTCTTTTGCCGCTAAGATTGCTTGTTATGCCGCTAAAGAGCTTGGATGTCAATGCAGGGATGACGGTAATGTTGTTGAAGTTTGTTGT	111	PCTQLDTTSFAAKIACYAAAELGCQCRDDGNVVEVCC	37	703
HB3.1	AGGAAGGAAACTTACAAGTCAAGAGAATTTGACTGCCAGTTGATTGCTATATTGGCAAGGGTAAATGGTTGTGAATCTTACGCAAACGGTACTGAGTGTGTTGTTATTTGT	111	RRETYKGREFDCQLIAILARVRGCESYAWGTECVVIC	37	703
HB3.2	ACTACTACTGAAAGAAAGCACGACAACCAGTTGGAAGCAGCAATTAAGTGCTTTATTGAGGCAATGTTGAAGCACTGCAAGTTGAGGATTGAGAACGTTATTTGCTACATATACTGT	117	TTTERKHDNQLEAAIKCFIRAMLKHCRLRIENVICYIYC	39	742
HB7.10	GGTTGTACTTCTTGCGACAATGCTAGAGCATTCAAGAGGATATTGTGCTTGATTTTGTGCAGGGCAAAGAAAACTTGCACAGAGTCTGGAGGAAAGATTAAGATGTGCGGT	111	GCTSCDNARAFKRILCLILCRAKETCTEGGGKIRMCG	37	703
HB7.11	ACTTGTACTACTGTTAAAGGATCTGATGAGAAAAAGGCTTTGAGATTGGCAATTTTGTGCTTGTCATTGAGGTGTATATGTTCTGTACATGGTCATTATGTTGTTATTTGCTGT	114	TCTTVRGSDERKALRLAILCLSLRCICSVRGHTVVICC	38	722
HB7.13	ACTTGTGCTAAGTTGTGTTCTACAGCTGAAGCACATAGAAGGTTGATATGTTTGGTATGCAGAGCTTCTGGTGAATGCAGAATTGACGGTAACTGTGTAGTATGTTGCTTT	111	TCAKLCSTAEARRRLICLVCRASGECRIDGNCVVCCP	37	703
HB7.14	GGTTGCCATTCTGTTGACGGTGACATGTCTTTGAAGAAGATTTTGTGCTTGATGGCATGTTTGGCTGCAAGATTGACAACTCACCAAAACGGTTTGTGTGTTTGCTGCGCT	111	GCHSVDGDMSLEKILCIMACLAARLTTHQRGLCVCCA	37	703
HB7.3	GGATGCAGGAGAAAGTGCGGAAGGTCAGAGTTGAGGAAGGTAATGTGCATGTTGTTGAAGATTGCTTGCACTGTTTCTAGGACAGAGAAGGACTGTTGCGTTTATTGCTGT	111	GCRRKCGRSELRKVMCHLLKIACTVSRTEKDCVYCC	37	703
HB7.5	GAACATCACGAAAGGAAGAACAGAAACCATGAGTGCGTAAAGTTGCACATTTTGGCAAGGGTAAGGGGATGTATTCACACAAGAAAAAACAATAGATGTAGAGTTAAATGT	111	EHHERKNGHHECVKLHILARVRGCINTRKNHHCRVKC	37	703
HB7.7	ACTTGCACAACTGTTAAGGGTTCAGACTTGATTCAGATGATAAAGTTGATGCATGAGTGCGCTAAGAAGAGGTGCATTTGCTCACATCACGGTCATTATGTTTGTTATATGCTGT	114	TCTTVRGSDLIQMIKLMHECARKRCICSHRGHTVVICC	38	722
HB7.8	ACTTGTGCTAAGATTTGTTCTACTGCTGATGCTCATAGACAATTGGTTTGCTTGGTATGTAGAGCTTCTGGTGAATGTAGAATTGATGGTAATTGTGTAGTTTGTTGTTTT	111	TCAKICSTADARRQLVCLVCRASGECRIDGNCVVCCP	37	703
HB7.9	ACTTGTGTTAGATATTTGAAAAGGGAATTTGCTTGCAAATTGATGGAGGTTTTGGCAAGATTTAGAGGATGTTTGTATAGGAGACACGGTAAAACTTGTGAATTGTGTTGT	111	TCVRVLKREFACKIMEVLARFRGCLYAHRGETCELCC	37	703

Shows the DNA and amino acid coding sequences of the 17 designs included in the SSM structural analysis. Their enrichment scores are in the online repository (see below).

The below RosettaScripts XML script was used to generate backbones from blueprint files.

Protocol 4.1 - Builder.xml

```
<dock_design>
<SCOREFXNS>
   <sfxn_std  weights=talaris2013/>
        <SFXN1 weights="fldsgn_cen">
                #Reweight scoretype="cenpack" weight="1.0" />
                <Reweight scoretype="hbond_sr_bb" weight="1.0" />
                <Reweight scoretype="hbond_lr_bb" weight="1.0" />
                <Reweight scoretype="atom_pair_constraint" weight="1.0" />
                <Reweight scoretype="angle_constraint" weight="1.0" />
                <Reweight scoretype="dihedral_constraint" weight="1.0" />
        </SFXN1>
</SCOREFXNS>
<FILTERS>
   <ScoreType name="hbond_sfn" scorefxn=sfxn_std score_type=hbond_lr_bb threshold=0/>
        <HelixKink name="hk1"
blueprint="./epigraftRosettaScripts_baseFiles_PBS/bigger.blueprint" />
        <SheetTopology name="sf1"
blueprint="./epigraftRosettaScripts_baseFiles_PBS/bigger.blueprint" />
        <SecondaryStructure name="ss1"
blueprint="./epigraftRosettaScripts_baseFiles_PBS/bigger.blueprint" />
        <CompoundStatement name="cs1">
                <AND filter_name="ss1" /> #secondary structure filter, checks against original
blueprint
                <AND filter_name="hk1" /> #helix kinked filter, NO KINKED HELIX
                <AND filter_name="sf1" /> #sheet topology filter, is correct strand pairing
obeyed as defined in blueprint
        </CompoundStatement>
        <ScoreType name="total_score_cen" score_type="total_score" scorefxn="SFXN2"
confidence="0" threshold="0" />
        <AverageDegree name=degree confidence=1 threshold=9.5/>
        <PackStat name=pack confidence=0/>
```

```
        <AtomicContactCount name=contact confidence=0/>
        <CavityVolume name=cavity confidence=0/>
</FILTERS>
<TASKOPERATIONS>
</TASKOPERATIONS>
<MOVERS>
        <Dssp name="dssp" />
        <SheetCstGenerator name="sheet_new1" cacb_dihedral_tolerance="0.1"
        blueprint="./epigraftRosettaScripts_baseFiles_PBS/bigger.blueprint" />
        <RemoveCsts name="sheet_rm1" generator="sheet_new1" />
        <SetSecStructEnergies name="set_ssene1" scorefxn="SFXN1"
        blueprint="./epigraftRosettaScripts_baseFiles_PBS/bigger.blueprint" />
        <BluePrintBDR name="bdr1" use_abego_bias="1" scorefxn="SFXN1"
        constraint_generators="sheet_new1" constraints_NtoC="-1.0"
        blueprint="./epigraftRosettaScripts_baseFiles_PBS/bigger.blueprint" />
        <ParsedProtocol name="build_dssp1" >
                <Add mover_name="bdr1" />  #builds backbone
                <Add mover_name="dssp" />  #evaluate secondary structure of newly built pose
                <Add filter_name="cs1" />  #verifies SS, helix kink, and strand topology are all
obeyed
                <Add filter_name="degree" /> #packing filter AverageDegree
        </ParsedProtocol>
</MOVERS>
<PROTOCOLS>
        <Add mover_name="set_ssene1" />  #read ss constraints from blueprint
        <Add mover_name="build_dssp1" /> #sets secondary structure constraints from
blueprint file
        <Add filter_name="total_score_cen" />
        <Add filter_name="cs1" />
        <Add filter_name="contact" />
        <Add filter_name="cavity" />
        <Add filter_name="pack" />
```

</PROTOCOLS>

</dock_design>

The below RosettaScripts XML script was used to add side chains and insert disulfides into successfully built backbones.

Protocol 4.2 - disulfide.xml

<dock_design>

<SCOREFXNS>

<sfxn_std weights=talaris2013/>

</SCOREFXNS>

<FILTERS>

<ResidueCount name="total_res" />

<ResidueCount name="total_cys" residue_types="CYD" />

<ResidueCount name="count_cys" max_residue_count=6 residue_types="CYD" /> #if fewer than 6 CYD returns true

<ResidueCount name="end_cys" min_residue_count=4 max_residue_count=6 residue_types="CYD" confidence=1/> #if fewer than 4 CYD kills trajectory

<ScoreType name="hbond_sfn" scorefxn=sfxn_std score_type=hbond_lr_bb threshold=0/>

<ScoreType name="dslf_fa13" scorefxn=sfxn_std score_type=dslf_fa13 threshold=0/>

<AverageDegree name=degree confidence=1 threshold=9.5/>

<PackStat name=pack confidence=0/>

<ExposedHydrophobics name=exposed confidence=0/>

<AtomicContactCount name=contact confidence=0/>

<CavityVolume name=cavity confidence=0/>

<CalculatorFilter name=bb equation="hbond / rescount" threshold="-0.30" confidence=1>

<VAR name="hbond" filter="hbond_sfn"/>

<VAR name="rescount" filter="total_res"/>

</CalculatorFilter>

<CalculatorFilter name=mean_dslf equation="dslf / cyscount" threshold="-0.35" confidence=1>

<VAR name="dslf" filter="dslf_fa13"/>

<VAR name="cyscount" filter="total_cys"/>

</CalculatorFilter>

</FILTERS>

<TASKOPERATIONS>

<LimitAromaChi2 name="limitchi2" include_trp="1" />

<LayerDesign name="layer_all" layer="core_boundary_surface_Nterm_Cterm" core_H="15" core_E="15" core_L="25" surface_H="60" surface_E="60" surface_L="40" pore_radius="2.0" verbose="true" />

<NoRepackDisulfides name="exemptdisulf" />

</TASKOPERATIONS>

<MOVERS>

<Dssp name="dssp" />

<DumpPdb name="dump" fname="pass" tag_time=True/>

<RemodelMover name="remodel" fast_disulf="True" match_rt_limit="1.5" quick_and_dirty="True" bypass_fragments="True" min_disulfides=2 max_disulfides=3/>

<RemodelMover name="remodel1" fast_disulf="True" match_rt_limit="3.5" quick_and_dirty="True" bypass_fragments="True" min_disulfides=1 max_disulfides=1/>

<If name="add_more_disulf" filter_name="count_cys" true_mover_name="remodel1" />

<FastDesign name="fastdes" task_operations="limitchi2,layer_all,exemptdisulf" scorefxn="sfxn_std" allow_design="1" clear_designable_residues="0" repeats="2" ramp_down_constraints="0" />

<FastDesign name="fastdes4" task_operations="limitchi2,layer_all,exemptdisulf" scorefxn="sfxn_std" allow_design="1" clear_designable_residues="0" repeats="4" ramp_down_constraints="0" />

<ParsedProtocol name="build_disulf">

<Add mover_name="remodel" /> #calls remodel to add one or two disulfides

<Add mover_name="fastdes"/> #round of fastdesign, possible replace with packrotomers

<Add mover_name="add_more_disulf" /> #count number of disulfidews and adds ONE more if needed

<Add mover_name="fastdes4"/> #more fastdesign

<Add filter_name="degree" /> #check for packing

<Add filter_name="end_cys"/> #has to have two cysteins

<Add filter_name="bb" /> #checks average backbone hydrogen energy, for better secondary structure

<Add filter_name="mean_dslf" /> #average of the disulfide energy per disulfide

</ParsedProtocol>

</MOVERS>

<PROTOCOLS>

<Add mover_name="build_disulf" />

<Add filter_name="total_res" />

<Add filter_name="total_cys" />

<Add mover_name="dssp" /> #checks secondary structure again, doesn't do anything

</PROTOCOLS>

</dock_design>

The below RosettaScripts XML script was used to add side chains to MotifDocked poly alanine helices.

Protocol 4.3 - design_asym.xml

<ROSETTASCRIPTS>

<SCOREFXNS>

<sfx_hard_surf weights="/work/sheffler/Dropbox/test/matdes/asym_iface/input/talaris2013.wts" symmetric=0 >
</sfx_hard_surf>

<sfx_hard weights="/work/sheffler/Dropbox/test/matdes/asym_iface/input/talaris2013.wts" symmetric=0 >
</sfx_hard>

<sfx_vanilla weights="talaris2013" />

<sfx_soft weights=soft_rep symmetric=0 />

</SCOREFXNS>

<TASKOPERATIONS>

<MotifResidues name=core_motifs mode="place" merge="union" dumpfile="" motif_sets="xs_scbb_aa1_resl0.8_smooth1.3_msc0.4_mbv1.0/xs_scbb_aa1_resl0.8_smooth1.3_msc0.4_mbv1.0.rpm.bin.gz" ex1=4 ex2=1 /> #recreate motifs

<ProteinInterfaceDesign name=frag_only design_chain1="0" design_chain2="1" jump="1" interface_distance_cutoff="15"/> #only design fragment side within 15A

<ReadResfile name=surf_resfile filename="/work/sheffler/Dropbox/test/matdes/asym_iface/input/surf.resfile" /> #read resfile generated from MotifResidues

<IncludeCurrent name=ic /> #include current rotamers

<DisallowIfNonnative name=dsgn_aa_core disallow_aas=CDGNPST /> #disallow core polars and others

<DisallowIfNonnative name=dsgn_aa_surf disallow_aas=CGP /> #disallow surface CGP

<RestrictToInterface name=interface_only jump=1 distance=10.0 /> #only design at interface both sides

<LimitAromaChi2 name=limitaro chi2max=110 chi2min=70 /> #avoids bad rotomers

<RetrieveStoredTask name=design_task task_name="design_task" /> #Retrieves a stored packer task from the pose's cacheable data; must be used in conjunction with the StoreTask mover. Allows the caching and retrieval of tasks such that a packer task can be defined at an arbitrary point in a RosettaScripts protocol and used again later. This is useful when changes to the pose in the intervening time may result in a different packer task even though the same task operations are applied.

<RestrictToRepacking name=repack_only /> #repack only no design

<SelectBySASA name=no_core_mono_repack mode="sc" state="monomer" probe_radius=2.2 core_asa=0 surface_asa=30 core=0 boundary=1 surface=1 verbose=1 />

<SelectBySASA name=core mode="sc" state="bound" core_asa=30 surface_asa=0 core=1 boundary=0 surface=0 verbose=1 />

<SelectBySASA name=surf mode="sc" state="bound" core_asa=0 surface_asa=30 core=0 boundary=0 surface=1 verbose=1 />

<OperateOnCertainResidues name=repack_target>

<RestrictToRepackingRLT/> #Turn off design on the positions selected by the accompanying ResFilter.

<ChainIs chain=A/>

</OperateOnCertainResidues>

</TASKOPERATIONS>

<FILTERS>

<ShapeComplementarity name=sc_filt0 jump=1 verbose=1 min_sc=0.0 write_int_area=1 cache=1 />

<ShapeComplementarity name=sc_filt1 jump=1 verbose=1 min_sc=0.65 write_int_area=1 cache=1 confidence=1 /> #why two shape complementarity filters?

<OligomericAverageDegree name=avg_deg threshold=8 distance_threshold=10.0 write2pdb=1 task_operations=design_task /> #Degree of connectivity of a subset of residues

<Ddg name=ddG_filt scorefxn=sfx_hard jump=1 repack=1 repeats=3 threshold=-10 confidence=1 />

<SymUnsatHbonds name=unsat_pols jump=1 cutoff=50 verbose=1 write2pdb=1 />

<RotamerBoltzmannWeight name=rotboltz task_operations=design_task radius=6.0 jump=1 unbound=1 scorefxn=sfx_hard temperature=0.8 repack=1 skip_ala_scan=1 write2pdb=1 />

<Sasa name=sasa_int_area threshold=700 upper_threshold=10000 hydrophobic=0 polar=0 jump=1 confidence=1 />

<AverageInterfaceEnergy name=air_energy task_operations=design_task scorefxn=sfx_hard cutoff=0 bb_bb=0 />

<ScoreType name=total_score scorefxn=sfx_hard score_type=total_score threshold=0 confidence=0/>

<CombinedValue name=ddg_cst_e confidence=0>

<Add filter_name=ddG_filt factor=1 />

<Add filter_name=total_score factor=1 />

</CombinedValue>

</FILTERS>

<MOVERS>

<StoreTaskMover name=store_design_task task_name="design_task" task_operations=interface_only,no_core_mono_repack,limitaro />

<PackRotamersMover name=design_soft scorefxn=sfx_soft task_operations=design_task,core,dsgn_aa_core,core_motifs,frag_only />

<PackRotamersMover name=design_hard scorefxn=sfx_hard task_operations=design_task,core,dsgn_aa_core,core_motifs,frag_only />

<PackRotamersMover name=design_elec scorefxn=sfx_hard_surf task_operations=design_task,surf,dsgn_aa_surf,surf_resfile,frag_only />

```
        <PackRotamersMover name=repack scorefxn=sfx_hard
task_operations=design_task,repack_only,frag_only />

        <TaskAwareMinMover name=min0 scorefxn=sfx_hard bb=0 chi=1 jump=0
task_operations=design_task />

        <TaskAwareMinMover name=min  scorefxn=sfx_hard bb=0 chi=1 jump=1
task_operations=design_task,frag_only />

        <TaskAwareMinMover name=minc scorefxn=sfx_hard bb=0 chi=1 jump=0
task_operations=design_task,core />

        <TaskAwareMinMover name=mins scorefxn=sfx_hard_surf bb=0 chi=1 jump=0
task_operations=design_task,surf  />

        <TaskAwareMinMover name=min1 scorefxn=sfx_hard bb=1 chi=1 jump=0
task_operations=design_task />

        <ParsedProtocol name=design_min_soft>

            <Add mover=design_soft />

            <Add mover=minc />

        </ParsedProtocol>

        <ParsedProtocol name=design_min_hard>

            <Add mover=design_hard />

            <Add mover=minc />

        </ParsedProtocol>

    <ParsedProtocol name=design_core>

        <add mover_name=min0 />

        <add mover_name=design_min_soft />

        <add mover_name=design_min_hard />

    </ParsedProtocol>

    <ParsedProtocol name=design_surf>

        <add mover_name=design_elec />

        <add mover_name=mins />

    </ParsedProtocol>

        <ParsedProtocol name=min_repack_min>

            <Add mover=min />

            <Add mover=repack />
```

```
                <Add mover=min />
        </ParsedProtocol>


                <MinMover name=bbrb_min_all_soft  scorefxn=sfx_soft  bb=1 chi=1 jump=1
cartesian=0 />
                <MinMover name=bbrb_min_all_hard  scorefxn=sfx_hard  bb=1 chi=1 jump=1
cartesian=0 />
                <MinMover name=bbrb_min_all_hard_void  scorefxn=sfx_hard_void  bb=1
chi=1 jump=1 cartesian=0 />
                <MinMover name=bbrb_min_all_vanilla  scorefxn=sfx_vanilla  bb=1 chi=1
jump=1 cartesian=0 />
                <RollMover name="random_rb_perturb" chain=1 random_roll=1
random_roll_angle_mag=4.0 random_roll_trans_mag=0.4 />
        </MOVERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
                <Add mover_name=random_rb_perturb />
                <Add mover_name=store_design_task />
                <Add mover_name=design_soft />   <Add mover_name=bbrb_min_all_soft />
                <Add mover_name=design_soft />   <Add mover_name=bbrb_min_all_hard />
                <Add mover_name=design_hard />   <Add mover_name=bbrb_min_all_hard />
                <Add mover_name=design_elec />
                <Add mover_name=bbrb_min_all_vanilla />
                <Add filter_name=sc_filt0 />
                <Add mover_name=bbrb_min_all_hard_void />
                <Add filter_name=sc_filt1 />
                <Add filter_name=ddG_filt />
                <Add filter_name=unsat_pols />
                <Add filter_name=sasa_int_area />
        </PROTOCOLS>
```

</ROSETTASCRIPTS>

The below RosettaScripts XML script was used to superposition and design the interface helices onto the generated scaffolds.

Protocol 4.4 - motifgraft.xml

```
<ROSETTASCRIPTS>

        <SCOREFXNS>
                <sfx_vanilla weights="/usr/lusers/achev/weights/talaris2013.wts"> </sfx_vanilla>
                <sfx_cart weights="talaris2013">
                        <Reweight scoretype="cart_bonded" weight="1.0"/>
      </sfx_cart>
        </SCOREFXNS>
        <TASKOPERATIONS>
                <IncludeCurrent name=ic />  #include current rotamers
                <DisallowIfNonnative name=dsgn_aa_core disallow_aas=CDGNPSTEQ />
#disallow core polars and others
                <DisallowIfNonnative name=dsgn_aa_surf disallow_aas=CGP /> #disallow
surface CGP
                <RestrictToInterface name=interface_only jump=1 distance=20 /> #only design
at interface both sides
                <LimitAromaChi2 name=limitaro chi2max=110 chi2min=70 />  #avoids bad
rotomers
                <RetrieveStoredTask name=design_task task_name="design_task" />
                <RestrictToRepacking name=repack_only />  #repack only no design
                <SelectBySASA name=no_core_mono_repack mode="sc" state="monomer"
probe_radius=2.2 core_asa=0 surface_asa=30 core=0 boundary=1 surface=1 verbose=1 />
                <SelectBySASA name=core mode="sc" state="bound" core_asa=30
surface_asa=0 core=1 boundary=0 surface=0 verbose=1 />
                 <SelectBySASA name=surf mode="sc" state="bound" core_asa=0
surface_asa=30 core=0 boundary=0 surface=1 verbose=1 />
```

<OperateOnCertainResidues name=no_repack_target>

<PreventRepackingRLT/> #Turn off design and repacking on the positions selected by the accompanying ResFilter.

<ChainIs chain=A/>

</OperateOnCertainResidues>

<RestrictIdentities name=no_glycan identities=AX1 prevent_repacking=1 />

<RestrictIdentities name=no_CYS identities=CYS prevent_repacking=1 />

</TASKOPERATIONS>

<MOVERS>

<RigidBodyTransMover name=unbound jump=1 distance=20/> #for MoveBeforeFilter, used n calculating no_glycan SASA score

</MOVERS>

<FILTERS>

<ShapeComplementarity name=sc_filt0 jump=1 verbose=1 min_sc=0.65 write_int_area=1 confidence=0/>

<ShapeComplementarity name=sc_filt1 jump=1 verbose=1 min_sc=0.65 write_int_area=1 confidence=1 /> #why two shape complementarity filters?

<OligomericAverageDegree name=avg_deg threshold=8 distance_threshold=10.0 write2pdb=1 task_operations=design_task /> #Degree of connectivity of a subset of residues

<Ddg name=ddG_filt scorefxn=sfx_vanilla jump=1 repack=1 repeats=3 threshold=-18 confidence=1 />

<BuriedUnsatHbonds name=buriedUnsatBonds scorefxn=sfx_vanilla jump_number=1 cutoff=9 confidence=0/>

<RotamerBoltzmannWeight name=rotboltz task_operations=design_task radius=6.0 jump=1 unbound=1 scorefxn=sfx_vanilla temperature=0.8 repack=1 skip_ala_scan=1 write2pdb=1 />

<Sasa name=sasa_int_area threshold=1500 upper_threshold=10000 hydrophobic=0 polar=0 jump=1 confidence=1 />

<TotalSasa name=boundSASA confidence=0 task_operations=no_glycan/> #no_glycan bound complex SASA

<MoveBeforeFilter name=unboundSASA mover=unbound filter=boundSASA/> #no_glycan unbound SASA, difference should be interface SASA

<ScoreType name=total_score scorefxn=sfx_vanilla score_type=total_score threshold=0 confidence=0/>

<Sasa name=sasa_hydro threshold=1000 upper_threshold=10000 hydrophobic=1 polar=0 jump=1 confidence=0 />

<Sasa name=sasa_polar threshold=1000 upper_threshold=10000 hydrophobic=0 polar=1 jump=1 confidence=0 />

<PackStat name="packstat" repeats="5" threshold="0.60" confidence="0"/>

</FILTERS>

<MOVERS>

<MotifGraft name="motif_grafting" context_structure="%%CONTEXT%%" motif_structure="%%MOTIF%%" RMSD_tolerance="1.0" NC_points_RMSD_tolerance="1.0" clash_score_cutoff="10" clash_test_residue="ALA" combinatory_fragment_size_delta="2:2" max_fragment_replacement_size_delta="0:0" full_motif_bb_alignment="1" allow_independent_alignment_per_fragment="0" graft_only_hotspots_by_replacement="0" only_allow_if_N_point_match_aa_identity="0" only_allow_if_C_point_match_aa_identity="0" revert_graft_to_native_sequence="1" allow_repeat_same_graft_output="0"/>

<StoreTaskMover name=store_design_task task_name="design_task" task_operations=interface_only,limitaro,no_repack_target,no_CYS />

<PackRotamersMover name=design_soft scorefxn=sfx_vanilla task_operations=design_task,core,dsgn_aa_core />

<PackRotamersMover name=design_hard scorefxn=sfx_vanilla task_operations=design_task,core,dsgn_aa_core />

<PackRotamersMover name=design_elec scorefxn=sfx_vanilla task_operations=design_task,surf,dsgn_aa_surf />

<PackRotamersMover name=repack scorefxn=sfx_vanilla task_operations=design_task,repack_only />

<TaskAwareMinMover name=min0 scorefxn=sfx_vanilla bb=0 chi=1 jump=0 task_operations=design_task />

<TaskAwareMinMover name=min scorefxn=sfx_vanilla bb=0 chi=1 jump=1 task_operations=design_task />

<TaskAwareMinMover name=minc scorefxn=sfx_vanilla bb=0 chi=1 jump=0 task_operations=design_task,core />

<TaskAwareMinMover name=mins scorefxn=sfx_vanilla bb=0 chi=1 jump=0 task_operations=design_task,surf />

<TaskAwareMinMover name=min1 scorefxn=sfx_vanilla bb=1 chi=1 jump=0 task_operations=design_task />

<ParsedProtocol name=design_min_soft>

```
            <Add mover=design_soft />
            <Add mover=minc />
        </ParsedProtocol>
        <ParsedProtocol name=design_min_hard>
            <Add mover=design_hard />
            <Add mover=minc />
        </ParsedProtocol>
          <ParsedProtocol name=design_core>
            <add mover_name=min0 />
            <add mover_name=design_min_soft />
            <add mover_name=design_min_hard />
          </ParsedProtocol>
          <ParsedProtocol name=design_surf>
            <add mover_name=design_elec />
            <add mover_name=mins />
          </ParsedProtocol>
        <ParsedProtocol name=min_repack_min>
            <Add mover=min />
            <Add mover=repack />
            <Add mover=min />
        </ParsedProtocol>


            <TaskAwareMinMover name=bbrb_min_all_soft scorefxn=sfx_vanilla bb=1
chi=1 jump=1 cartesian=0 task_operations=interface_only />
            <TaskAwareMinMover name=bbrb_min_all_hard scorefxn=sfx_vanilla bb=1
chi=1 jump=1 cartesian=0 task_operations=interface_only/>
            <TaskAwareMinMover name=bbrb_min_all_hard_void scorefxn=sfx_vanilla
bb=1 chi=1 jump=1 cartesian=0 task_operations=interface_only/>
            <TaskAwareMinMover name=bbrb_min_all_vanilla scorefxn=sfx_cart bb=1
chi=1 jump=1 cartesian=1 task_operations=interface_only/>
        </MOVERS>
```

```
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
                <Add mover_name="motif_grafting"/>
                <Add mover_name=store_design_task />
                <Add mover_name=design_soft />   <Add mover_name=bbrb_min_all_soft />
                <Add mover_name=design_soft />   <Add mover_name=bbrb_min_all_hard />
                <Add mover_name=design_hard />  <Add mover_name=bbrb_min_all_hard />
                <Add mover_name=design_elec />
                <Add mover_name=bbrb_min_all_vanilla />
                <Add filter_name=sc_filt0 />
                <Add mover_name=bbrb_min_all_hard_void />
                <Add filter_name=sc_filt1 />
                <Add filter_name=ddG_filt />
                <Add filter_name=buriedUnsatBonds />
                <Add filter_name=sasa_int_area />
                <Add filter_name=boundSASA />
                <Add filter_name=unboundSASA />
                <Add filter_name=rotboltz />
                <Add filter_name=sasa_hydro/>
                <Add filter_name=sasa_polar/>
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

Online repository

https://drive.google.com/open?id=0Bz97EhooEr7XYUdzdTktOXFma0E&authuser=0

This repository contains:

complete list of initial 12,383 designs ordered in excel format:

18March14_Array_synthesis_order.xlsx

raw deep sequencing analysis after initial screen in excel format:

Array_synthsis_miseq_analysis.xlsx

-raw deep sequencing analysis after SSM screen in excel format:

Flu_SSM_Array_miseq_analysis.xlsx

heatmaps from SSM analysis in powerpoint format:

flu_ssm_heatmaps.pptx

References

1. Fleishman, S. J. *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science (New York, N.Y.)* **332,** 816–21 (2011).

2. Whitehead, T. a *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature biotechnology* **30,** 543–8 (2012).

3. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature methods* **7,** 741–6 (2010).

4. Araya, C. L. & Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends in biotechnology* **29,** 435–42 (2011).

5. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nature protocols* **1,** 755–68 (2006).

6. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science (New York, N.Y.)* **244,** 1081–5 (1989).

7. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science (New York, N.Y.)* **247,** 1306–10 (1990).

8. Pál, G., Kouadio, J.-L. K., Artis, D. R., Kossiakoff, A. A. & Sidhu, S. S. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *The Journal of biological chemistry* **281,** 22378–85 (2006).

9. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness – epistasis link shapes the fitness landscape of a randomly drifting protein. **444,** 929–932 (2006).

10. Fleishman, S. J. *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science (New York, N.Y.)* **332,** 816–21 (2011).

11. Fleishman, S. J. *et al.* RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PloS one* **6,** e20161 (2011).

12. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* **487,** 545–74 (2011).

13. Joughin, B. A., Green, D. F. & Tidor, B. Action-at-a-distance interactions enhance protein binding affinity. *Protein science : a publication of the Protein Society* **14,** 1363–9 (2005).

14. Marshall, S. A., Vizcarra, C. L. & Mayo, S. L. One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. *Protein science : a publication of the Protein Society* **14,** 1293–304 (2005).

15. Dutta, S. *et al.* Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. *Journal of molecular biology* **398,** 747–62 (2010).

16. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79,** 1061–78 (2011).

17. Ekiert, D. C. *et al.* Antibody recognition of a highly conserved influenza virus epitope. *Science (New York, N.Y.)* **324,** 246–51 (2009).

18. Sui, J. *et al.* Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nature structural & molecular biology* **16,** 265–73 (2009).

19. Throsby, M. *et al.* Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. *PloS one* **3,** e3942 (2008).

20. Corti, D. *et al.* A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science (New York, N.Y.)* **333,** 850–6 (2011).

21. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 7896–901 (2011).

22. Pitt, J. N. & Ferré-D'Amaré, A. R. Rapid construction of empirical RNA fitness landscapes. *Science (New York, N.Y.)* **330,** 376–9 (2010).

23. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature biotechnology* **27,** 1173–5 (2009).

24. Shultzaberger, R. K., Malashock, D. S., Kirsch, J. F. & Eisen, M. B. The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. *PLoS genetics* **6,** e1001042 (2010).

25. Wu, X. *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science (New York, N.Y.)* **333,** 1593–602 (2011).

26. Chao, G., Cochran, J. R. & Wittrup, K. D. Fine epitope mapping of anti-epidermal growth factor receptor antibodies through random mutagenesis and yeast surface display. *Journal of molecular biology* **342,** 539–50 (2004).

27. Kunkel, T. a, Roberts, J. D. & Zakour, R. a. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods in enzymology* **154,** 367–82 (1987).

28. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least Angle Regression. 1–44 (2003).

29. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C.-M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein engineering, design & selection : PEDS* **23,** 155–9 (2010).

30. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. **41,** 207–234 (2005).

31. Rohl, C. a, Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods in enzymology* **383,** 66–93 (2004).

32. Sitkoff, D., Ben-Tal, N. & Honig, B. Calculation of Alkane to Water Solvation Free Energies Using Continuum Solvent Models. *The Journal of Physical Chemistry* **100,** 2744–2752 (1996).

33. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79,** 830–8 (2011).

34. Sitkoff, D., Sharp, K. A. & Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *The Journal of Physical Chemistry* **98,** 1978–1988 (1994).

35. Richards, F. M. Areas, volumes, packing and protein structure. *Annual review of biophysics and bioengineering* **6,** 151–76 (1977).

36. McCoy, A. J. *et al.* Phaser crystallographic software. *Journal of applied crystallography* **40,** 658–674 (2007).

37. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* **66,** 213–21 (2010).

38. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* **66,** 486–501 (2010).

39. Strong, M. *et al.* Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 8060–5 (2006).

40. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology* **238,** 777–93 (1994).

41. Sheriff, S., Hendrickson, W. A. & Smith, J. L. Structure of myohemerythrin in the azidomet state at 1.7/1.3 A resolution. *Journal of molecular biology* **197,** 273–96 (1987).

42. Delano, W. L. *The PyMOL molecular graphics system*. (2002).

43. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica. Section D, Biological crystallography* **66,** 12–21 (2010).

44. Nguyen, J. T. *et al.* Triple combination of oseltamivir, amantadine, and ribavirin displays synergistic activity against multiple influenza virus strains in vitro. *Antimicrobial agents and chemotherapy* **53,** 4115–26 (2009).

45. Smee, D. F., Huffman, J. H., Morrison, A. C., Barnard, D. L. & Sidwell, R. W. Cyclopentane neuraminidase inhibitors with potent in vitro anti-influenza virus activities. *Antimicrobial agents and chemotherapy* **45,** 743–8 (2001).

46. Nguyen, J. T. *et al.* Triple combination of amantadine, ribavirin, and oseltamivir is highly active and synergistic against drug resistant influenza virus strains in vitro. *PloS one* **5,** e9332 (2010).

47. Schreiber, G. & Fersht, A. R. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *Journal of molecular biology* **248,** 478–86 (1995).

48. Schreiber, G. & Fersht, A. R. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry* **32,** 5145–5150 (1993).

49. Buckle, A. M., Schreiber, G. & Fersht, A. R. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-.ANG. resolution. *Biochemistry* **33,** 8878–8889 (1994).

50. Schreiber, G., Frisch, C. & Fersht, A. R. The role of Glu73 of barnase in catalysis and the binding of barstar. *Journal of molecular biology* **270,** 111–22 (1997).

51. Weiss, M. S. & Hilgenfeld, R. On the use of the merging R factor as a quality indicator for X-ray data. *Journal of Applied Crystallography* **30,** 203–205 (1997).

52. Murphy, S. L., Xu, J., Kochanek, K. D. & Statistics, V. *National Vital Statistics Reports Deaths : Final Data for 2010*. **61,** (2013).

53. Thompson, W. W. *et al.* Mortality Associated With Influenza and Respiratory Syncytial Virus in the United States. *JAMA* **289,** 179–186 (2013).

54. Johnson, N. P. a. S. & Mueller, J. Updating the Accounts: Global Mortality of the 1918-1920 “Spanish” Influenza Pandemic. *Bulletin of the History of Medicine* **76,** 105–115 (2002).

55. Dawood, F. S. *et al.* Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet infectious diseases* **12,** 687–95 (2012).

56. Viboud, C. & Simonsen, L. Global mortality of 2009 pandemic influenza A H1N1. *The Lancet infectious diseases* **12,** 651–3 (2012).

57. Burch, J. *et al.* Prescription of anti-influenza drugs for healthy adults: a systematic review and meta-analysis. *The Lancet. Infectious diseases* **9,** 537–45 (2009).

58. Wright P. F., Neumann G., K. Y. . *Orthomyxoviruses*. 536–549 (2006).

59. Gerdil, C. The annual production cycle for influenza vaccine. *Vaccine* **21,** 1776–1779 (2003).

60. Webster, R. G., Sharp, G. B. & Claas, E. C. Interspecies transmission of influenza viruses. *American journal of respiratory and critical care medicine* **152,** S25–30 (1995).

61. Fiore, M., Fry, A. & Shay, D. Antiviral agents for the treatment and chemoprophylaxis of influenza. *Centers for Disease …* **60,** 28 (2011).

62. Hu, Y. *et al.* Association between adverse clinical outcome in human disease caused by novel infl uenza A H7N9 virus and sustained viral shedding and emergence of antiviral resistance Yunwen. *The Lancet* **381,** 2273–2279

63. Moscona, A. Oseltamivir resistance--disabling our influenza defenses. *The New England journal of medicine* **353,** 2633–6 (2005).

64. Ekiert, D. C. *et al.* A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science (New York, N.Y.)* **333,** 843–50 (2011).

65. Lingwood, D. *et al.* Structural and genetic basis for development of broadly neutralizing influenza antibodies. *Nature* **489,** 566–70 (2012).

66. Whittle, J. R. R. *et al.* Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 14216–21 (2011).

67. Magadán, J. G. *et al.* Biogenesis of influenza a virus hemagglutinin cross-protective stem epitopes. *PLoS pathogens* **10,** e1004204 (2014).

68. Hansel, T. T., Kropshofer, H., Singer, T., Mitchell, J. A. & George, A. J. T. The safety and side effects of monoclonal antibodies. *Nature reviews. Drug discovery* **9,** 325–38 (2010).

69. Chames, P., Van Regenmortel, M., Weiss, E. & Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *British journal of pharmacology* **157,** 220–33 (2009).

70. Perchiacca, J. M. & Tessier, P. M. Engineering aggregation-resistant antibodies. *Annual review of chemical and biomolecular engineering* **3,** 263–86 (2012).

71. Tabrizi, M., Bornstein, G. G. & Suria, H. Biodistribution mechanisms of therapeutic monoclonal antibodies in health and disease. *The AAPS journal* **12,** 33–43 (2010).

72. Daugherty, A. L. & Mrsny, R. J. Formulation and delivery issues for monoclonal antibody therapeutics. *Advanced drug delivery reviews* **58,** 686–706 (2006).

73. Wang, W., Wang, E. Q. & Balthasar, J. P. Monoclonal antibody pharmacokinetics and pharmacodynamics. *Clinical pharmacology and therapeutics* **84,** 548–58 (2008).

74. Harding, F. A., Stickler, M. M., Razo, J. & DuBridge, R. B. The immunogenicity of humanized and fully human antibodies: residual immunogenicity resides in the CDR regions. *mAbs* **2,** 256–65

75. Procko, E. *et al.* A computationally designed inhibitor of an epstein-barr viral bcl-2 protein induces apoptosis in infected cells. *Cell* **157,** 1644–56 (2014).

76. Dahiyat, B. I. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **278,** 82–87 (1997).

77. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–227 (2012).

78. Goldstein, S., Pheasant, D. & Miller, C. The charybdotoxin receptor of a Shaker K+ channel: Peptide and channel residues mediating molecular recognition. *Neuron* **12,** 1377–1388 (1994).

79. Craik, D., Daly, N. & Waine, C. The cystine knot motif in toxins and implications for drug design. *Toxicon* **39,** (2001).

80. Werle, M., Kafedjiiski, K., Kolmar, H. & Bernkop-Schnürch, a. Evaluation and improvement of the properties of the novel cystine-knot microprotein McoEeTI for oral administration. *International journal of pharmaceutics* **332,** 72–9 (2007).

81. Lavergne, V. Cysteine-rich mini-proteins in human biology. *Current topics in medicinal* ... 1514–1533 (2012). at <http://www.ingentaconnect.com/content/ben/ctmc/2012/00000012/00000014/art00005>

82. Silverman, A. P., Levin, A. M., Lahti, J. L. & Cochran, J. R. Engineered cystine-knot peptides that bind alpha(v)beta(3) integrin with antibody-like affinities. *Journal of molecular biology* **385,** 1064–75 (2009).

83. Kimura, R. H., Cheng, Z., Gambhir, S. S. & Cochran, J. R. Engineered knottin peptides: a new class of agents for imaging integrin expression in living subjects. *Cancer research* **69,** 2435–42 (2009).

84. Molinski, T. F., Dalisay, D. S., Lievens, S. L. & Saludes, J. P. Drug development from marine natural products. *Nature reviews. Drug discovery* **8,** 69–85 (2009).

85. Venkatraman, J. Design of folded peptides. *Chemical reviews* 69–85 (2001). at <http://pubs.acs.org/doi/pdf/10.1021/cr000053z>

86. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, N.Y.)* **302,** 1364–8 (2003).

87. Chen, R. & Chung, S.-H. Molecular dynamics simulations of scorpion toxin recognition by the Ca(2+)-activated potassium channel KCa3.1. *Biophysical journal* **105,** 1829–37 (2013).

88. Jaravine, V., Nolde, D. & Reibarkh, M. Three-dimensional structure of toxin OSK1 from Orthochirus scrobiculosus scorpion venom. *Biochemistry* **2960,** 1223–1232 (1997).

89. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6,** 343–5 (2009).

90. Orr-Weaver, T. L., Szostak, J. W. & Rothstein, R. J. Yeast transformation: a model system for the study of recombination. *Proceedings of the National Academy of Sciences of the United States of America* **78,** 6354–8 (1981).

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

92. Chetverin, A. B. & Kramer, F. R. Oligonucleotide arrays: new concepts and possibilities. *Bio/technology (Nature Publishing Company)* **12,** 1093–9 (1994).

93. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nature biotechnology* **28,** 1295–9 (2010).

94. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic acids research* **38,** 2522–40 (2010).

95. Borovkov, A. Y. *et al.* High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic acids research* **38,** e180 (2010).

96. Schwartz, J. J., Lee, C. & Shendure, J. Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nature methods* **9,** 913–5 (2012).

97. Quan, J. *et al.* Parallel on-chip gene synthesis and application to optimization of protein expression. *Nature biotechnology* **29,** 449–52 (2011).

98. King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science (New York, N.Y.)* **336,** 1171–4 (2012).

99. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510,** 103–8 (2014).

100. Dreyfus, C., Ekiert, D. C. & Wilson, I. A. Structure of a classical broadly neutralizing stem antibody in complex with a pandemic H2 influenza virus hemagglutinin. *Journal of virology* **87,** 7149–54 (2013).

101. Corti, D. *et al.* A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science (New York, N.Y.)* **333,** 850–6 (2011).

102. Dreyfus, C. *et al.* Highly conserved protective epitopes on influenza B viruses. *Science (New York, N.Y.)* **337,** 1343–8 (2012).

103. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PloS one* **6,** e24109 (2011).

104. Azoitei, M. L. *et al.* Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science (New York, N.Y.)* **334,** 373–6 (2011).

105. VanAntwerp, J. J. & Wittrup, K. D. Fine affinity discrimination by yeast surface display and flow cytometry. *Biotechnology progress* **16,** 31–7 (2000).

106. Qian, Z.-G., Xia, X.-X., Choi, J. H. & Lee, S. Y. Proteome-based identification of fusion partner for high-level extracellular production of recombinant proteins in Escherichia coli. *Biotechnology and bioengineering* **101,** 587–601 (2008).

107. Marblestone, J. G. *et al.* Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein science : a publication of the Protein Society* **15,** 182–9 (2006).

108. Bahl, C. D., MacEachran, D. P., O'Toole, G. A. & Madden, D. R. Purification, crystallization and preliminary X-ray diffraction analysis of Cif, a virulence factor secreted by Pseudomonas aeruginosa. *Acta crystallographica. Section F, Structural biology and crystallization communications* **66,** 26–8 (2010).

109. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods* **11,** 801–7 (2014).

110. Hay, A. J. & Hayden, F. G. Oseltamivir resistance during treatment of H7N9 infection. *Lancet* **381,** 2230–2 (2013).

111. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein science : a publication of the Protein Society* **4,** 2411–23 (1995).

112. Noble, J. E., Knight, A. E., Reason, A. J., Di Matola, A. & Bailey, M. J. A. A comparison of protein quantitation assays for biopharmaceutical applications. *Molecular biotechnology* **37,** 99–111 (2007).

113. Mir-Shekari, S. Y., Ashford, D. A., Harvey, D. J., Dwek, R. A. & Schulze, I. T. The glycosylation of the influenza A virus hemagglutinin by mammalian cells. A site-specific study. *The Journal of biological chemistry* **272,** 4027–36 (1997).

114. De Vries, R. P. *et al.* The influenza A virus hemagglutinin glycosylation state affects receptor-binding specificity. *Virology* **403,** 17–25 (2010).