Directed evolution and *de novo* design for improved pathogen-targeting protein drugs

Jorgen W Nelson

A dissertation

Submitted in partial fulfillment of the

Requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

David Baker, Chair

Douglas M Fowler

Jay Shendure

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Directed evolution and de novo *design for improved pathogen-targeting protein drugs*

Jorgen W Nelson

Chair of Supervisory Committee

Professor David Baker

Biochemistry

Infectious diseases continue to claim millions of lives, and protein design with Rosetta is quickly becoming a contributor to the fight against these diseases. My dissertation has focused on leveraging recently developed high-throughput synthesis and screening technologies to improve existing designed proteins for use as pathogen-targeting drugs as well as to develop new design methods for infectious disease targets. First, I present my effort to develop techniques to improve two existing influenza-targeting proteins into viable protein drugs, in terms of higher stability and stronger influenza neutralization and how, with collaborators, I developed one of these improved proteins into the first designed protein to ever cure an infectious disease in an animal model. Second, I present my efforts to develop a new protein design method that incorporates existing known structural motifs into fully *de novo* protein scaffolds at massive scale, and use this method to test the hypothesis that structure-based design can stabilize a malaria surface epitope in order to build a better malaria vaccine. It is hoped that these contributions will bring designed protein drugs closer to pharmaceutical relevance and help reduce the burden of infectious disease worldwide.

Table of Contents

Acknowledgements

I would primarily like to thank the many research partners and collaborators that I have had over the course of my graduate work. Christy Tinberg and Aaron Chevalier guided me through the early years in the lab, and for that I will always remain grateful.

The Rosetta macromolecular modeling suite was critical for much of the work I've done, and I thank both the development team and those who have helped me learn how to use it. Specifically helpful for learning Rosetta and other computational work were Daniel-Adriano Silva, Lucas Nivon, Vikram Mulligan, Brian Coventry, Brian Weitzner, TJ Brunette, Jeremy Mills, Tom Linsky, Will Sheffler, Possu Huang, Rocco Moretti, Fabio Parmeggiani, Eva-Maria Strauch, Gustav Oberdorfer, Alex Ford, Yang Hsia, Neil King, Scott Boyken, Brian Koepnick and Derrick Hicks.

I have also learned a vast number of new lab techniques since coming to the Baker lab, and for particular help with experimental work I'd like to thank Lauren Carter, Rashmi Ravichandran, Yang Hsia, Jeremy Mills, Una Natterman, Jiayi Dou, Gabe Rocklin, Matt Bick, Anindya Roy, Inna Goreshnik, Stephen Rettie, Fabio Parmeggiani, Gabe Butterfield, Cassie Bryan, Eva-Maria Strauch, Kara Lau, Jacob Bale, Franziska Seeger, and Jack Montana Lalonde.

The administrative wing of the lab has also been extremely helpful in keeping everything running smoothly and ensuring that I've never felt financially constrained. Thank you to Lance Stewart, Ratika Krishnamurty, Michelle Matsnuaga, and Mike Murphy.

One of the greatest perks of working in the Baker lab is that sometimes it feels like the entire world is eager to work with you. I have had many excellent external collaborators over the years, and would particularly like to thank Caitlin Anderson, Carly Holstein, and Paul Yeager,

Jason Klein and Jay Shendure, Connor Weidle and Marie Pancera, Peter Lee and Ian Wilson, Joe Francica, Azza Idris, and Bob Seder, and Julia Frei, Ariel Wirchnianski, and Jon Lai.

I thank my thesis committee members Doug Fowler, Jay Shendure, and Ning Zheng for taking time out of their schedules to mentor me and to read this document.

My advisor David Baker has been wonderfully patient and supportive over the last six years in the lab. He has never discouraged me from pursuing my own ideas wherever they might lead, and I don't know if I've ever met anyone else who manages to combine the kind of burning passion that David has for our work with such an easygoing manner and true generosity of spirit.

For helping to keep me relatively sane through this process I'd like to credit my hetero life-mate Gordon Ross, Distractions, Inc., and coffee.

Finally I thank my family- in particular my parents David and Claudia who have been highly supportive of me over the years, up to and including helping to proofread this dissertation, my brother Erik, the only one of us four sensible enough to *not* go to PhD school, and my grandmother Marlys, who got her wish that I graduate "before [she] die[s]."

Introduction

In the past fifteen years, the ability to design and test *de novo* proteins created by the Rosetta molecular modeling suite has progressed from a single example, Top7[1] to the large-scale generation of thousands of hyperstable proteins[2]. At the same time, rational design of protein-protein interactions has progressed from making small numbers of mutations to switch the affinity of a natural interaction[3] to being able to design[4], build[5], and test[6], in parallel, thousands of putative binding proteins. All of this has been possible because of, over a relatively short period of time, order-of-magnitude increases in the capabilities of key enabling technologies like high-throughput DNA sequencing, low-cost DNA synthesis, high-performance computing, and fluorescence-activated cell sorting (FACS).

Hand-in-hand with the rapid improvement in protein design capabilities, the field has broadened its horizons as far as the practical problem space that could be addressed with designed proteins. The protein design community has evolved from considering only proof-of-principle demonstrations of the technology to taking seriously the idea that rationally designed proteins might soon have a role to play in treating some of mankind's most recalcitrant diseases and maybe even stemming the decline in modern pharmaceutical innovation[7]. Given these newly broadened ambitions and the general need for more drugs coming down the pipeline, it has become important to ask how new high-throughput methods could be used to improve the drug-like properties of existing designed proteins, and to improve the design process to address pressing global health needs.

Early iterations of designed proteins that sought to treat infectious diseases encountered difficulties with protein binding strength, as well as stability. The first series of experiments described in this dissertation sought a means to take existing designed protein binders and use

recently developed high-throughput sequencing techniques to improve these "drug-like" properties, as well as the first ever attempts to treat an infectious disease in an animal model with a Rosetta-designed protein. As the technology of protein design improved, new methods allowed more difficult infectious disease targets to be addressed, using fully *de novo* proteins that incorporate the design principles of good drug-like molecules from the beginning of the process. In the second set of experiments described, I will introduce a new protein design methodology that should allow many new pharmaceutical targets to be addressed, as well as its application to a specific structural hypothesis that may allow for the design of a new generation of better malaria vaccines.

Section 1: Improving the drug-like properties of designed influenza binders

Designed influenza binders as potential drug candidates

Influenza poses a significant public health threat each year and pandemics, such as the 2009 H1N1 outbreak, are inevitable. The influenza hemagglutinin antigen (HA) is a glycoprotein on the surface of the influenza virus consisting of a highly variable globular head domain (HA1) and a stalk domain (HA2) that is more conserved between influenza strains[8]. Genetically, influenza viruses comprise 2 phylogenetic groups (Groups 1 and 2) consisting of 16 influenza subtypes and numerous drifted and shifted genetic variants or strains within each subtype. Vaccination can prevent influenza infection but current vaccines induce antibodies primarily against the highly variable immunodominant HA1 region that undergoes antigenic drift each year so annual vaccination is strain specific, providing little or no protection against drifted strains within the same subtype or shifted strains that lead to pandemics[9,10]. Influenza vaccines are also less effective in certain populations, such as the elderly, infants or immune-compromised individuals. For this reason, antivirals that can be self-administered by either intranasal or oral routes have been marketed to treat or reduce influenza disease. However, recent studies show that currently marketed influenza antivirals provide little or no efficacy, incur high influenza resistance rates and cause significant adverse effects[11]. New antivirals that can broadly protect against a wide range of influenza variants are urgently needed to supplement the protective effects of vaccines and improve the population's resistance against seasonal influenza and future pandemics. Monoclonal antibodies (mAbs) that bind the conserved HA stalk and potently neutralize diverse influenzas have been produced but due, in part, to the high cost for their production, they are primarily being developed for intravenous treatment of the most severe, individual influenza cases in hospital settings[12,13].

In pursuit of these goal, in 2011 the first designed proteins were described that could bind to a key area of the stem region of the influenza hemagglutinin protein (**Figure 1.1a**). Binding to this region by antibodies was known to lead to viral inactivation in cell culture models[14]. The two proteins described in that study, HB80 and HB36, were relatively weak binders when compared to the antibodies by which they were inspired, but further experimentation with directed evolution created a second generation which showed stronger binding potency and the ability to neutralize viruses *in vitro*[15]. However, when these improved proteins were tested *in vivo* by intravenous injection into mice, they showed no therapeutic efficacy (data not shown). From this we concluded that the proteins were either not strong enough binders, or some other consideration, like stability, or the lack of antibody-like Fc-mediated immune system recruitment, was preventing them from exerting a therapeutic effect.

A viable protein antiviral must, in addition to broad and potent binding to the target of interest, be highly stable, both *in vivo* and *in vitro*, to ensure that it is robust to storage and transport to point-of-care and that it can perform its task without being degraded once in the body.[16] It also must maintain this stability without triggering an immune response against itself. We also predict that, based on previous research, highly stabilized protein scaffolds will allow a greater range of functions and higher affinity protein-protein interactions. It is therefore critical to understand and determine stability in designed proteins and how to improve engineered protein stability throughout the design process.

In the following study, we describe the use of recently developed deep mutational scanning (DMS) techniques to improve the stability of HB80 and the binding capabilities of HB36, as well as successful use of the improved HB36 as an *in vivo* protein therapeutic.

Deep mutational scanning to assess stability-improving mutations

The region of HA to which HB80 and HB36 bind is known to be necessary for influenza infection, and several neutralizing antibodies have been described that bind to this region.[17] HB80 and HB36 have both shown binding to intact flu viruses and reduction of virus infection in cell culture. The two proteins are a good choice of model system because though they both have very similar binding interfaces with identical 'hotspot' residues and similar binding affinities to HA, they have very different stabilities, with temperatures of unfolding (T_m) different by over 20°C. (**Figure 1.1b**).

The most common means of experimentally searching for stabilized mutants of a protein is directed evolution: creating a large library of mutations and then selecting them functionally until the library has converged to one or a few clones. Directed evolution has the advantage of being agnostic to assumptions about which residues might be important, and has been applied to stabilize several engineered scaffolds.[18,19] Traditional directed evolution approaches, however, have significant limitations. In general, directed evolution experiments on a mutation library will converge to one or a few single clones, which may or may not represent the absolute best clone, due to the fact that a combinatorial library can reach immense sizes with just a handful of mutations, meaning that many possible combinations might be untested, providing only limited stability landscape data.[20]

Recently, much more comprehensive new approaches based on deep sequencing have allowed the evaluation of the fitness effects of every point mutation in a protein simultaneously by comparing the abundances of each mutation before and after functional selection, using DNA sequencing. [6,21] This technique, deep mutational scanning (DMS), has been used extensively in recent years to comprehensively assess fitness landscapes in both engineered and natural proteins, where it has been found that highly-enriched mutations can, for the most part, be

combined to enhance function additively.[15,22,23] The DMS technique has also been used to assess entire protein stability landscapes specifically, albeit by various indirect means.[24,25] These techniques have their own limitations, however, being variously applicable to only very small proteins or only large, multi-domain proteins that are irreversibly unfolding. As an initial step towards better understanding and methods for stability improvement in designed proteins, we have developed a DMS-based approach that explicitly measures the stability effects of mutations in any selectable protein without regard to the protein's size or kinetics or that mutation's effects on the protein's function. We have tested its efficacy on the previously described HA binder, HB80. The development of this approach allows us to avoid the sparse data provided by directed evolution, and will also overcome some of the limitations of previous indirect DMS-based methods for comprehensive stability landscape measurements of other designed proteins.

Mutagenesis, yeast display, and Illumina sequencing

We subjected plasmids encoding the designed hemagglutinin-binding protein HB80.4 to a site-saturation mutagenesis protocol based on the method of Kunkel,[26] yielding a library of plasmids in which each codon was singly substituted to each other possible codon, yielding 1060 individual mutants. Using yeast display and fluorescence-activated cell sorting (FACS), we selected subpopulations of the mutant library that bound strongest to hemagglutinin subtype SC1918/H1 in five replicate libraries at five different temperatures, representing a range around the temperature at which purified HB80.4 is half-unfolded by circular dichroism spectroscopy (T_m). We extracted plasmid DNA from the unselected population and the five selected populations, and the HB80.4 mutant genes were sequenced using 250bp paired-end Illumina sequencing on a MiSeq instrument. Sequencing results of the unselected library indicates that 955 mutants are present, with an average coverage of 243 reads.

Functional score calculation

Similar to previous approaches, comparing the frequencies of each individual mutant in the selected libraries against the same mutant in the unselected library allows the calculation of an enrichment value, which represents how favored or disfavored that mutation is. An innovation in our analysis is that instead of using a simple enrichment value, we have created a functional score that is calculated as the slope of a mutation's log2 enrichment or depletion as the temperature of the yeast display step increases. We hypothesize that with an increase in temperature, affinity effects due to improved or disrupted binding site complementarity should stay constant, while affinity effects due to improved or decreased binder stability should be exacerbated as the temperature of the selection rises and more stable mutants come to dominate the binding population. Calculated functional scores vary from -1.40 to 1.21 relative to wild type, with a mean across all mutants of -0.11, suggesting that most mutations are mildly destabilizing, by this measure (**Figure 1.2a**).

High-scoring mutant testing

In order to evaluate whether the functional score is able to successfully identify stabilizing mutations, we picked 7 individual high-scoring mutations to test functionally. Each mutation chosen had the highest functional score of codon in its position in the gene. Mutant plasmids were produced by Kunkel's method, and mutant proteins were expressed in and purified from E. coli. Mutants were tested for their unfolding temperature using circular dichroism (CD) spectroscopy, and their protease sensitivity was tested by digestion with porcine pancreatic trypsin and mouse serum. All seven high-scoring mutants tested showed superior stability by at least one measure, and most were more stable by all three measures (**Figure 1.2b**). To ensure that our assay achieved its goal of identifying stabilizing mutants that do not disrupt

binding, we also tested all seven mutants by yeast display and FACS sorting for the proteins' HA-binding dissociation constants, which were all found to bind HA as well as the non-stabilized protein.

Cross-strain optimization of HB36.5

The previously described designed HA binding protein, HB36.5, is stable, as measured by thermal denaturation and protease resistance (**Figure 2.1**). These are desirable properties in a protein therapeutic, but HB36.5 does not bind tightly to many HA subtypes, another desirable drug property, since a good influenza therapeutic should not need to be seasonally reformulated, but rather available on hand for both seasonal and pandemic flu strains. To improve binding affinity against multiple HAs in parallel, we performed deep mutational scanning of the HB36.5 protein against seven different Group 1 HA strains. Because of the specific orientation in which HB36.5 binds, binding to most Group 2 HAs is probably impossible due to an occluding stem-region glycosylation that they share. We constructed a site-saturation mutagenesis library of HB36.5 using Kunkel mutagenesis, in which each amino acid was individually mutated to each other possible amino acid. This library was transformed into surface display-competent yeast, and we then sorted the library in two rounds of yeast display and FACS at increasing temperatures for binding to each different HA strain individually.

All of these libraries, one unselected and fourteen selected, were processed for plasmid DNA and sequenced on an Illumina Miseq platform. Sequencing yielded abundances of each substitution in each library, and enrichment values were calculated by taking the slope of the log2 enrichment in substitution abundances across both rounds of sorting. (**Figure 1.3b)** A positive slope indicates positive functional selection, while a negative slope indicates negative selection. For the most part, mutations in the HA-HB36.5 binding interface were highly

conserved. However, some positions, generally not within the binding interface, have high average enrichment values across all seven subtypes, indicating that substitutions in these positions improve binding against many HA subtypes. (**Figure 1.3a**) Several substitutions at the binding interface favor only one or several HA subtypes. (**Figure 1.3c**) To improve binding to many different HA subtypes, we made a combinatorial library of HB36.5 mutations that were positively selected across all subtypes, over a total of 12 mutated positions. We performed three rounds of sorting by yeast display for binding to A/South Carolina/1/1918 HA until the library converged. This new protein, which we call HB36.6 has 9 total substitutions relative to HB36.5.

Biochemical characterization of HB36.6

To confirm that this strategy resulted in broader binding, we compared HB36.5 and HB36.6 for equilibrium binding affinity against 6 HA subtypes by biolayer interferometry (BLI). These experiments revealed that HB36.6 had improved affinity against all subtypes, some dramatically (**Figure 1.4d**), with greater than 40-fold and 10-fold affinity increases for H2 and H5, respectively. Structural studies by cryo-electron microscopy revealed that HB36.6 retains binding in the designed location of the stem region of HA (**Figure 1.4a-c**). This means that the optimization process likely did not change the ability of HB36.6 to inhibit the conformational change in HA that leads to virus neutralization. HB36.6 is both a very strong HA binder and a proteolytically stable protein, so we hypothesized that if administered mucosally in an *in vivo* infection model, it might persist in the lung long enough to show therapeutic or prophylactic effects against multiple influenza strains.

HB36.6 provides prophylactic and therapeutic protection from influenza *in vivo*

To determine if HB36.6 can protect against influenza infection *in vivo*, we gave Balb/c mice a single intranasal (IN) dose of 6.0 mg HB36.6/kg body-weight mouse at 2, 24, 48 or 72

hours prior to challenge with a lethal dose (10 times the 50% mouse lethal dose or 10 MLD_{50}) of H1N1 A/California/07/2009 (CA09) virus, a highly virulent group 1 pandemic influenza strain that leads to rapid weight loss and death in mice within 4-7 days post-infection (p.i.). When administered up to 48 hours before challenge, a single pre-exposure dose of HB36.6 afforded complete protection with 100% survival and no weight loss whereas all untreated controls (Ctr) exhibited >30% weight loss and no survival (**Figure 1.5a**). Protection was dependent on the IN route of delivery because the same dose of HB36.6 delivered intravenously (IV) provided no protection. Lower doses of 3.0, 1.0, and 0.1 mg/kg administered IN two hours prior to lethal challenge with CA09 also resulted in 100% survival with little (0.1mg/kg) or no (1.0 and 3.0 mg/kg) weight loss whereas controls and only those mice that received the lowest dose tested (0.01 mg/kg) exhibited rapid weight loss and succumbed to the infection within 7 days p.i. (**Figure 1.5b**).

To determine if HB36.6 can provide broad protection against genetically distinct strains in vivo, we inoculated mice IN with HB36.6 two hours before challenge with either CA09 or PR8, a highly virulent H1N1 mouse-adapted influenza virus from group 1 that is 82% divergent from CA09. HB36.6 provided the same degree of protection against PR8 as for CA09 with 100% of the mice surviving and no significant weight loss (**Figure 1.5c**). This result is consistent with the in vitro results showing that HB36.6 broadly binds many Group 1 HAs.

We next investigated HB36.6 as a post-exposure therapeutic. We challenged mice with CA09 virus and then either treated with a single IN dose of 3 mg/kg HB36.6 on day 0 (2 hours p.i.), +1, +2 or +3 p.i. or four daily IN doses administered on days +1-4 p.i. HB36.6 reduced weight loss and afforded complete recovery and protection from lethality in 100% of mice when administered daily for 4 days or as a single inoculation administered on day 0 or +1 p.i. (**Figure**

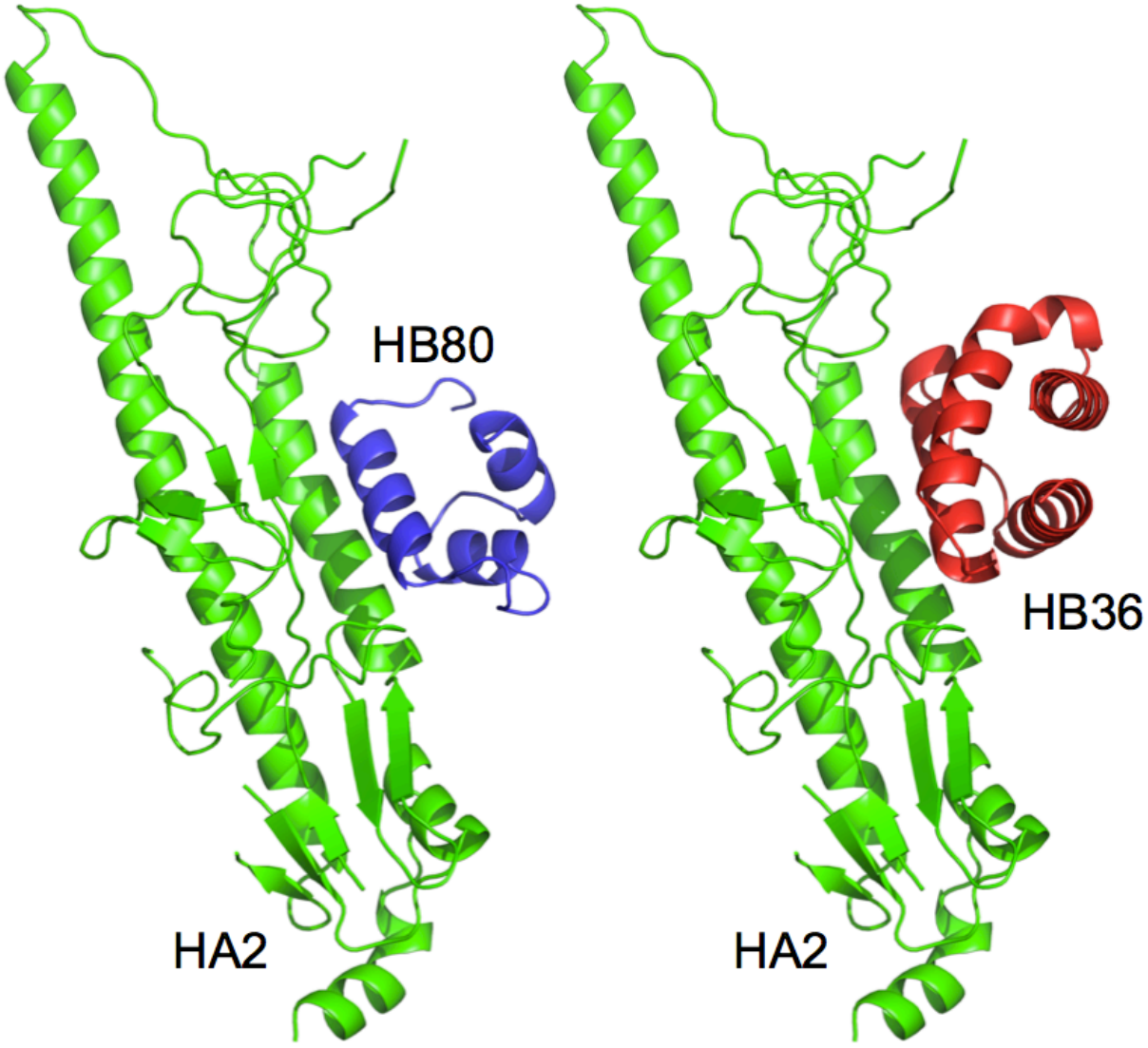
1.5d) There was no significant difference in protection from weight loss between mice receiving daily doses on days +1-4 p.i. versus a single dose at day +1 p.i. suggesting that a single dose within 1 day post-exposure is sufficient to achieve 100% protection. Although mice that received HB36.6 at day +2 or +3 p.i. eventually succumbed to their infection, a single dose at these later time-points also provided a benefit in that the majority of these mice survived up to days 8 or 9 p.i. whereas 100% of the controls succumbed within 4-7 days p.i. (2 d.p.i., p= 0.0006; 3 d.p.i, p=0.0031 compared to control) (**Figure 1.5d**). The protection observed was specific for HB36.6 binding to the HA since daily administration of the scaffold protein (PDB ID 1u84) that HB36.6 is modeled on provided no protection (**Figure 1.5d**). In addition, antiviral cytokines were not detected in lungs collected 24 hours after IN administration of 3.0 mg/kg HB36.6 indicating protection was also not mediated by nonspecific induction of anti-viral cytokines in the lungs (not shown).

Discussion and future work

Our previous studies showed that computationally designed proteins optimized for high affinity binding to a viral protein can neutralize viruses in vitro[14,15]. Here, we report the first proof-of-concept that this model of computational-based drug discovery can lead to a highly effective antiviral that provides robust protection in vivo. We show that HB36.6, computationally-designed to bind the conserved HA stem of influenza A is able to bind and neutralize all group 1 viruses in vitro and a single intranasal dose afforded protection against two widely divergent group 1 influenza strains in vivo indicating that the broad specificity of HB36.6 observed in vitro translated to broad protection in vivo. This broad efficacy would be critical especially during a pandemic when pre-existing immunity is non-existent, and there is not a suitable vaccine. Prophylactic protection when HB36.6 was administered IN up to two days

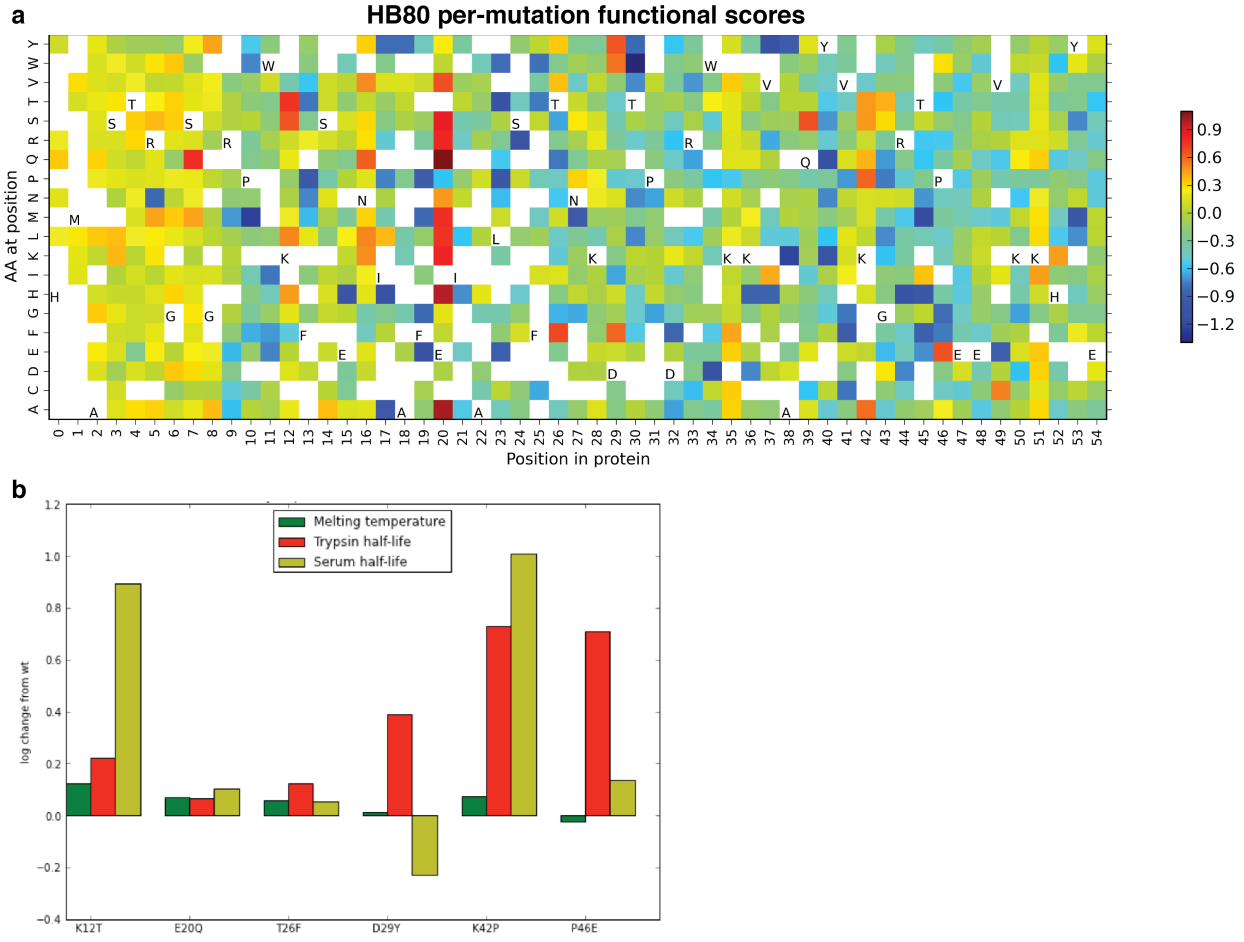
before infection appears to be mediated by limiting or blocking viral replication at the respiratory site of exposure whereas therapeutic protection when HB36.6 was administered IN up 1 day post-infection is likely achieved by curtailing the spread of the virus into the lower respiratory tract and limiting inflammation and disease. While clinical testing in humans is needed to establish whether the duration of protection would be the same, this relatively short duration of prophylactic activity might limit utility in ordinary seasonal flus or in those not at high risk of imminent infection.

Figure 1.1



Two designed protein binders, HB80 and HB36, inactivate influenza hemagglutinin by binding the conserved stem region. Design models of the two protein binders show that they bind to the same area of the stem region of the influenza hemagglutinin HA2 subunit. Binding in this location is thought to prevent the pH-mediated structural change of the protein that enabled endosomal escape, thereby preventing the virus' RNA from escaping degradation in the mature lysosome.

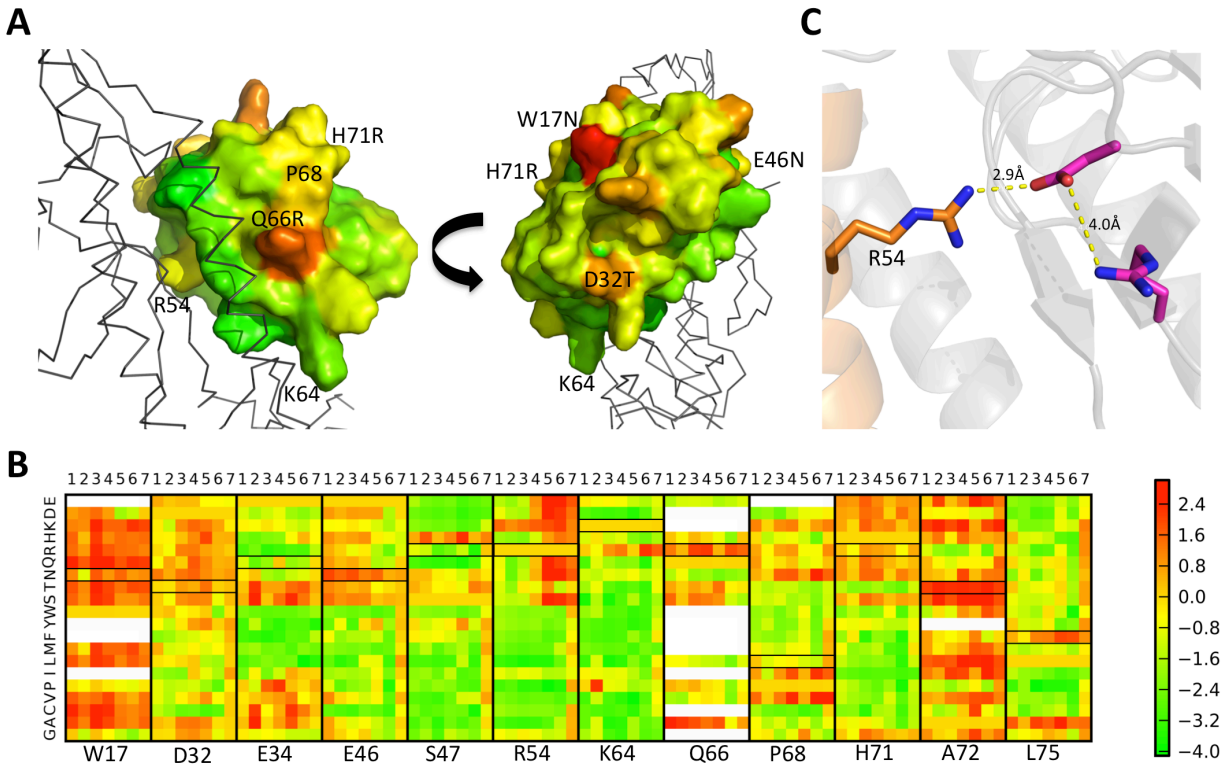
Figure 1.2



Temperature-ramped yeast display sorting reveals stabilizing mutations in HB80. a) Regression of log2 mutation abundances over the five collected temperature points reveals a

mutation landscape where most substitutions are destabilizing, but some of them are putatively strongly stabilizing. Testing seven of these putative stabilizing mutations biochemically for three different measures of protein stability: melting temperature, trypsin resistance, and mouse serum digestion resistance (**b**) shows that all seven are stabilized by at least one measure, and 5/7 are stabilized by all three measures.

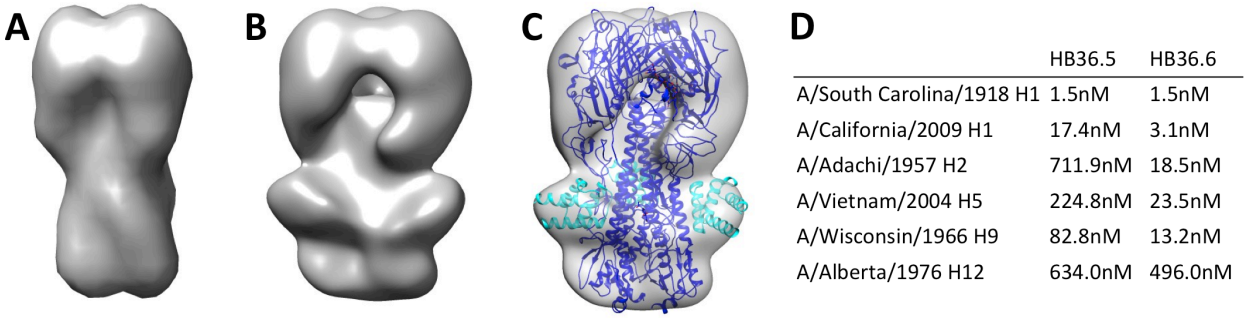
Figure 1.3



HA subtype selections reveal both specificity-enhancing mutations and overall affinity-increasing substitutions. (**A**) HB36.5, in complex with HA, colored by average residue enrichment in FACS sorts against the 7 different HA subtypes. Residues that were mutated to

obtain HB36.6 are labeled in black. (**B**) Enrichment of substitutions at 12 key positions in HB36.5 in selections against each HA strain. Labels at the bottom indicate position in HB36.5; numbers at the top represent the different flu strains (Subtype 1: A/South Carolina/1/1918, 2: A/California/04/2009, 3: A/Vietnam/1203/2004, 4: A/Indonesia/05/2005, 5: A/Adachi/2/1957, 6: A/turkey/Wisconsin/1/1966, 7: A/duck/Alberta/60/1976). At most positions, the enrichment profiles against the different HA strains are similar, but at several they are quite distinct. At Arg 54 for example, the arginine is highly conserved in selections for binding against HAs 1-4, but is outcompeted by smaller charged/polar residues in selections against HAs 5-7 (red region at upper right of R54 panel). White cells indicate insufficient data (<15 sequences in the input library); black boxes indicate the residue identities in HB36.6. (**C**) Origin of HA strain dependence of substitutions at HB36.5 position R54. R54 forms salt bridges with Asp and Arg residues in HAs 1-4. In HAs 5-7, the Asp is mutated to a Glu, disrupting the salt bridge with Arg54 leading to a preference for smaller polar residues.

Figure 1.4

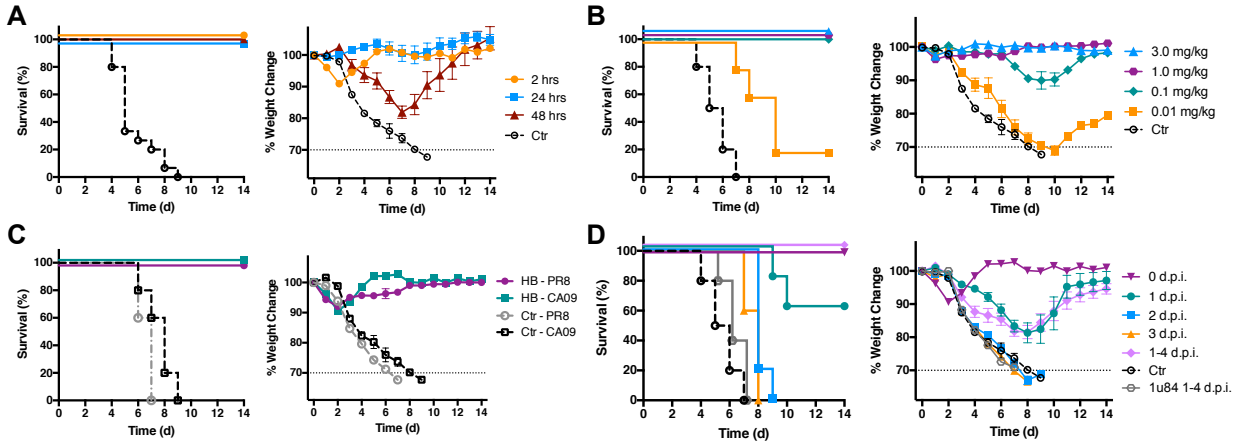


	HB36.5	HB36.6
A/South Carolina/1918 H1	1.5nM	1.5nM
A/California/2009 H1	17.4nM	3.1nM
A/Adachi/1957 H2	711.9nM	18.5nM
A/Vietnam/2004 H5	224.8nM	23.5nM
A/Wisconsin/1966 H9	82.8nM	13.2nM
A/Alberta/1976 H12	634.0nM	496.0nM

Characterization of HB36.6. (A) Structure of HA (PDB 3R2X) with HB36.3 removed and filtered to 20 Å resolution. (B) EM structure of HA PR8 bound to HB36.6. (C) Structure of HA

(blue) in complex with HB36.3 (cyan) (PDB 3R2X) docked into EM reconstruction in (B). HB36.3 fits well into the extra density in the stem region. (D) Equilibrium binding constants determined by biolayer interferometry for HB36.5 and HB36.6 against 6 HAs demonstrate broad improvements against a variety of Group 1 strains.

Figure 1.5



Intranasal delivery of HB36.6 affords prophylactic and therapeutic protection against lethal Influenza virus challenge. (**A**) Survival and weight change in BALB/c mice (N=10 per group) that received 6 mg/kg of HB36.6 administered intranasally (IN) at 2, 24, or 48 hours before challenge with 10 MLD_{50} CA09 virus. (**B**) Survival and weight change in BALB/c mice (N=5 per group) that received 0.01–3 mg/kg IN doses of HB36.6 2 hours before challenge with 10 MLD_{50} of CA09 virus (**C**) Survival and weight change in BALB/c mice (N = 5 per group) that received 3 mg/kg of HB36.6 (HB) IN 2 hours before IN infection with 10 MLD_{50} of H1N1 CA09 virus or 6 MLD_{50} H1N1 A/PR/8/34 (PR8). (**D**) Survival and weight change in BALB/c mice (N = 10 per group) that received 3 mg/kg of HB36.6 IN on day 0 (2 hours post-infection) or +1, +2, or +3 days post-infection (d.p.i.) with 10 MLD_{50} CA09 virus. Shown is mean and SEM.

Section 2: Supplementary information for improving the drug-like properties of designed influenza binders

Materials and methods

Yeast display k_D titrations

Wild type HB80.4 and HB36.5 and the transformed HB36.5 site-saturation mutagenesis (SSM) yeast display libraries were inoculated into 1mL of synthetic dextrose casamino acids (SDCAA) medium supplemented with carbenicillin and chloramphenicol and grown overnight at 30°C, 250rpm. Cells were pelleted by centrifugation, resuspended in 200μL of synthetic galactose casamino acids (SGCAA), 40μL of the resuspended cells were inoculated into 960 additional μL of com- plete SGCAA and induced ~24h at 18°C, 250rpm. Cells were collected by centrifugation, washed with Phosphate Buffered Saline (PBS), 0.1% w/v bovine serum albumin (BSA) (PBSF), and diluted to optical density 600nm (OD600) of 2.0. $1.5x10^5$ cells were mixed with purified biotinylated hemagglutinin (HA) in PBSF individually at a range of concentrations spanning the construct's predicted k_D and incubated at 22°C for 30m. After labeling with HA, the cells were collected by centrifugation, washed once with PBSF, and incubated with 0.6μL of FITC-labeled anti-CMyc antibody and 0.25μL phycoerythrin (PE)-labeled streptavidin on ice for 10m. Cells were collected, washed with PBSF, and resuspended into 200μL of PBSF. Fluorescence of 50,000 cells from each titration point was measured on an Accuri C6 flow cytometer with a 488nm laser for excitation and a 575nm band pass filter for emission. Negative controls for binding were induced cells with no HA labeling. BD Cflow software was used to measure the total PE fluorescence of the displaying cell population and a custom MATLAB non-linear curve fitting script was used to derive equilibrium binding constants for each hemagglutinin subtype.

Site-saturation mutagenesis library construction

HB36.5 in pETCON plasmid was mutagenized individually via Kunkel's method [42] in 86 consecutive codon positions using NNK degenerate primers purchased from Integrated DNA Technologies (Coralville, IA). Primers were designed using Firnberg's method. The theoretical library size of all 86 reactions was 1720 amino-acid sequences. Kunkel's reactions were purified with QiaQuick columns (Qiagen, Hilden, Germany) and pooled in groups of 12 (codon positions 1–12, 13–24, 25–36, 37–48, 49–60, 61–72, 73–86). 1μL of each plasmid pool was trans- formed by electroporation into XL10 Gold electrocompetent cells (Stratagene, La Jolla, CA) with a minimum efficiency of 3x105 Colony Forming Units (CFUs) per pool (>100-fold cover- age).

Liquid cultures were grown overnight in terrific broth (TB) and harvested using a Qia-Prep miniprep kit (Qiagen, Hilden, Germany). Mutated genes were amplified from each plasmid pool by adding 1μL of plasmid to 10μL of 5x Phusion Buffer, 1μL of 10mM dNTPs, 2.5μL of 10μM upGS primer, 2.5μL of 10μM downCMyc primer, and 0.5μL of Phusion polymerase in 50μL total volume. The reaction used 30 cycles of PCR (98°C 10s, 65°C 15s, 72°C 15s). PCR product was purified with a QiaQuick kit and transformed into EBY100 S. cerevisiae using Chao's method along with gel-purified pETCON vector digested with NdeI/XhoI (NEB, Waltham, MA).

Yeast display HA selections

A transformed HB36.5 SSM yeast display library was sorted in two rounds. For each round, cells were grown in 10mL of SDCAA overnight at 30°C, collected by centrifugation, and induced in SGCAA at 18°C for ~24 hours. Cells were collected by centrifugation, washed with PBSF, and ~$4x10^6$ cells labeled with purified biotinylated HA at a concentration half of the kD determined by yeast display titrations or, if no kD could be determined, at 500nM. In the first

round of sorting, primary labeling proceeded for 30m at 22°C as described for titrations. In the second round, to select for mutations that improved binding but did not destabilize the protein, primary labeling was performed for 30m at 37°C. Secondary labeling was done with 1.2μL anti-c-Myc tag antibody conjugated to fluorescein isothiocyanate (FITC) and 0.5μL streptavidin-phycoerythrin (SAPE) in a total volume of 100μL PBSF on ice for 10m. In each sorted population, the top ~5% of FITC-displaying cells were collected.

Illumina sequencing

Plasmid DNA was prepared as previously described[14]. Genes were amplified from the plasmid by adding 36.5μL of purified plasmid to 10μL of 5x Phusion master mix, 1μL each of pETCON_inner_fwd and pETCON_inner_rev, 1μL of 10mM dNTPs, and 0.5μL of Phusion polymerase (Thermo). The reaction used 30 cycles of PCR (98°C 10s, 58°C 15s, 72°C 15s). Correctly sized products were gel extracted using a Qiaquick gel extraction kit (Qiagen). 10μL of gel extracted reaction product were added to 10μL of 5x Phusion master mix, 1μL each of miseq_outer_fwd and miseq_outer_rev with the correct library barcode, 1μL of 10mM dNTPs, and 0.5μL of Phusion polymerase in 50μL, and amplified again with 30 cycles of PCR (98°C 10s, 58°C 15s, 72°C 15s). The two primer sets have overhangs that add Illumina sequencing primer binding sites, barcode sequences, and flow cell adaptors to the gene to be sequenced.

They additionally add 12 entirely degenerate bases at the beginning of the forward and reverse read, ensuring adequate diversity for the Illumina basecalling algorithms. This enabled the DNA pools to be prepared and sequenced in two runs of paired-end 251bp mode on an Illumina MiSeq (Illumina, San Diego, CA) using a standard MiSeq kit and protocols. Pools were mixed to have 6.5% of the total loaded DNA from the unselected pool, 6.5% of the total loaded

DNA from each first-round selected pool, and 3.25% from each second-round selected pool, with 35% Illumina PhiX control DNA to increase diversity and data quality.

Deep mutational scanning data processing

Raw sequence files were processed into fastq format, split by barcode, allowing up to 1 mismatch, and adapter sequence was removed using Illumina OLB 1.9.4. Split library sequences were processed using scripts from Enrich 0.2 to yield mutation counts in each library. Counts for each sorted library were converted to log2 enrichment relative to the unselected library using custom scripts. An enrichment value was calculated by linear regression of enrichment for each individual substitution at each round. The slope of the regressed line is the enrichment value[24].

Combinatorial library construction and selection

Twelve positions in HB36.5 that contained substitutions highly enriched against many or all tested subtypes were mutated in a combinatorial library with a total sequence diversity of 108. This library was constructed using recursive PCR assembly, as described below, with the only difference being that the assembly oligos contained degenerate codons designed using GLUE [27] to maximize enriched amino acid codon representation. This library was transformed into yeast using Chao's method[28] with a transformation efficiency around 10^7, and sorted by three rounds of yeast display for binding to A/South Carolina/1/1918 HA until sequence convergence was achieved. The final converged sequence, HB36.6, had 9 total mutations relative to HB36.5.

Recursive PCR assembly

The gene for HB36.6 with 40bp of additional pETFLAG overlap sequence, to allow homologous recombination, was assembled via recursive PCR. Sequences were designed using DNA- Works[29] and purchased from Integrated DNA Technologies, Inc. (Coralville, IA). The outermost two primers were diluted to 5μM and mixed, while the inner primers were diluted to

0.5μM and mixed. A 10μL volume of outer primer mix was added to 12.7μL of inner primer mix, along with 1μL of 10mM DNTPs, 6μL of 5x Phusion buffer, and 0.3μL of Phusion polymerase (NEB, Waltham, MA) for a final volume of 30μL. Product was assembled with 30 rounds of PCR (98°C 30s, 58°C 30s, 72°C 30s). A second round of PCR was used to further amplify correctly assembled product. A 1.25μL aliquot of the first PCR reaction product was added to 5μL of 5x Phusion Buffer, 0.75μL 10mM DNTPs, 2μL of outer primer mix, and 0.25μL of Phusion polymerase in 25μL. The same PCR conditions were used for a further 30 rounds and the product was purified using a QiaQuick PCR cleanup kit (Qiagen, Hilden, Ger- many) and eluted in EB. Gibson assembly[30] was used to insert the assembled gene into gel- purified pETFLAG vector digested with NdeI/XhoI (NEB, Waltham, MA). A 1.5μL aliquot of cut vector at 20ng/μL was added to 1μL of assembled gene and 7.5μL of Gibson enzyme mix (all enzymes from NEB, Waltham, MA). The reaction was incubated at 50°C for 1h and 2μL was transformed into 20μL of XL10 Gold chemically competent E. coli and plated onto a kanamycin agar plate. Plasmid sequences were confirmed by colony PCR and Sanger sequencing, and colonies with correctly assembled plasmid were grown in TB and harvested using a QiaPrep miniprep kit (Qiagen, Hilden, Germany).

Protein expression and purification

HB36.6 in a pETFLAG vector was expressed in Rosetta2 (DE3) E. coli cells. HB36.6 used throughout these studies contained the FLAG tag (DYKDDDDK). Cells were grown in 500μL aliquots of medium salt aspartate-glucose (MDG) non-inducing media supplemented with kanamycin at 37°C, 250rpm overnight. Each vial was used to inoculate 500mL of ZYM-5052 auto-induction media[31], which was grown for ~48 hours at 22°C, 250rpm. Cells were harvested by centrifugation and resuspended in 25mL of lysis buffer (50mM Tris, 300mM NaCl, 30mM

imidazole, pH 8.2) with half of a dissolved complete, ethylenediaminetetracetic acid (EDTA)-free protease inhibitor tablet (Roche, Basel, Switzerland) and supplemented with DNAse and lysozyme at ~1mg/mL. Resuspended cells were lysed via sonication with a Qsonica Q500 (Fisher Scientific, Hampton, New Hampshire) at 70% power for 10 minutes (20s on/20s off) on ice. Insoluble cell debris was removed by centrifugation for 30m at 40,000g. Supernatant was applied to gravity-flow columns containing 2.5mL of Ni-NTA resin (Qiagen, Hilden, Ger- many) pre-equilibrated with lysis buffer. Protein was washed with 25mL wash buffer (50mM Tris, 300mM NaCl, 75mM imidazole, pH 8.2) and eluted with 10mL elution buffer (50mM Tris, 300mM NaCl, 300mM imidazole, pH 8.2). Protein was concentrated to ~20mg/mL using a Vivaspin 10kD MWCO centrifugal concentrator (Sartorius Stedim, Goettingen, Germany) at 4000g at 4°C. Imidazole was removed by dialysis (2x 4L buffer) into 50mM Tris, 300mM NaCl, pH 8.2 at 4°C. Concentration was determined by absorbance at 280nm on a NanoDrop spec- trophotometer (Thermo Scientific, Waltham, Massachusetts) using extinction coefficients cal- culated from amino acid sequences. For in vivo experiments, proteins were further processed by incubating with magnetic bacterial endotoxin removal beads. Beads were removed with a magnet and then samples were centrifuged to ensure complete removal. Endotoxin lipopolysaccharide (LPS) levels were reduced to <110 EU/mg (Miltenyi Biotec, San Diego, California).

A larger scale purification procedure for HB36.6 was used to produce the protein for the mouse study comparing HB36.6 to oseltamivir. Rosetta2 (DE3) E. coli cells carrying the HB36.6 pETFLAG vector were cultured in 50 ml of ZY Broth (10 g/l tryptone, 5 g/l yeast extract) with 15 ug/ml kanamycin at 37°C, at 250rpm overnight. The overnight culture was used to seed (3 ml starter culture per bottle) multiple airlift fermentation bottles, each containing 2 liters of autoinduction media ZYP-5052[31] supplemented with 15 ug/ml kanamycin, and the cultures were

sparged with air for 72 hours at 25°C using an airlift LEX Bioreactor Sys- tem (Harbinger Biotech, Ontario Canada)[32]. Cells were harvested at 4°C from culture media in 2 liter centrifuge buckets, pelleted at 4000g using a Sorvall RC 12 BP centrifuge fitted with an H-12000 swinging-bucket rotor. The masses of the cell pastes were measured to verify proper growth (usually in the range of 20–30 g per 2 liter culture). The pelleted cells were flash frozen in liquid nitrogen and stored at -80°C. Frozen cell pellets from the equivalent of 2 liters of auto- induction culture were thawed by addition of 20 ml of Lysis Buffer (25 mM HEPES pH 7.0, 500 mM NaCl, 5% Glycerol, 0.5% CHAPS, 30 mM Imidazole, 10 mM MgCl2, 1 mM TCEP, 250 ug/ml AEBSF, and 0.025% Azide) and 15 minute incubation at 37°C in a water bath, followed by addition of 0.01g of lysozyme and incubation for an additional 15 minutes in the 37°C water bath. The thawed cell pellet is gently resuspended with a spatula and transferred into a 500 mL beaker inside an ice bath filled with 180 ml of Lysis Buffer. The resuspended pellet is lysed in the ice bath by sonication for 30 minutes at 100W with "10 s ON and 20 s OFF" cycle. After sonication, the crude lysate is clarified with 20ul (25 units/ul) of Benzonase and incubated at room temperature for 40 minutes in 250 ml centrifuge bottles using a Stuart SRT1 rotating mixer. The clarified crude extract is then centrifuged at 14,000 rpm for 1 hour at 4°C using a Sorvall SLA-1500 Rotor and the supernatant is transferred into a clean reservoir. Using an ÄKTAexplorer 100 (GE Healthcare), the supernatant with soluble protein is pumped at 5 ml/ min aver a 5 ml Ni-NTA His-Trap FF column (GE Healthcare) pre-equilibrated with Wash Buffer (25 mM HEPES pH 7.0, 500 mM NaCl, 5% Glycerol, 30 mM Imidazole, 1 mM TCEP, and 0.025% Azide). The column is washed with 20 column volumes (CVs) of Wash Buffer and the bound protein is then eluted with 20 ml of Elution Buffer 1 (Wash Buffer + 250 mM imidazole), followed by 20 ml of Elution Buffer 2 (Wash Buffer + 320 mM imidazole). Each imidazole elution pool was then

further purified by preparative size exclusion chromatography using an ÄKTAexplorer 100, and a HiLoad 26/60 Superdex 75 preparative-grade column (GE Healthcare) equilibrated with SEC Buffer (25 mM HEPES, 0.5 M NaCl, 5% Glycerol, 1 mM TCEP, pH 7)[33]. Following SDS-PAGE analysis, the peak SEC fractions of HB36.6 were pooled together with additional SEC Buffer such that a HB36.6 was at final concentration of 1–2 mg/ml. Using a Millipore Tangential Flow Filtration System (Millipore Labscale system), the pooled HB36.6 was concentrated to approximately 5 mg/mL using a 5kDa MWCO PES 10cm2 membrane, then diafiltered at a constant volume against Storage Buffer (20mM Tris-HCl, 300mM NaCl pH 8) for five diavolumes at 2-8C. Transmembrane pressure was held to 20psi. Concentrated HB36.6 samples were analyzed by SDS-PAGE to verify purity of the samples at >98% and mass spectrometry to verify identity. Endotoxin levels were also measured using the Endosafe PTS reader (Charles River Laboratories), with a result of 17.2 EU per mg of HB36.6. Finally the purified HB36.6 was dispensed into 2 ml vials and flash frozen in liquid nitrogen followed by storage at -80°C.

Circular dichroism spectroscopy

Purified proteins were tested for folding and denaturation temperature using an Aviv 420 circular dichroism (CD) spectrometer (Aviv Biomedical, Lakewood, NJ). Protein was diluted to 0.6mg/mL in PBS and CD absorbance was measured at 205nm at 25°C. Absorbance was characteristic of a structured α-helical protein. To test the thermal denaturation temperature of HB36.6, absorbance at the 222nm α-helix peak was measured in 2° increments between 15° and 95°C.

Protease resistance assays

Proteins to be tested were diluted to 1mg/mL in PBS and Gibco 0.25% trypsin/phenol red (Life Technologies, Carlsbad, CA) was diluted to 0.005% in PBS. Equal parts protein solution

and trypsin were mixed and incubated at 37°C. Time points were taken by removing 6μL aliquots of solution, mixing with 6μL of 2x SDS loading buffer (100mM Tris, 4% SDS, 0.2% bromphenol blue, 20% glycerol, pH 6.8) and incubating for 2m at 85°C. Denatured aliquots were stored at -20°C until being loaded on a 12% NuPage Bis-Tris gel (Life Technologies, Carlsbad, CA) and run at 150V in 1x MES buffer (50 mM MES, 50 mM Tris Base, 0.1% SDS, 1 mM EDTA, pH 7.3). Negative controls were one lane with undigested protein and one lane with trypsin but no test protein. Gels were scanned and bands were quantified using ImageJ. Band size as a percentage of the undigested negative control was fit by non-linear regression using a custom MATLAB script to derive protein digestion half-lives.

BioLayer Interferometry (BLI)

Titrations were performed on an OctetRed96 BLI system (ForteBio, Menlo Park, CA) using streptavidin-coated probe tips. Tips were equilibrated for 10m in 50x kinetics buffer, designed to reduce nonspecific binding (PBS, pH 7.4, 0.5% w/v BSA, 0.05% v/v Tween 20) and loaded with 25-40nM biotinylated HA of one of six subtypes in 50x kinetics buffer for 15m. Following a 10m wash and 10m baseline reading in the 50x kinetics buffer, association rates were measured by incubating each tip for 30m in different concentrations of purified HB36.5 or HB36.6 protein spanning the predicted k_D for the given HA subtype, diluted in 50x kinetics buffer. Dissociation was measured by then incubating the tips in 50x kinetics buffer for a further 30m. HB36.6 did not show well-defined off-rates, so equilibrium binding constants were computed from the maximum steady-state response reached during the association phase. The limit of detection for the Octet instrument used in these experiments is around 1nM and the equilibrium binding experiments on the Octet yielded values of HB36.5 and HB36.6 against SC1918 at that threshold and within a standard error of each other. However, HB36.6 is a

stronger binder against this strain with lower k_{off}, though not a different k_{on}. Unfortunately, due to equipment limitations, a specific measurement of how much stronger was not possible. The k_D values for HB36.6-SC1918 were as low as 1pM but, since the dissociation curves are shallow, any minor deviation would cause enormous changes in the calculated k_D. Thus equilibrium values are reported. Nonlinear regression curve fitting was done with a custom MATLAB script.

In vitro antiviral neutralization

MDCK (Madin Darby canine kidney) from American Type Culture Collection (ATCC, Manassas, VA) were grown in Growth medium comprising minimum essential medium (MEM) with non-essential amino acids, 5% FBS and 0.22% NaHCO3. Influenza A/California/ 07/2009 (H1N1), A/Puerto Rico/08/1934 (H1N1), A/New Caledonia/20/1999 (H1N1), A/ Hong Kong/213/2003 (H5N1), A/Nanchang/933/1995 (H3N2), A/Brisbane/10/2007 (H3N2), and A/Hong Kong/33982/2009 (H9N2), were obtained from the Center for Disease Control (Atlanta, GA). Influenza A/Duck/MN/1525/81 was kindly provided by Robert Webster

(St. Jude Children's Research Hospital, Memphis. TN). The viruses were prepared in Madin Darby canine kidney (MDCK) cells, placed in ampules and frozen at -80°C. Cells are seeded to 96-well flat-bottomed tissue culture plates at the proper cell concentration to establish confluent cell monolayers and incubated overnight at 37°C. Various dilutions of test compound were added to each well. Ribavirin (1-β−D-ribofuranosyl-1,2,4-triazole-3-carboxamide), FI6v3, and HB36.6 were tested in half-log increments from 320 μg/ml and below. Virus was added to test compound wells and to virus control wells at about 50–100 cell culture infectious dose per ml. The virus titer was determined by a prior titration, where the most diluted virus stock causes 100% cytopathic effect (CPE) in all wells at the particular virus dilution. Test medium without virus was added to all toxicity control wells and to cell control wells. The plates were incubated

at 37°C for 72 hours. Sterile neutral red (0.034% in saline solution) was then added to each well. After two hours at 37°C, all medium was removed and the cells washed with PBS and inverted to drain. Neutral red was extracted from the cells by adding an equal volume mixture of absolute ethanol and Sörensen's citrate buffer, pH 4.2. The contents of each well are mixed gently and the optical density (O.D.) values of each well are obtained by reading the plates at 540 nm with a microplate reader.

Negative-stain sample preparation and imaging

Complexes of HA and HB36.6 were prepared for electron microscopy studies by diluting to 2.1 μg/ml in Tris-buffered saline and applied to freshly glow discharged carbon coated 400 mesh copper grids for 20 seconds. Two rounds of a 3μl droplet of 2% uranyl formate were applied and immediately blotted followed by a third 3μl droplet blotted after 1 minute. Grids were viewed using the FEI Tecnai T12 electron microscope operating at 120 kV accelerating voltage at 52,000 x magnification resulting in a pixel size 2.05 Å at the specimen level. Images were acquired on a Tietz 4k x 4k complementary metal-oxide-semiconductor (CMOS) camera using Leginon[34,35] MSI-raster 3.0 software package at a defocus of ~1.0 μm. Microscope magnifications were calibrated using a catalase crystal prior to data collection.

EM data processing and 3D volume reconstruction

Particles were picked automatically using DoG Picker[36] and boxed into 96x96 pixel boxes and aligned using Xmipp CL2D clustering alignment. Ten *ab initio* models of each complex were created using EMAN2CL with C3 symmetry and based on 17 2D class averages of PR8 in complex with HB36.6. Initial models of complexes were then refined against 10,005 raw particles using EMAN. The resolution of the final model was determined to be ~22 Å using

an FSC cut-off of 0.5. The UCSF Chimera "Fit in Map" function was used to dock structural models into the EM maps.

HB36.6 and oseltamivir administration and influenza challenge

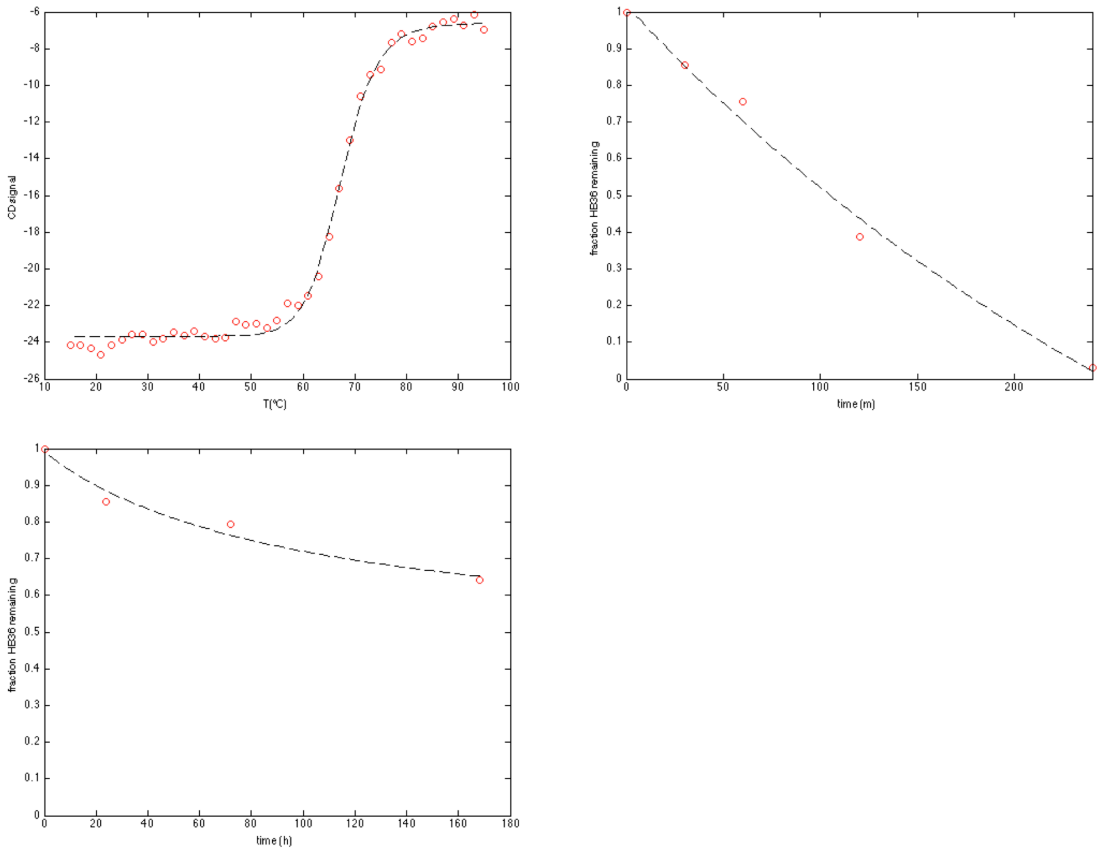
Animal studies approved by the University of Washington and Utah State University Institu- tional Animal Care and Use Committee. Female, 6–8 week-old BALB/c mice were randomly separated in to groups, anesthetized and intranasally administered protein binder (HB36.6) at concentrations varying from 0.01 to 6.0 mg/kg. Two to forty-eight hours later, the mice were anesthetized with 2.5% isoflurane and challenged IN with 3–10 MLD50 (fifty percent mouse lethal dose) of A/California/04/09 (H1N1) (CA09), A/PR/8/34 (H1N1) (PR8) or A/Duck/MN/ 1525/81 (H5N1) (MN81). In a therapeutic setting, mice received the protein binder 0 (2 hours post-infection), +1, +2, +3, or +4 days post infection. The mice were monitored daily for weight loss and survival until 14 days post-infection. Animals that lost more than 30% of their initial body weight were euthanized by carbon dioxide in accordance with our animal protocols. Oseltamivir-treated mice received 2.5 mg/kg of oseltamivir (Roche, Palo Alto, CA) twice daily for

5 days (total of 10 doses) by oral gavage. Oseltamivir was dissolved in water prior to adminis- tration. The SCID (Non-Obese Diabetic (NOD), Severe Combined Immunodeficiency (SCID) gamma, strain NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ) mice and the MyD88-/- (strain B6.129P2 (SJL)-Myd88tm1.1Defr/J) mice were purchased from Jackson Laboratory. At least five mice per group were used for each experiment. All mice used for the experiments are included for analyses. For mouse experiments, researchers were not blinded to animal identity.

Statistical and power analyses

All of the analyses were performed using Graphpad Prism version 5.01. A Student's t test (to compare two samples) and analysis of variance (ANOVA) (to compare multiple samples) were used for statistical analysis. Survival analyses were performed by using the Kaplan-Meier log-rank test. A P value of <0.05 was considered to be significant. For mice, the minimum group size was determined using weight loss data with 100% of control mice becoming infected with CA09. Based on a standard deviation of 2% in weight loss, a group size of n = 5 yields >80% power to detect a minimum of a 10% difference between groups in weight loss using a two-sized t-test with an alpha value of 0.05.

Figure 2.1



Stability of HB36.6. A) Melting curves obtained by circular dichroism at 222nm for HB36.6. Nonlinear regression indicate a T_m of 67.3°C. **B**) Trypsin digestion curve for HB36.6. The curve indicates a digestion half-life for HB36.6 around 100m. **C**) Serum digestion curve for HB36.6. The digestion was halted before a half-life for HB36.6 could be determined, but established that it is greater than 1 week (168 hrs).

Table 2.1 HB36.5 yeast display kDs

Subtype	kD
A/South Carolina/1/1918	7.64 nM
A/California/04/2009	10.95 nM
A/Adachi/2/1957	>128 nM
A/Indonesia/05/2005	16.8 nM
A/Vietnam/1203/2004	92.45 nM
A/turkey/Wisconsin/1/1966	9.63 nM
A/duck/Alberta/60/1976	>128 nM

Table 2.2 Flow cytometry statistics

Target	sort #	label	labeling temperature	# of cells analyzed	# of cells collected	% of cells collected
A/South Carolina/1/1918	1	4 nM	22°C	4,307,076	250,000	5.80%
	2	4 nM	37°C	1,750,000	100,000	5.71%
A/California/04/2009	1	5 nM	22°C	4,463,611	250,000	5.60%
	2	5 nM	37°C	1,740,493	100,000	5.74%
A/Adachi/2/1957	1	500 nM	22°C	3,988,455	250,000	6.28%
	2	500 nM	37°C	2,000,000	100,000	5.00%
A/Indonesia/05/2005	1	8 nM	22°C	4,403,380	250,000	5.68%
	2	8 nM	37°C	1,590,000	100,000	6.29%
A/Vietnam/1203/2004	1	45 nM	22°C	4,235,994	250,000	5.90%
	2	45 nM	37°C	1,650,000	100,000	6.06%
A/turkey/Wisconsin/1/1966	1	5 nM	22°C	4,054,467	250,000	6.17%
	2	5 nM	37°C	1,740,000	100,000	5.75%
A/duck/Alberta/60/1976	1	500 nM	22°C	5,066,844	250,000	4.93%
	2	500 nM	37°C	2,060,000	100,000	4.85%

Table 2.3 List of oligos used for construction of HB36.5 site-saturation mutagenesis library.

Name	Sequence
HB36.5_KSSM_0	GTCCATTGCGTTMNNCATATGGCTAGCC
HB36.5_KSSM_1	GACCGTCCATTGCMNNAGACATATGGCT
HB36.5_KSSM_2	ATTGTTGACCGTCCATMNNGTTAGACATATGGCT
HB36.5_KSSM_3	TTGTTGACCGTCMNNTGCGTTAGACATATGGCT
HB36.5_KSSM_4	CTGTTTAATTGTTGACCMNNCATTGCGTTAGAC
HB36.5_KSSM_5	GTTTAATTGTTGMNNGTCCATTGCGTTAGACATATGGC
HB36.5_KSSM_6	ATCTGTTTAATTGMNNACCGTCCATTGCGTTAGACA
HB36.5_KSSM_7	TAATCTGTTTAAMNNTTGACCGTCCATTGCG
HB36.5_KSSM_8	TAATAATCTGTTMNNTTGTTGACCGTCCATTGCG
HB36.5_KSSM_9	CAATAATAATCTMNNTAATTGTTGACCGTCCATTGCG
HB36.5_KSSM_10	CCATTCCAATAATAAMNNGTTTAATTGTTGACCGTCC
HB36.5_KSSM_11	CCATTCCAATAAMNNTCTGTTTAATTGTTGACCGTCC
HB36.5_KSSM_12	TCCCAGGCACCTATCCATTCCAAMNNTAATCTGTTTAA
HB36.5_KSSM_13	TCCCAGGCACCTATCCATTCMNNTAATAATCTGTT
HB36.5_KSSM_14	TCCCAGGCACCTATCCAMNNCAATAATAATCTG
HB36.5_KSSM_15	GGGTCCCAGGCACCTATMNNTTCCAATAATAA
HB36.5_KSSM_16	GTCCCAGGCACCMNNCCATTCCAATAATAATCT
HB36.5_KSSM_17	AGGGTCCCAGGCMNNTATCCATTCCAA
HB36.5_KSSM_18	CAAACCAAAAGGGTCCCAMNNACCTATCCATTC
HB36.5_KSSM_19	AAACCAAAAGGGTCMNNGGCACCTATCCA
HB36.5_KSSM_20	CCAAACCAAAAGGMNNCCAGGCACCTAT
HB36.5_KSSM_21	TTTACCCAAACCAAAMNNGTCCCAGGCACC
HB36.5_KSSM_22	CTTTACCCAAACCMNNAGGGTCCCAGGCAC
HB36.5_KSSM_23	GTCTTTACCCAAMNNAAAAGGGTCCCAGGCAC
HB36.5_KSSM_24	AAGCGTCTTTACCMNNACCAAAAGGGTCC
HB36.5_KSSM_25	ATAAGCGTCTTTMNNCAAACCAAAAGGGTCCC
HB36.5_KSSM_26	CGTAATCATAAGCGTCMNNACCCAAACCAAA
HB36.5_KSSM_27	GCTTCGTAATCATAAGCMNNTTTACCCAAACCA
HB36.5_KSSM_28	GCTTCGTAATCATAMNNGTCTTTACCCAAACC
HB36.5_KSSM_29	CAGCTTCGTAATCMNNAGCGTCTTTACCCA
HB36.5_KSSM_30	TGCAGCTTCGTAMNNATAAGCGTCTTTACCCA
HB36.5_KSSM_31	AAACTTCTGCAGCTTCMNNATCATAAGCGTC
HB36.5_KSSM_32	AAACTTCTGCAGCMNNGTAATCATAAGCGTC
HB36.5_KSSM_33	CTTCAAAACTTCTGCMNNTTCGTAATCATAAGCGTC
HB36.5_KSSM_34	TACAGCCTTCAAAACTTCMNNAGCTTCGTAATCA
HB36.5_KSSM_35	GTATACAGCCTTCAAAACMNNTGCAGCTTCGTA
HB36.5_KSSM_36	GTATACAGCCTTCAAMNNTTCTGCAGCTTCG
HB36.5_KSSM_37	GTATACAGCCTTMNNAACTTCTGCAGCTTC
HB36.5_KSSM_38	CAGTTTCGTATACAGCMNNCAAAACTTCTGC

HB36.5_KSSM_39	CAGATTCAGTTTCGTATACMNNCTTCAAAACTTCTGC
HB36.5_KSSM_40	GCAGATTCAGTTTCGTAMNNAGCCTTCAAAAC
HB36.5_KSSM_41	GTCAAATGCAGATTCAGTTTCMNNTACAGCCTTCAA
HB36.5_KSSM_42	AAATGCAGATTCAGTMNNGTATACAGCCTTC
HB36.5_KSSM_43	GTCAAATGCAGATTCMNNTTCGTATACAGCCTTC
HB36.5_KSSM_44	CTAAGTCAAATGCAGAMNNAGTTTCGTATACAGCC
HB36.5_KSSM_45	GCTAAGTCAAATGCMNNTTCAGTTTCGTATACAGC
HB36.5_KSSM_46	CTCATGGCTAAGTCAAAMNNAGATTCAGTTTCG
HB36.5_KSSM_47	CTCATGGCTAAGTCMNNTGCAGATTCAGTTTC
HB36.5_KSSM_48	TCTCATGGCTAAMNNAAATGCAGATTCAG
HB36.5_KSSM_49	GTATTCTCATGGCMNNGTCAAATGCAGATTCAG
HB36.5_KSSM_50	ATGTATTCTCATMNNTAAGTCAAATGCAGATTCAGTTTCG
HB36.5_KSSM_51	CCAATGTATTCTMNNGGCTAAGTCAAATGCAG
HB36.5_KSSM_52	TATCCAATGTATMNNCATGGCTAAGTCAAATGCAG
HB36.5_KSSM_53	GCGAAGTTGTATATCCAATGMNNTCTCATGGCTAA
HB36.5_KSSM_54	GCGAAGTTGTATATCCAMNNTATTCTCATGGC
HB36.5_KSSM_55	GAATGCGAAGTTGTATATMNNATGTATTCTCATGGC
HB36.5_KSSM_56	AATGCGAAGTTGTAMNNCCAATGTATTCTCATGGC
HB36.5_KSSM_57	GAATGCGAAGTTMNNTATCCAATGTATTCTCATGGC
HB36.5_KSSM_58	TTTGAATGCGAAMNNGTATATCCAATGTATTCTCATGGC
HB36.5_KSSM_59	GTCTTTTGAATGCMNNGTTGTATATCCAATGTATTCTC
HB36.5_KSSM_60	GCGAATGGTATTTGTCTTTTGAAMNNGAAGTTGTATATCC
HB36.5_KSSM_61	GTGGGCGAATGGTATTTGTCTTTTMNNTGCGAAGTTGTA
HB36.5_KSSM_62	GAATGGTATTTGTCTMNNGAATGCGAAGTTG
HB36.5_KSSM_63	GTGGGCGAATGGTATTTGMNNTTTGAATGCGAA
HB36.5_KSSM_64	GTGGGCGAATGGTATMNNTCTTTTGAATGCG
HB36.5_KSSM_65	GCGTGGGCGAATGGMNNTTGTCTTTTGAA
HB36.5_KSSM_66	TTTGAGCGTGGGCGAAMNNTATTTGTCTTTTGA
HB36.5_KSSM_67	CTTTTGAGCGTGGGCMNNTGGTATTTGTCT
HB36.5_KSSM_68	CTTTTGAGCGTGMNNGAATGGTATTTGTCTTTTGAATGCG
HB36.5_KSSM_69	CAACTTTTGAGCMNNGGCGAATGGTATTTGTCT
HB36.5_KSSM_70	TTCTGGCCAACTTTTGMNNGTGGGCGAATGGT
HB36.5_KSSM_71	TCTGGCCAACTTMNNAGCGTGGGCGAA
HB36.5_KSSM_72	ATCTTCTGGCCAAMNNTTGAGCGTGGGCG
HB36.5_KSSM_73	AATTCCAATAATCTTCTGGCMNNCTTTTGAGCGTG
HB36.5_KSSM_74	CAATAATCTTCTMNNCAACTTTTGAGCGTGGGC
HB36.5_KSSM_75	AATTCCAATAATCTMNNGGCCAACTTTTGAGCGTGGGCG
HB36.5_KSSM_76	CAATTCCAATAAMNNTCTGGCCAACTTTTGAGCGTGGGCG
HB36.5_KSSM_77	CTTGTTTCAATTCCAAMNNTCTTCTGGCCAAC
HB36.5_KSSM_78	CGGCTTGTTTCAATTCMNNTAATCTTCTGGCC
HB36.5_KSSM_79	CGGCTTGTTTCAAMNNCAATAATCTTCTGGCC
HB36.5_KSSM_80	GGCGGCTTGTTTMNNTTCCAATAATCTTCTGGC
HB36.5_KSSM_81	AGGAGGCGGCTTGMNNCAATTCCAATAATCT

HB36.5_KSSM_82	GTGAGGAGGCGGCMNNTTTCAATTCCAA
HB36.5_KSSM_83	TGGTGAGGAGGCMNNTTGTTTCAATTCCAA
HB36.5_KSSM_84	AATGGCAATGGTGAGGAMNNGGCTTGTTTCAA
HB36.5_KSSM_85	AATGGCAATGGTGAMNNGGCGGCTTGTTTC

Table 2.4 List of oligos for construction of HB36.5 combinatorial library

Name	Sequence
hb36_stbl1	CTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTA
	AATCTGTTTAATTGTTGACCGTCCATTGCGTTAGACATATGGCTAG
hb36_stbl2	CCGACCCTCCGCCT
	TGGACGGTCAACAATTAAACAGATTATTATTGGAANNKATAGGTG
hb36_stbl3	CCTGGGACCCTTTTG
	AACTTCTGCAGCKNNGTAWRNATAAGCGTCTTTACCCAAACCAAA
hb36_stbl4	AGGGTCCCAGGCACC
	NYWTACNNMGCTGCAGAAGTTTTGAAGGCTGTATACGAAACTVN
hb36_stbl5	MVRWGCATTTGACTTA
	YYTGAATGCGAAGTTGTATATCCAATGTATWYBCATGGCTAAGTC
hb36_stbl6	AAATGCWYBKNBAGT
	TTGGATATACAACTTCGCATTCARRAGANNKATAVHHTTCGCCVDS
hb36_stbl7	NNKCAAAAGKYNGC
	TGGTGAGGAGGCGGCTTGTTTCAATTCCAATAATCTTCTGGCNRMC
hb36_stbl8	TTTTGMNNSHBGGC
	AAGCCGCCTCCTCACCATTGCCATTAGAAGGTAGTCTCGAGCTCGA
hb36_stbl9	GGGGGGCGGATCCG
hb36_stbl10	CTTCAGAAATAAGCTTTTGTTCGGATCCGCCCCCCTC

Table 2.5 List of oligos used for gene assembly of HB36.6

Name	Sequence
HB36.6_asbly1	CTCTGGTGGAGGCGGTAGCG
	CGTCCATCGCGTTAGACATGTGGCTAGCCGACCCTCCGCCTCC
HB36.6_asbly2	GCTACCGCCTCCACCAG
	ACATGTCTAACGCGATGGACGGTCAGCAGCTGAACCGTCTGCT
HB36.6_asbly3	GCTGGAAAACATCGGTG
	TAGGTGTACGCGTCTTTACCCAGACCGAACGGGTCCCACGCAC
HB36.6_asbly4	CGATGTTTTCCAGCAGC
	TGGGTAAAGACGCGTACACCTACCAGGCGGCGGAAGTTCTGA
HB36.6_asbly5	AAGCGGTTTACGAAACCA
	GATCCAGTGGATACGCATCGCCAGGTCGAACGCACGGTTGGTT
HB36.6_asbly6	TCGTAAACCGCTTTCAG
	GCGATGCGTATCCACTGGATCTACAACTTCGCGTTCAAACGTC
HB36.6_asbly7	GTATCCCGTTCGCGCGT
	CGCCTGTTTCAGTTCCAGCAGACGACGCGCGAATTTCTGGGTA
HB36.6_asbly8	CGCGCGAACGGGATACG

HB36.6_asbly9	TGCTGGAACTGAAACAGGCGGCGTCTTCTCCGCTGCCGCTGGA AGGTTCTCTCGAGGGGG
HB36.6_asbly10	CTTCAGAAATAAGCTTTTGTTCGGATCCGCCCCCCTCGAGAGA ACCTTCC
HB36.6_asbly1_pE TFLAG	TATAGCGATGGATTACAAGG
HB36.6_asbly2_pE TFLAG	CGTCCATCGCGTTAGACATGTGGGATCCTTTATCGTCGTCATC CTTGTAATCCATCGCTA
HB36.6_asbly9_pE TFLAG	TGCTGGAACTGAAACAGGCGGCGTCTTCTCCGCTGCCGCTGGA AGGTTCTCTCGAGCACC
HB36.6_asbly10_p ETFLAG	AGCAGCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAGAGA ACCTTCCA

Table 2.6 List of oligos used for cloning and library preparation

Name	Sequence
fwd_pETCON_col	TGACAACTATATGCGAGCAAATCCCCTCAC
rev_pETCON_col	GTACGAGCTAAAAGTACAGTGGGAAC
upGS	GGACAATAGCTCGACGATTGAAGGTAGATACCCATA
downCMyc	AGATCTCTATTACAAGTCCTCTTCAGAAATAAGCTTTTGTTC
T7_fwd	TAATACGACTCACTATAGG
T7_rev	GCTAGTTATTGCTCAGCGG
pETCON_fwd_inner_amp	CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNN GGAGGGTCGGCTAGC
pETCON_rev_inner_amp	GAGTTCAGACGTGTGCTCTTCCGATCT NNNNNNNNNNNN GCTTTTGTTCGGATCCGC
MiSeq_fwd_outer_amp	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT
Miseq_rev_outer_amp_TSBC01	CAAGCAGAAGACGGCATACGAGAT ATCACG GTGACTGGAGTTCAGACGTGTGCTCTTCCG
Miseq_rev_outer_amp_TSBC02	CAAGCAGAAGACGGCATACGAGAT CGATGT GTGACTGGAGTTCAGACGTGTGCTCTTCCG
Miseq_rev_outer_amp_TSBC03	CAAGCAGAAGACGGCATACGAGAT TTAGGC GTGACTGGAGTTCAGACGTGTGCTCTTCCG
Miseq_rev_outer_amp_TSBC04	CAAGCAGAAGACGGCATACGAGAT TGACCA GTGACTGGAGTTCAGACGTGTGCTCTTCCG
Miseq_rev_outer_amp_TSBC05	CAAGCAGAAGACGGCATACGAGAT ACAGTG GTGACTGGAGTTCAGACGTGTGCTCTTCCG
Miseq_rev_outer_amp_TSBC06	CAAGCAGAAGACGGCATACGAGAT GCCAAT GTGACTGGAGTTCAGACGTGTGCTCTTCCG
Miseq_rev_outer_amp_TSBC07	CAAGCAGAAGACGGCATACGAGAT CAGATC

	GTGACTGGAGTTCAGACGTGTGCTCTTCCG CAAGCAGAAGACGGCATACGAGAT ACTTGA
Miseq_rev_outer_amp_TSBC08	GTGACTGGAGTTCAGACGTGTGCTCTTCCG

Section 3: High-throughput motif-driven *de novo* malaria vaccine design

High-throughput motif-driven *de novo* design

Early designed binding proteins like HB80 and HB36 used natural proteins as "scaffolds" on which to add binding interactions, either in the form of individual "hotspot" residues [14], or larger continuous binding peptide "motifs"[37]. Using natural proteins limits the potential addressable problem space for binder design, since the geometry of the binding patch to be targeted may not be well-fit by the natural proteins for which crystal structures are available, or the hotspots/motifs to be grafted may not fit well in the available natural scaffold proteins. This is especially true since there is a selection bias in that available natural proteins are only those that are of interest to structural biologists and can be crystallized readily, neither of which are properties relevant to interface design. Additionally, natural proteins, being the products of evolution, often have undesirable biochemical properties that remain after design, such as low thermal stability or unwanted enzymatic activity. One means of addressing some of these deficiencies, while remaining within the framework of hotspot and motif grafting, is to generate and test for folding large sets of *de novo* proteins, and then use them as a set of available scaffolds on which to perform grafting-based design. While these *de novo* scaffold sets avoid the problem of evolutionary baggage of natural proteins, they still may have tertiary structure topologies or surface geometries that cause them to be inappropriate for some targets. In addition, these *de novo* scaffold sets are created using 'idealized' protein design rules [38,39] with minimally sized loops and a high percentage of secondary structure, which means binding motifs that are extended loops or hotspot geometries that do not fit easily onto secondary structure elements are difficult to fit into the scaffold. Because many natural protein interactions of potential therapeutic interest use extended loops for binding, a methodology that combines the

biochemical advantages of *de novo* designed proteins with the ability to accommodate the frequently irregular topologies of binding motifs should be a powerful advancement in developing *de novo* binding proteins as pharmaceuticals.

To that end, we have attempted to create such a methodology by adapting large-scale *de novo* design to build custom scaffolds around a given binding motif. Instead of starting with application-agnostic scaffolds and then attempting to accommodate a binding motif within them, our technique begins with the motif in its exact binding conformation and then uses Rosetta design to build a scaffold around it, from the beginning, in the context of binding to the target of interest. This improves both accommodation of the motif, especially motifs with extended loop conformations, and also geometric complementarity to the binding surface of the target protein.

There are many possible applications of this novel design technique. The ability to specifically tailor *de novo* scaffolds to the unique shape of the motifs they incorporate could open up multiple new avenues for drug design. A major advantage over using peptide motifs alone is that small peptides are generally unstructured or only partially structured in solution, but take on a defined structure upon binding to their targets. Ordering the peptides in this way imposes an entropic cost by reducing their degrees of conformational freedom in the bound state relative to the solution state. This entropic cost can lead to relatively weak binding of peptides to their targets. While in natural protein-protein interactions weak binding can be an asset to achieve fine-grained cellular control, many drug targets, like pathogen neutralization sites[13], basal insulin[40], or cancer-specific surface markers[41], benefit from binding that is as strong and specific as possible. For these types of targets, scaffolding existing, known protein-peptide interactions is doubly useful: first because by compensating for the entropic cost of association a larger motif-incorporating designed protein should be a stronger binder, even without additional

binding interactions other than the peptide motif of interest, and second because protein-protein interactions do not need to be designed from scratch, but can use well-validated known interactions or even mined from easily-synthesized peptide libraries. While there is a wide range of potential applications for this protein design technique, we chose to focus first on building *de novo* scaffolds around a motif from a surface protein of the *Plasmodium falciparum* parasite that may have potential as an immunogen for next-generation malaria vaccines.

Malaria vaccines as an application of high-throughput motif-driven *de novo* design

Malaria is a mosquito-borne protozoal disease that continues to inflict an enormous health burden on the world, most particularly children in Sub-Saharan Africa, with an annual death toll of somewhere between 1 and 3 million[42]. While public health measures such as bed nets, insecticide spraying, and antimalarial drugs have significantly reduced the worldwide malaria burden in recent years, political dysfunction in vulnerable regions, rising global temperatures, and increasing levels of drug resistance among circulating parasites mean that discovering an effective vaccine remains the best solution for finally eliminating the disease.

Plasmodium falciparum (Pf) circumsporozoite protein (CSP) is a promising target for pre-infection malaria vaccines. PfCSP is a 42kDa membrane protein that coats the surface of the protozoan during its sporozoite phase, the segment of the life cycle between the mosquito bite and hepatic invasion. (**Figure 3.1a**) PfCSP consists of three main segments: an N-terminal domain, a C-terminal domain, and an extended series of repeated sequence of the form Asn-Ala-Asn-Pro (NANP), with variable lengths between strains, but generally around 175 residues. (**Figure 3.1b**) PfCSP is a promising vaccine target; previous work has identified a number of anti-PfCSP antibodies that show the ability to neutralize sporozoites and prevent hepatic invasion. Deactivating sporozoites halts malaria progression to the symptomatic stage and also

prevents transmission via future mosquito bites. One major challenge is that the parasite load needed to progress to a full infection is very low- even a single un-neutralized sporozoite can lead to liver stage infection. This reality means that an effective vaccine must raise a very high titer of neutralizing antibodies that persist over a long period. The only current widely-circulated pre-infection malaria vaccine, called RTS,S, is based on a fusion of the C-terminal portion of wild-type PfCSP plus 72 residues of NANP repeat with the hepatitis B virus surface antigen (HBsAg) protein. This construct is co-expressed with free HBsAg in yeast, spontaneously forming virus-like particles (VLP), whose multivalency is thought to increase vaccine efficacy.[43] While RTS,S is able to provide some protection, it only a modest efficacy of about 35% in its target population of children, and requires boosters[44], which can be a serious challenge in the low-resource contexts in which malaria is most prevalent. While RTS,S raises high titers of anti-Pf antibodies, its weak efficacy may be due to only a small percentage of those raised antibodies being able to neutralize the sporozoite.

The importance of neutralizing antibodies for vaccine efficacy against a variety of human pathogens is well established.[45] A major challenge in designing modern vaccines, then, for the pathogens that remain resistant to traditional vaccine design strategies, is to enhance the ability of new vaccines to elicit neutralizing antibodies. One way this has been accomplished in the past is by stripping target proteins of the immunodominant domains that raise strong antibody titers but few neutralizing ones, leaving only the more vulnerable targets on the protein.[46] Another method is to engineer constructs to expose neutralization epitope targets that are ordinarily hidden from the immune system until proteolysis or conformational change just before cellular invasion[47]. While the potential of approaches like these is not nearly exhausted, vaccine engineering is frequently performed in a sequence-based manner, rather than a structural one.

In the specific case of PfCSP, the existing RTS,S vaccine is sequence-based, but there is compelling evidence to suggest that structural approaches to malaria vaccine design have been under-explored, in part due to a historical lack of structure-based design tools. Previous studies have used chemical cross-linking to stabilize short peptides based on the NANP repeat of the PfCSP, and have successfully used those immunogen constructs to raise neutralizing anti-sporozoite antibodies[48,49]. Advances in atomically-accurate structure-based *de novo* design in Rosetta, as well as in large-scale DNA synthesis and testing, present the opportunity to begin to use sophisticated computational techniques to advance large-scale structure-based vaccine design and testing.

In the following study we sought to design *de novo* proteins that can stabilize a structurally labile 15-mer epitope (NPDPNANPNVDPNAN, "peptide 21") found in the junction between the N-terminal and repeat regions of the PfCSP protein. This junctional epitope is the target of several highly potent neutralizing antibodies, but some evidence suggests that the ability of a given antibody targeting this sequence to neutralize Plasmodium sporozoites depends on the epitope's specific structural conformation[50,51]. One such study has described two different anti-PfCSP antibodies which both target this specific junctional epitope with similar binding strength, but one antibody is neutralizing and the other is not. X-ray crystallography revealed that while the epitope sequence recognized by the two antibodies is identical, the two antibodies bind to different structural conformations of the epitope (**Figure 3.1c**). Molecular dynamic simulations show that the conformation bound by the neutralizing antibody is the much rarer one in solution. Assuming the epitope's solution conformations are found in similar ratios when the epitope is incorporated in the PfCSP protein, limited epitope residence time in the neutralizing conformation might lead to the relatively low fraction of neutralizing antibodies raised against

this area of the PfCSP protein, and correspondingly low efficacy of immunization. We hypothesized, therefore, that a *de novo* protein capable of stabilizing and presenting the junctional epitope in the conformation recognized by these neutralizing antibodies might be able to test the currently unproven relative contributions of sequence and structural binding effects, and if structure is the more important, raise a higher fraction of neutralizing antibodies than the currently-existing RTS,S vaccine, particularly when displayed multivalently on a VLP.

Computational design strategy

While the pfCSP junctional motif (peptide 21) is targeted and bound by CIS42 and CIS43 with similar strengths, it adopts two different structural conformations in each binding interaction. Molecular dynamics simulations revealing that the conformation bound by CIS43 is a rare one in solution relative to that bound by the non-neutralizing CIS42 antibody[50]. Rosetta energy calculations supported both of these results, finding the CIS42-binding conformation to be more energetically favorable, but antibody-motif interaction energies to be roughly the same. The challenge turning peptide 21 into a specific immunogen in the CIS43-binding conformation is therefore not to optimize for binding affinity, but rather to use a well-structured *de novo* scaffold to overcome the energetic barrier of taking on the disfavored peptide conformation.

The β-1 turn structure of CSP NANP repeat motifs like peptide 21 does not lend itself to motif grafting methodologies that have been used in the past for design of binding proteins[37]. In general, successful binding proteins have a high degree of secondary structure, and the fragment to be replaced is itself a canonical piece of secondary structure[4]. To overcome this limitation, we constructed *de novo* protein scaffolds around the peptide 21 motif, using a similar blueprint-based methodology described previously[38,52]. This approach allows us to create highly non-canonical structures that nevertheless are designed with the same stability/foldability metrics that

previous authors have determined are key to successful protein design. A key innovation we applied is to automatically generate blueprint topologies, rather than manually curating them, by using motif grafting to determine where in a given topology the motif might comfortably fit, generalizing that topology with the grafted motif into a Rosetta blueprint, and using the blueprint to design many new proteins of the same topology.

The peptide 21 motif was matched for fit against 1984 previously reported small *de novo* protein scaffolds by motif grafting, yielding 2633 matches (each scaffold could be matched in multiple locations). A typical match from this algorithm consists of either one of the input alpha helices or one of the beta strands entirely replaced by the peptide 21 motif. The matched scaffold, complete with the motif included, is then analyzed to determine the Ramachandran angles of each residue and decomposed into a generic Rosetta blueprint. These blueprints are then run through the RosettaRemodel *de novo* fragment assembly protocol, which attempts to build a coherent protein structure that does not clash with itself but still meets compactness metrics as well as forms proper hydrogen bonding patterns and pairing between the secondary structure elements. If successful, the algorithm returns a polyvaline backbone scaffold that can then be used for downstream sequence design.

Since the motif was not perturbed at all from its position in the CIS43-peptide21 crystal structure during the backbone design process, we could test completed scaffolds for fit with CIS43 by simply superimposing the structure with the Fab from the crystal structure and checking for excessive clashing. While it is likely the case that a successful immunogen design need not bind CIS43 in particular, but only display the correctly shaped peptide 21 epitope in a binding-competent position, we would not have been able to test any proteins that didn't bind CIS43, hence they were eliminated at this stage. After filtering, we were left with about 45000

designed *de novo* scaffolds, overwhelmingly of the 3-helix bundle topology, with one helix replaced by the β turn peptide 21 motif. As a final step, the 10,000 backbones with the greatest Rosetta-calculated shape complementarity (all 3-helix bundles) were fed back into the fragment picker as seeds for another set of blueprints in which the original backbone remains fixed, but an extra helix is added to either the N- or C- terminus, and again filtered for compactness and clashing, yielding a four-helix bundle with one helix replaced by the motif.

On the assumption that very small proteins with a large disordered region would not remain folded without disulfide staples, we designed all of the scaffolds to incorporate between one and two disulfide bonds. These bonds were allowed to be placed anywhere in the sequence outside of the peptide21 motif, as well as positions within the motif itself that had been found to not affect CIS43 binding by alanine scanning. Otherwise, we designed sequences by standard published Rosetta FastDesign methods and filtered designed sequences for their overall Rosetta calculated stability as measured by total folding free energy, as well as binding affinity at least as strong as the original peptide21 motif.

High throughput synthesis and screening

After sequence design, we reverse-translated the 10948 designed protein sequences with the best Rosetta energies into DNA, and ordered their synthesis as an array-synthesized oligonucleotide pool. We PCR-amplified this pool and transformed it into yeast cells with a plasmid coding for inducible surface display. After determining that the k_D of binding of the pool in the yeast surface display system against both antibodies is roughly 500pM (**Figure 4.1**), we performed FACS against a fluorescently tagged version of CIS43 at 500pM. Any sequences that were successfully synthesized should contain the peptide21 motif, which we already know to bind CIS43, so the rationale of this first sort was to filter out mis-synthesized genes or designs

that lock the peptide21 motif into the CIS42 conformation. The resulting library from this first sort was then sorted against fluorescently labeled CIS42 and CIS43 at three different concentrations each, 5nM, 500pM, and 50pM, in order to select for antigen designs that show differential binding against the two antibodies. (**Figures 4.2**) We additionally performed a FACS assay with the library sorted against three different concentrations of trypsin, in order to select for designs that were stable to proteolytic degradation (**Figure 4.3**). All sorts were performed in triplicate to improve statistical robustness against experimental error. Sorting revealed that even at picomolar concentrations of IgG, nearly the entire library showed strong binding to IgG43, while a significant fraction of the library showed no apparent binding at concentrations up to 5nM (**Figure 3.3a**).

As the next step we extracted plasmids from each input and sorted pool, and performed high-throughput sequencing to count relative representation of each design. Enrichment scores of designs in each condition were calculated by simply averaging the proportion of the design in each condition across all three biological replicates, and calculating the slope of the change in representation from the input pools with increasing sort stringency (increasing concentrations of CIS42, decreasing concentrations of CIS43). This analysis revealed that designs mainly fell into two populations: those with roughly equal enrichment against CIS42 and CIS43 and those with only significant enrichment against CIS43. (**Figure 3.3b**) In order to deconvolute these two groups from one another we used a Gaussian mixture model trained on strength of enrichment against CIS42 to assign designs to their most probable group. Of the 10593 designs that were assigned to clusters, 3595 (33.9%) were of the cluster with no significant CIS42 binding, while the remaining 6998 (66.1%) showed equal binding to both antibodies (**Figure 3.3c**).

In order to test which Rosetta score components were most important in establishing the difference between these two clusters, we tested combinations of score terms as binary classifiers in logistic regressions for their ability to discriminate between the two clusters. A combination of six Rosetta score terms had roughly 75% accuracy in discriminating between CIS43-specific and nonspecific designs. A single score term measuring disulfide bond quality, was able to achieve nearly 60% discrimination (**Figure 3.3d**). In addition to logistic regression highlighting the importance of disulfide bonding for successfully ablating CIS42 binding, multiple sequence alignments of the two clusters were performed with Clustal Omega,[53] followed by hidden Markov model-based sequence motif generation using WebLogo[54]. These sequence motifs revealed no significant sequence-based patterns in the two clusters except for a very strong tendency for non-CIS42 binders to have a cysteine mutation in the fourth position of the peptide 21 motif, and a second cysteine positioned six residues after the motif (**Figure 4.4**). In fact, this pattern was found in 3305/3691 (89.5%) of the designs in the CIS 43-specific cluster, and 2244/7036 (31.8%) of the designs in the non-specific cluster. Those six residues in successful immunogen design models make up two helical turns and form a disulfide bond with the cysteine at the beginning of the peptide 21 epitope, physically constraining the majority of the motif into the CIS43-binding conformation by way of forcing the motif termini to be a certain distance apart from one another (**Figure 4.5**).

Individual characterization of designed immunogens

We chose five individual sequences for experimental characterization from those that showed no enrichment in the CIS42 screen (i.e. those designs whose library prevalence dropped to zero after sorting even with the highest concentration of CIS42 IgG), high enrichment scores in the CIS43 screen, and high enrichment values in the trypsin screen (**Figure 4.4**). Three

sequences were chosen from the set of sequences with the longer αααα topology and two sequences with the shorter ααα topology. After synthesizing each sequence using PCR assembly of 60-mer single-stranded oligonucleotides (IDT), each sequence was transformed into a yeast surface display plasmid construct in the pETCON3 plasmid in EBY100 yeast, and tested for individual antigenicity against CIS42 and CIS43. We also cloned the same five genes into pET29b plasmids (Novagen), expressed the proteins individually via autoinduction in Lemo21 E. coli, purified them using Ni-NTA affinity purification, and tested them for antigenicity against CIS42 and CIS43 in solution by both ELISA and biolayer interferometry (BLI). All three methodologies showed very strong binding to CIS43 and negligible binding to CIS42 (**Figure 3.4**).

The five chosen antigen constructs are very small proteins (5-6.5kDa). It is therefore unlikely that they would persist long enough in an animal after intramuscular injection to raise a significant immune response and we would therefore be unable to test the structural immunization hypothesis. To remedy this size limitation, as well as potentially improve the immunogenicity of the designed antigens by virtue of high multivalency, we designed fusion constructs to incorporate each of the five *de novo* antigens by genetic fusion into one component of two different previously described *de novo* two-component self-assembling icosahedral virus-like particles (VLPs), I53-40 and I53-47[55]. Each individual component of these constructs was expressed in E. coli using IPTG induction, purified using Ni-NTA affinity purification followed by size exclusion chromatography (SEC). The two components were then mixed together to assemble them into mature particles and purified once more by SEC, from which VLP assembly efficiency could be measured. Once assembled, these constructs yield particles with masses of between 2.3 and 2.8MDa and 60 copies of the immunogen per assembly. The immunogen-

displaying VLPs show similar antibody-binding patterns to the free immunogens, though with stronger binding against all antibodies, likely due the avidity effects of having 60 immunogens in close proximity.

Discussion and future work

In this study we have shown a highly automated and massively parallel approach for successful scaffolding and stabilization of an interesting and potentially important structural conformation of a highly irregular protein motif that is not stable or persistent on its own in solution. The results obtained from this particular application in this advance in protein design methodology help to validate the hypothesis that the neutralization properties of certain antibodies that bind to the junctional epitope region of the pfCSP antibody are not due to different binding strengths, but rather to the specific arrangement of the protein when they bind. In addition to showing the intended designed binding pattern against CIS42 and CIS43, testing with additional neutralizing and non-neutralizing antibodies against the pfCSP protein shows that the designed immunogens show little to no binding to other known non-neutralizing antibodies, but intriguingly they do show binding to two additional antibodies that have been previously reported to be neutralizing, including one that is traditionally thought to bind to the NANP repeat region of the protein, not the junctional epitope[56]. While not conclusive, these results suggest that the mechanism of neutralization for antibodies other than CIS43 might also be mediated by structure rather than sequence, or even that cross-reactivity with the CIS43-binding conformation of the junctional epitope is an important mechanism of neutralization for all neutralizing repeat region-targeting antibodies.

This work was presaged by a series of experiments in which structural information was used to create covalently-stabilized versions of the NANP β-turn motif from the central repeat

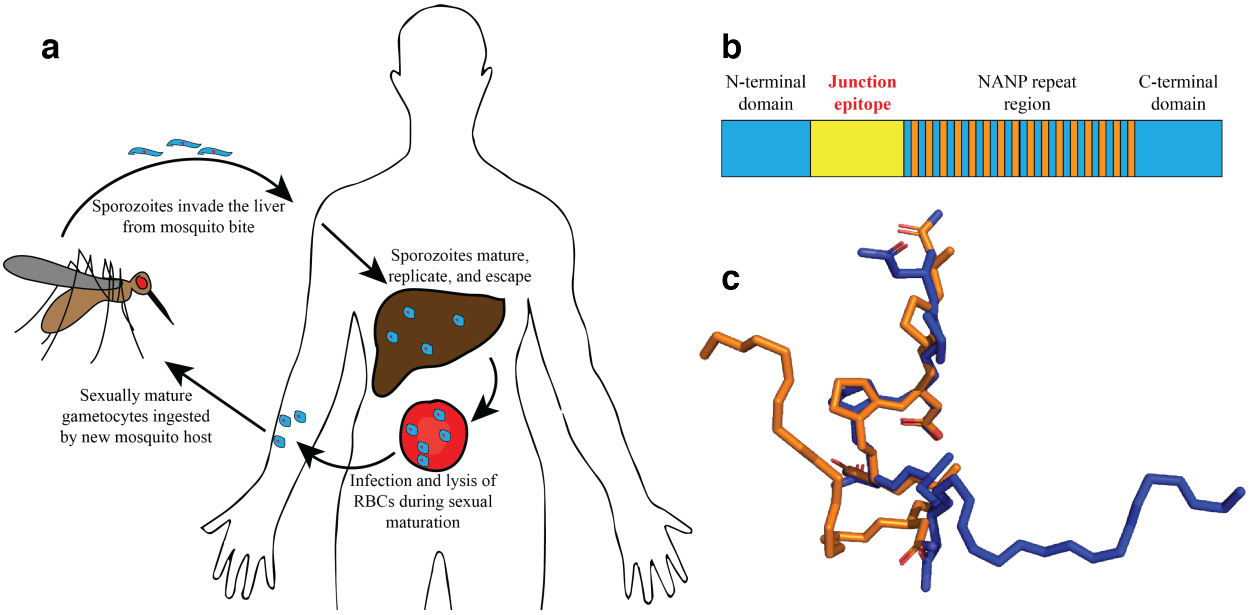
region of the circumsporozoite protein. [48,57] Polyclonal sera from rabbits immunized with these constructs showed strong cross-reactivity with the structured pfCSP protein. In light of these past studies in which covalent cross-linking showed such successes, it is particularly interesting that the placement of disulfides turned out to be such an important component of successful motif stabilization. Our methodology, which incorporates covalent cross-linking along with modern protein design, could be the final step to move towards a fully functional structure-based malaria immunogen.

In order to test this hypothesis, additional work will be required. We have begun efforts to immunize mice with the VLP-fused versions of these immunogens, as well as additional VLPs displaying only peptide21 or truncated versions of the entire wild-type pfCSP protein. After immunization, the mice will be bled and their total IgG fraction extracted. We will use the same designed immunogens as probes to detect "CIS43-like" antibodies and see if they represent a higher fraction of the total IgG fraction in the mice immunized with designed immunogens vs. those immunized with sequence-based CSP constructs. The final test will then be to see whether the mouse serum is able to protect human liver cells *in vitro* from invasion by P. falciparum sporozoites, and whether the mice themselves are more protected from infection by P. berghei (the rodent strain of malaria) than traditionally immunized mice.

While the hypothesis that a protein stabilizing the pfCSP junctional epitope in the "neutralization-competent" conformation would raise a higher titer of neutralizing antibodies than sequence-based vaccines or pfCSP on its own remains untested, we are optimistic that these constructs, if displayed multivalently and mixed with the right adjuvants, have a possibility of being the first in a new generation of rationally designed vaccines that rely on display of known structural epitopes, rather than previous structural approaches, to generate high quality immune

responses in patients. Beyond the narrow application of scaffolding the pfCSP epitope, being able to specifically tailor *de novo* scaffolds to the unique shape of the motifs they incorporate could open up many new avenues for drug design. Many biological interactions are mediated through small peptides and scaffolding them to reduce the entropic cost of adopting their binding conformation can help achieve strong and specific binding proteins from relatively limited structural information.

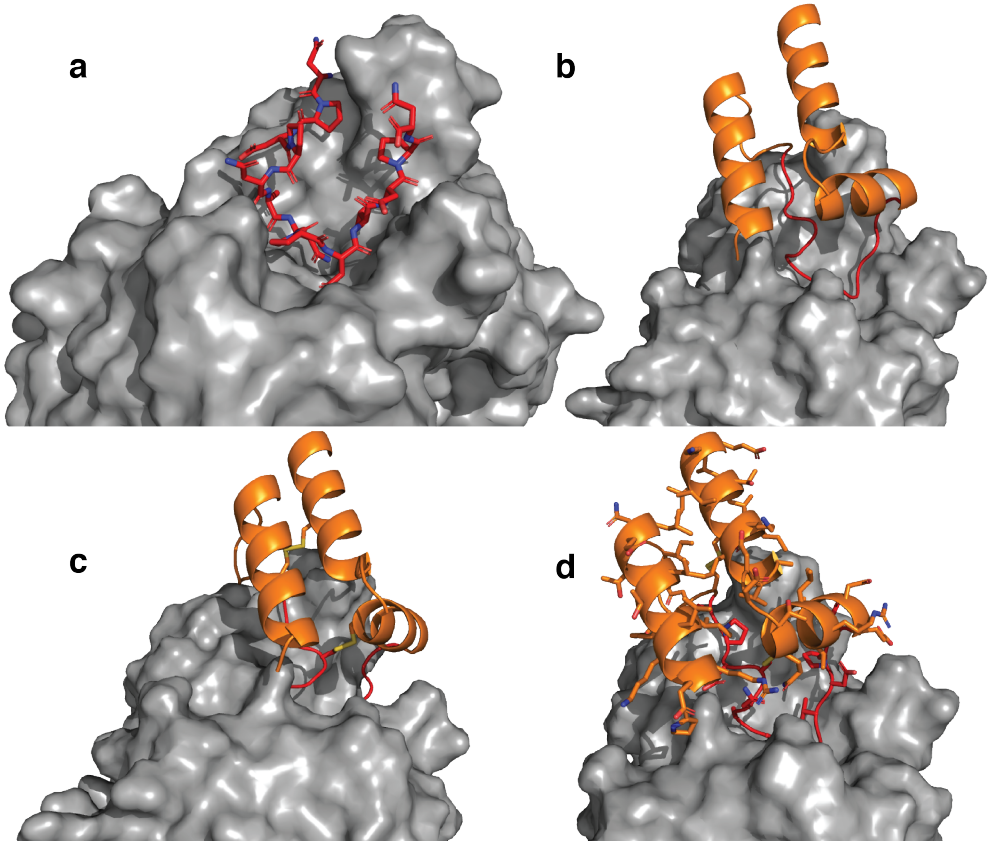
Figure 3.1



Overview of the circumsporozoite protein target motif. **a,** A simplified outline of the P. falciparum life cycle, showing the sporozoite life stage during which the CSP protein mediates liver invasion. **b,** While the exact structure of the CSP protein is not known, it can be roughly divided into several domains. The junctional epitope against which the neutralizing antibody CIS43 shows binding is a region between the N-terminal domain and the long stretch of NANP repeats in the middle of the protein. **c,** Structural comparison of the 14-mer junctional epitope in

the CIS43-binding (orange) and CIS42-binding (blue) conformations. The first seven residues are identically aligned within the margin of error of the crystal structures, but the second seven take on two different β-turn conformations.

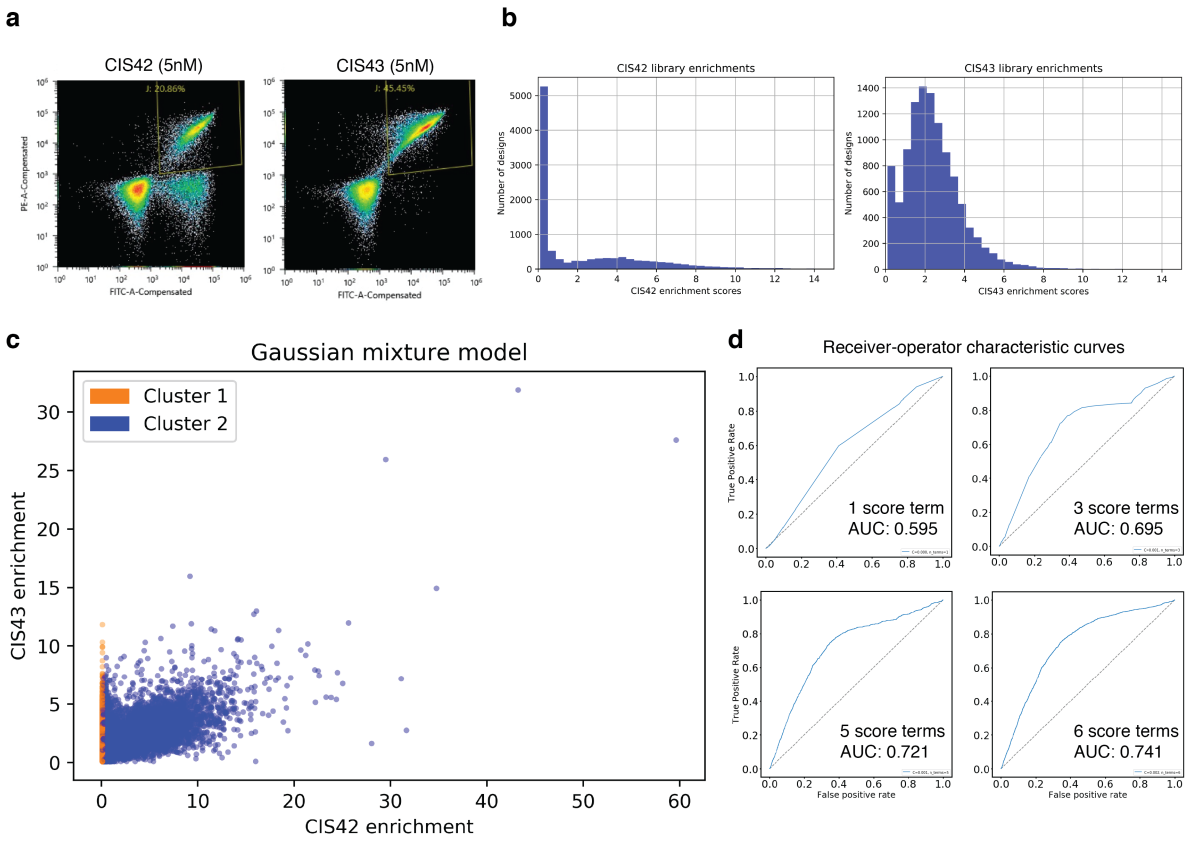
Figure 3.2



Overview of the motif-driven *de novo* design protocol. The crystal structure **a,** of the CSP motif in complex with CIS43 is prepared for Rosetta design and incorporated into **b,** a *de novo* designed polyvaline backbone constructed by RosettaRemodel fragment assembly from a topology-specifying blueprint file. Backbones that do not clash with CIS43 are stapled with **c,** 1-2 disulfides in either *de novo* positions or motif positions previously found to be unimportant for

binding affinity by alanine scanning. Finally, positions other than the motif and disulfides are **d,** designed to maximize folding free energy and maintain binding energy.

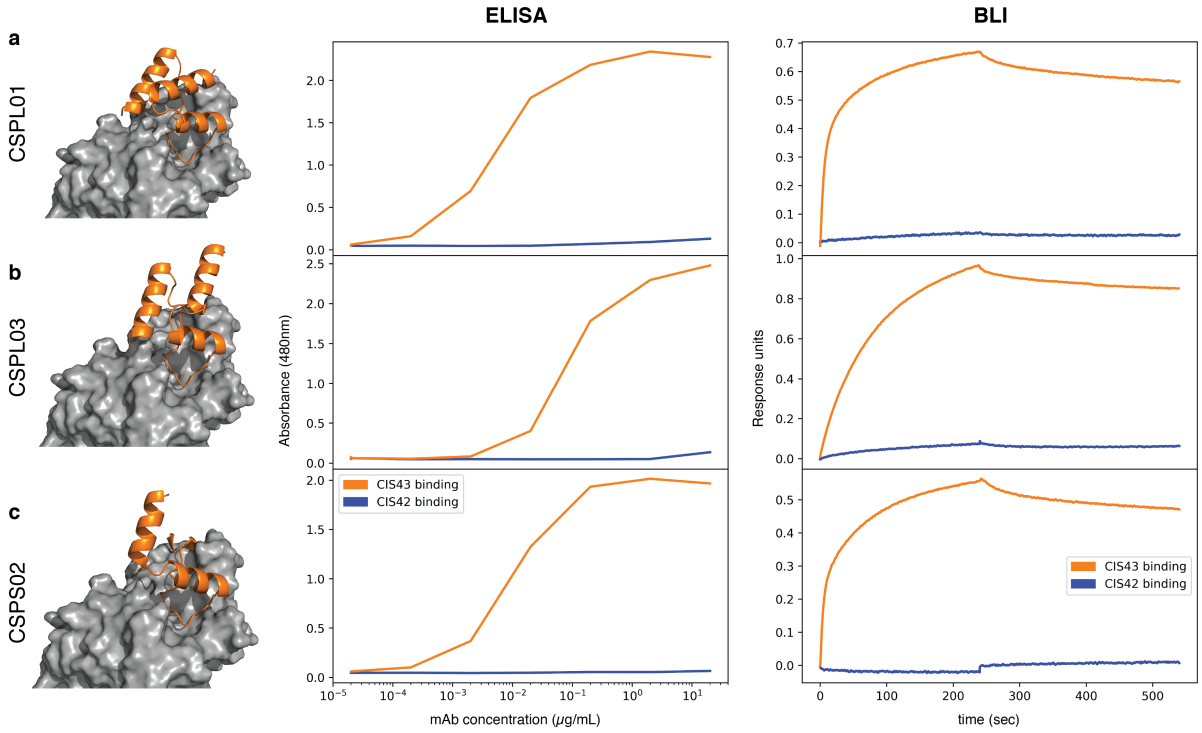
Figure 3.3



Testing designs by FACS sorting and deep sequencing. The pool of designs was tested by FACS against both CIS42 and CIS43 antibodies, showing **a,** almost universal binding to relatively low concentrations of CIS43, which is to be expected since all designs include a CIS43-binding epitope. Only 2/3 of the pool, however, is seen to bind to CIS42 at the same concentration of antibody, which should only be the case if the binding epitope is being held only in the CIS43-binding conformation. After deep sequencing, analysis of enrichment scores indicate **b,** that CIS43 enrichment shows a single distribution, while CIS42 shows a much more

bimodal pattern, with one distribution centered about an enrichment score of zero. Clustering samples based on their CIS42 enrichment using a Gaussian mixture model **c,** shows two primary groups defined almost entirely by whether they show zero enrichment against CIS42 or not. Binary classifiers built from combinations of Rosetta score terms and tested in logistic regression models can distinguish between these two clusters with very high accuracy, as shown by receiver operating characteristic plots **d,** using a combination of ten score terms, but also achieve better than even discrimination using only one score term measuring the quality of disulfide bond geometry.

Figure 3.4



Biochemical testing of selected immunogen designs. Computational design models (left), ELISA binding curves (middle), and biolayer interferometry (BLI) binding curves (right) are shown for **a,** CSPL01 **b,** CSPL03 and **c,** CSPS02 immunogens. Both ELISA and BLI show strong binding for all three immunogens against the CIS43 antibody, but weak or negligible binding against the CIS42 antibody.

Section 4: Supplementary Information for high-throughput motif-driven *de novo* malaria vaccine design

Materials and Methods

Energetic Analysis of Circumsporozoite Protein Motifs

We confirmed that this result is reflected by the Rosetta energy function by calculating the Rosetta monomer energy per-residue for both peptide conformations: +0.9 REU/residue for the CIS43-bound conformation and -0.4 REU/residue for the CIS42-bound conformation after tightly heavy-atom constrained energy minimization to reduce potential high-energy crystallographic artifacts. Since the peptide is a 14-mer, this calculation represents an 18.2 REU total difference between the energies of the two conformations.

We also sought to confirm that Rosetta calculations recapitulate the similar experimentally determined binding affinities of CIS43 and CIS42 to peptide 21. The change in total energy score between the unbound and bound conformations were found to be -37.45 REU ddG for the peptide 21-CIS43 complex and -32.33 REU ddG for the peptide 21-CIS42 complex.

Scaffold topology identification and design by motif grafting

We searched existing sets of natural and designed scaffold proteins that could accommodate the CIS43-bound conformation of peptide 21 using the `MotifGraft` backbone-matching algorithm.[37] A backbone-matching search of all monomeric proteins in the PDB found no 12-14-mer segments in any deposited proteins with a backbone RMSD within 1.5Å of the peptide 21 motif. However, searching for 12-14-mers with N- and C- termini within 1.5Å of those of the peptide 21 motif and without regard for backbone clashes with the scaffolding protein discovered 26 matches in 14 different natural protein chains. (**Protocol 4.1**) Performing the same search on a set of experimentally validated stable designed protein scaffolds returned

1787 matches with a ααα topology, 88 with αββα topology, 444 with βαββ topology, and 314 with ββαββ topologies.[2] Matches that did not clash with the CIS43 Fab crystal structure were carried forward to sequence design as a comparison set against the motif-driven *de novo* design protocol's results. 102 of these passed filtering metrics and were synthesized as DNA for testing.

Generation of blueprints for fragment assembly from motif graft matching results

We decomposed matching scaffolds incorporating the peptide 21 motif into Rosetta blueprint files[52] by classifying each non-motif residue by its ABEGO bin and secondary structure type, while keeping the motif residues fixed in their exact conformations from the CIS43 crystal structure. We generated 2634 unique blueprints in this way as seeds for motif-driven *de novo* scaffold design.

***De novo* motif-driven scaffold generation**

We used the blueprints generated from motif grafting results as guides for the RosettaRemodel fragment assembly algorithm, which picked 9-mer and 3-mer fragments from deposited structures in the PDB so as to match the secondary structure and topology specified in the blueprint file for every position except for those of the peptide 21 motif, whose residue identities and ABEGO angles were drawn directly from the relaxed CIS43-peptide 21 crystal structure. Each pose returned from fragment assembly was filtered for adherence to the blueprint specification, proper pairing in hydrogen bond formation in sheets and helices, and helix kinking. Successful poses were output as polyvaline scaffolds except in motif locations, which retained their native sequences. (**Protocol 4.2**) The ααα blueprint topologies were considerably more successful than the others in generating well-formed scaffolds, with 44390 PDB file results, as compared to 158 with an αββα topology, 365 with a βαββ topology, and none with ββαββ topologies.

***De novo* sequence design**

We tested properly formed scaffolds for clash and shape complementarity against the CIS43 antibody by simply superimposing the two structures. (**Protocol 4.4**) Because the motif around which the scaffold was designed did not move at all during the design process, the scaffold PDB files already existed in the same coordinate frame as the CIS43 target, and no docking algorithm was required.

To design the sequence of superimposed polyvaline backbone scaffolds that passed the clash check with CIS43 we first used the RosettaScripts `disulfidize` application to staple the backbone with either one or two disulfide bonds. Disulfides were allowed to be placed anywhere within the non-motif backbone, as well as motif residues that had previously been found to be relatively unimportant for binding strength by alanine scanning mutagenesis. We then designed the sequence of stapled backbones outside of the motif and the designed cysteine residues using two rounds of `FastDesign` plus additional minimization steps with decreasing coordinate constraints. The backbones were designed directly in their superimposed configuration against the CIS43 Fab in order to maximize complementarity to the binding surface of the antibody, with the first round of `FastDesign` designing the entire scaffold without allowing backbone perturbation and the second round focusing only on the binder-antibody interface and allowing backbone perturbation. Allowed per-position residues were restricted to only those found canonically in natural proteins by secondary structure type and burial, as defined by number of side chain neighbors within 4Å. Rosetta score function beta_nov16 was used throughout. Sequences which passed filtering thresholds of -2.2 REU/residue total monomer score, -0.2 pabb/residue, and a ddg roughly less than or equal to the original peptide 21/CIS43 complex (-38.0 REU) passed on to the next round. (**Protocol 4.5**) We sorted designed sequences that passed

all filtering criteria by total score/residue, and chose the top 8500, all of the ααα topology, to be synthesized for testing. In addition to these short designs, the highest-scoring 10000 designs were fed back into the blueprint and scaffold generation pipeline with an additional helix added to either to N- or C- terminus, to produce longer αααα scaffolds. We performed sequence design on these scaffolds with the same method as the shorter set, and synthesized the 2346 lowest score/residue results.

Parallel gene synthesis and yeast transformation

For the 10948 designed sequences that passed all filtering metrics, we reverse-translated the amino acid sequences into DNA using custom software that accounted for yeast codon usage, repeat sequence content, and other measures of synthesis. We also added adapters for amplification and yeast display in the pETCON3 system. (**Table 4.1**)

All DNA sequences were submitted for synthesis by microarray by Agilent Biotechnologies and returned as a pool of mixed oligonucleotides. We PCR amplified the pooled sequences at large scale using KAPA HiFi polymerase (Roche) on a CFX96 thermocycler (BioRad), using primers that add additional sequence complementarity to the pETCON3 plasmid,[4] in preparation for yeast display testing (NGS_fwd_extend_adaptor and NGS_rev_extend_adaptor, **Table 4.1**). We performed a first-round 10μL scouting qPCR of 35 cycles to determine the cycle at which the reactions reached an amplification fluorescence inflection point, and then a full-size 50μL amplification reaction was performed, removing samples from the thermocycler when they reached that cycle.

	time	temperature
step 1	3:00	98°

step 2		0:20	98°
step 3		0:15	65°
step 4		0:15	72°
	GOTO 2		
step 5	34x		
step 6		5:00	72°

We separated the PCR products on a 2% agarose electrophoresis gel and extracted the amplified bands using a Qiaprep gel extraction kit (Qiagen). 12 micrograms of gel-purified DNA were co-transformed with 4 micrograms of pETCON3 double-digested by NdeI XhoI restriction enzymes (New England Biolabs) for homologous recombination into Invitrogen EBY100 yeast, according to a previously published protocol, yielding a transformation efficiency of .[58]

Yeast surface display k_D titration

Yeast clones or libraries were grown overnight to saturation in 2mL cultures of SDCAA medium in shaking incubators at 30°C. Cultures were then centrifuged for 3 minutes at 4000g to pellet, and resuspended in 2mL of SGCAA. These were then grown again to saturation overnight and harvested by centrifugation for 3 minutes at 4000g. After a wash with 1mL of ice-cold 50mM $NaPO_4$, 150mM NaCl, 1% w/v bovine serum albumin, pH 7.4 (PBSF 1%), cells were resuspended in PBSF 1% to an OD600 of 2.0, as determined with a NanoDrop 8000 spectrophotomer (Thermo Fisher). We incubated IgGs CIS42 and CIS43 with a 5:1 molar ratio of biotinylated Protein ZZ:antibody for 15 minutes on ice, and then mixed 200,000 cells from the resuspended yeast libraries/clones with Protein ZZ-labeled IgG to concentrations of 0nM, 1nM,

5nM, 20nM, and 100nM for 30m on ice. The cells were then washed with 200μL of PBSF 1% and resuspended in 50μL of PBSF 1% and 0.1μL each of fluorescein isothiocyanate (FITC)-labeled anti-Myc antibody (Abcam) and phycoerythrin (PE)-labeled streptavidin (Thermo Fisher) for an additional 10m on ice. After a second 200μL PBSF 1% wash yeast cells were kept as pellets until immediately before flow cytometry. After resuspension in 100 μL of PBSF 1%, 50,000 cells were analyzed by flow cytometry for FITC and PE fluorescence using an Accuri C6 flow cytometer. The library/clone k_D values were determined by fitting an exponential decay model to the total FITC fluorescence in each sample using SciPy Python package.

High-throughput yeast display affinity screening

Following the same growth, induction, and labeling protocol as for k_D titrations, the library of transformed antigen designs was grown in triplicate and incubated with biotin-labeled CIS42 and CIS43 IgG at various concentrations. Libraries were then sorted by FACS on a Sony SH800 cell sorter, For each of the initial 500pM filtering sorts, 35,000,000 cells were labeled and sorted, collecting all cells showing PE fluorescence above the level of the negative control (anti-Myc only, no SAPE). After a second round of growth and induction, 4,000,000 cells from each of the three sorted libraries were labeled with CIS43 and CIS42 at three different antibody concentrations each: 50pM, 500pM, and 5nM.

Yeast display protease resistance screening

Following the same growth and induction protocol as for k_D titrations, the library of transformed antigen designs was grown in triplicate and 20,000,000 cells were resuspended in PBS to an OD600 of 4.0. The libraries were then mixed with equal volumes of 0.25% trypsin-EDTA (Sigma-Aldrich) diluted to 1/3, 1/27, and 1/243, for final dilutions of 1/6, 1/56, and 1/486 and an OD600 of 2.0 in a total volume of 250μL. After 5 minutes of incubation, 1mL of ice-cold

PBSF 1% was added to stop the reaction, and cells were pelleted and resuspended in 1mL ice-cold PBSF 1% 4 times. The cells were then resuspended in 1/50 FITC-labeled anti-Myc antibody diluted in PBSF 1% and sorted by FACS on a Sony SH800. All cells showing FITC fluorescence above the level of the negative control (induced but unlabeled and untrypsinized cells) were collected.

Next-generation DNA sequencing

Plasmids were prepared from yeast libraries as previously described,[22]and genes were amplified out from the plasmids using pETCON_inner_fwd and pETCON_inner_rev (**Appendix 4.1**) primers using the same PCR protocol as described for DNA production for gene synthesis. After the first amplification, products were amplified a second time and barcoded using the miseq_outer_fwd and miseq_outer_rev primers, with barcodes TSBC01-TSBC17 and TSBC28-42 (**Appendix 4.1**). Libraries were purified by AMPure magnetic beads (New England Biolabs), quantitated using a Qubit spectrophotometer (Thermo Fisher), and sequenced in 318bp paired-end mode on a Miseq instrument (Illumina) with 600bp v3 sequencing kit, according to standard methods. The sequencing run yielded 22686833 reads passing the MCS software filters and matching any of the library barcodes within 1 mismatch. The forward and reverse reads obtained for each yeast library were fused using PEAR[59], and enrichment scores calculated using custom software.

Monomeric antigen protein production

Monomeric *de novo* antigens were produced as genetic fusions to the DsbC protein disulfide bond isomerase protein from S. cerevisiae to enhance expression and encourage proper disulfide bond formation when expressed intracellularly in E. coli. The antigen sequence was cloned into a custom vector containing an N-terminal 10x His tag, followed by a TEV protease

cleavage site, an SGG tripeptide linker, and finally the antigen itself. Following previously published autoinduction protocols,[31] the plasmid was transformed into Lemo21(DE3) E. coli cells (New England Biolabs), starter cultures were grown overnight in MDG media, and then diluted 1:1000 into 1L of ZYM-5052 media and grown in 2.8L shake flasks at 225 RPM for 24 hours at 25°C. After growth, cells were collected by centrifugation at 4000g RCF for 15m. Cell pellets were resuspended in 35mL of 50mM Tris, 500mM NaCl, 30mM imidazole, 0.75% CHAPS, pH 8.0 (lysis buffer) and lysed at 18,000 PSI in an M-110P Microfluidizer (Microfluidics).

Cell lysate was collected and His-tagged proteins were isolated by affinity chromatography by gravity flow over 2mL of freshly charged Ni-NTA agarose resin (Qiagen). Resin was washed with 10CV of lysis buffer and then protein was collected with 5CV of 50mM Tris, 500mM NaCl, 500mM imidazole, 0.75% CHAPS, pH 8.0 (elution buffer). Collected Ni-NTA purified protein was dialyzed 1:100 overnight into 100mM NaCl, 20mM NaPi, pH 7.4 (cleavage buffer). After determining dialyzed protein concentration by A280, TEV protease was added at a 1:20 ratio of protease:dialyzed protein and allowed to digest at room temperature for 2h. Cleaved antigen protein was then isolated by concentrating the sample to 2mL with Amicon 3kDa centrifugal concentrators (Millipore) and injecting it onto a Superdex S75 gel filtration column (Sigma Aldrich) on an ÄKTA Pure FPLC (GE Life Sciences) and collecting the relevant protein fraction.

Antigen-fused VLP production

Antigen-displaying VLPs were produced as genetic fusions between the antigen and either the trimeric (I53-40) or the pentameric (I53-47) component of the VLP. Constructs were cloned into pET29b plasmids with C-terminal 6x His tags. The plasmid was transformed into

Lemo21(DE3) E. coli cells (New England Biolabs), starter cultures were grown overnight in LB media, and then diluted 1:50 into 1L of LB media and grown in 2.8L shake flasks at 225 RPM at 37°C until an OD600 of 1.2 was reached. 1mM IPTG was added at this point and growth continued at 18°C (pentamers) or 37°C (trimers) overnight. After growth, cells were collected by centrifugation at 4000g RCF for 15m. Cell pellets were resuspended in 35mL of 50mM Tris, 500mM NaCl, 30mM imidazole, 0.75% CHAPS, pH 8.0 (lysis buffer) and lysed at 18,000 PSI in an M-110P Microfluidizer (Microfluidics).

Cell lysate was collected and His-tagged proteins were isolated by affinity chromatography by gravity flow over 2mL of freshly charged Ni-NTA agarose resin (Qiagen). Resin was washed with 10CV of lysis buffer and then protein was collected with 5CV of 50mM Tris, 500mM NaCl, 500mM imidazole, 0.75% CHAPS, pH 8.0 (elution buffer). After concentration to under 2mL using 10kDa centrifugal concentrators, soluble multimers were collected by injection onto a Superdex S200 gel filtration column (Sigma Aldrich) with 50mM Tris, 500mM NaCl, 0.75% CHAPS, pH 8.0 (assembly buffer) and the relevant fractions collected. Trimeric and pentameric components were then concentrated such that the total combined volume once mixed would not exceed 2mL and mixed at a 1:1 molar ratio at room temperature overnight to assemble. After assembly, a final SEC filtering run was performed on a Superose S6 column in PBS + 5% glycerol to isolate assembled cages, and the relevant fractions were collected.

ELISA antigen testing

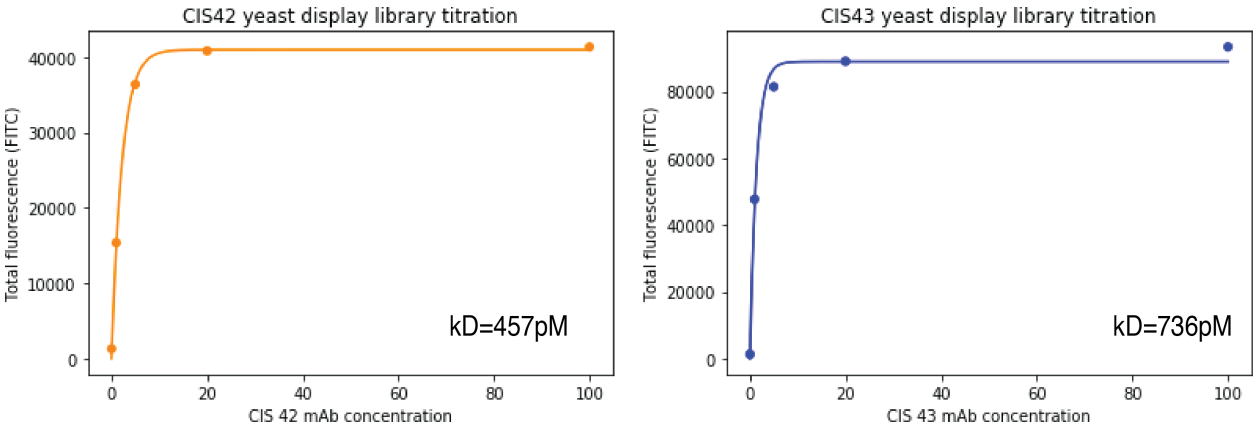
MaxiSorp ELISA plates (Thermo Scientific Nunc) were coated with 100 µl of immunogen (1 µg/ml) per well for 1 h at room temperature according to the manufacturer's instructions (KPL). Coated plates were blocked with 100 µl of 1× blocking solution for 1 h at room temperature,

which was followed by incubation with 100 μl of PfCSP monoclonal antibodies, mock transfection filtrate or control antibodies (VRC 01, a human anti-HIV-1 IgG1 as an isotype-matched negative control44; 2A10, a mouse monoclonal antibody specific for the NANP-repeat region of PfCSP20,21) at varying concentrations (0.00006–5.0 μg/ml). After 1 h, plates were incubated with 100 μl/well of 1.0 μg/ml peroxidase-labeled goat anti–human IgG antibody (KPL, cat. no. 04-10-06). Plates were washed six times with PBS-Tween between each step. After a final wash, samples were incubated for about 15 min with the ABTS peroxidase (KPL, cat. no. 50-62-10) or Ultra TMB ELISA (Invitrogen, cat. no. 34029) substrate. The optical density was read at 405 or 450 nm after addition of stopping solution (100 μl/well).

BLI antigen testing

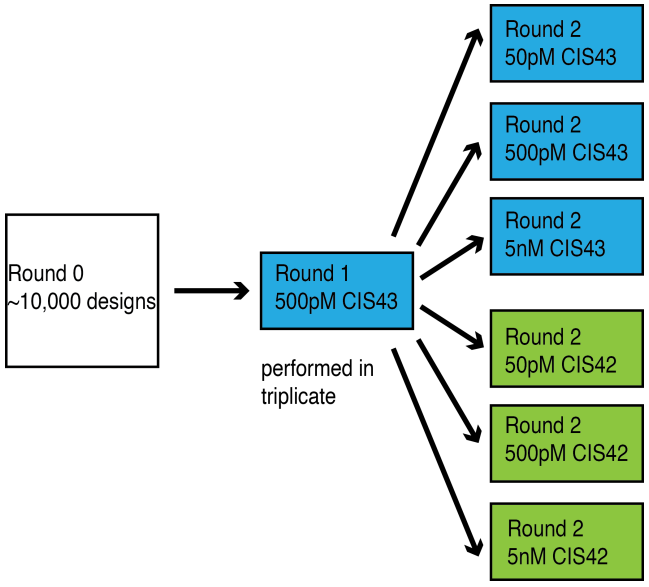
Antibody binding kinetics were measured using biolayer interferometry on an Octet Red384 instrument (fortéBio) using streptavidin-capture biosensors (fortéBio). PfCSP monoclonal antibody solutions were plated in solid black tilt-well 96-well plates (Geiger Bio-One); assays were performed with agitation at 30 °C. Loading of biotinylated rPfCSP or peptides 21 and 29 was performed for 300 s, followed by dipping the biosensors into buffer (PBS + 1% BSA) for 60 s to assess baseline assay drift. Association with whole IgG (serially diluted from 267 to 33 μM) was done for 300 s, followed by a dissociation step in buffer for 600s. Background subtraction of nonspecific binding was performed through measurement of association in buffer alone. Data analysis and curve fitting were performed using Octet software, version 7.0. Experimental data were fitted with the binding equations describing a 1:1 heterologous ligand interaction. Global analyses of the complete data sets, assuming binding was reversible (full dissociation), were carried out using nonlinear least-squares fitting allowing a single set of binding parameters to be obtained simultaneously for all concentrations of a given monoclonal antibody dilution series.

Figure 4.1



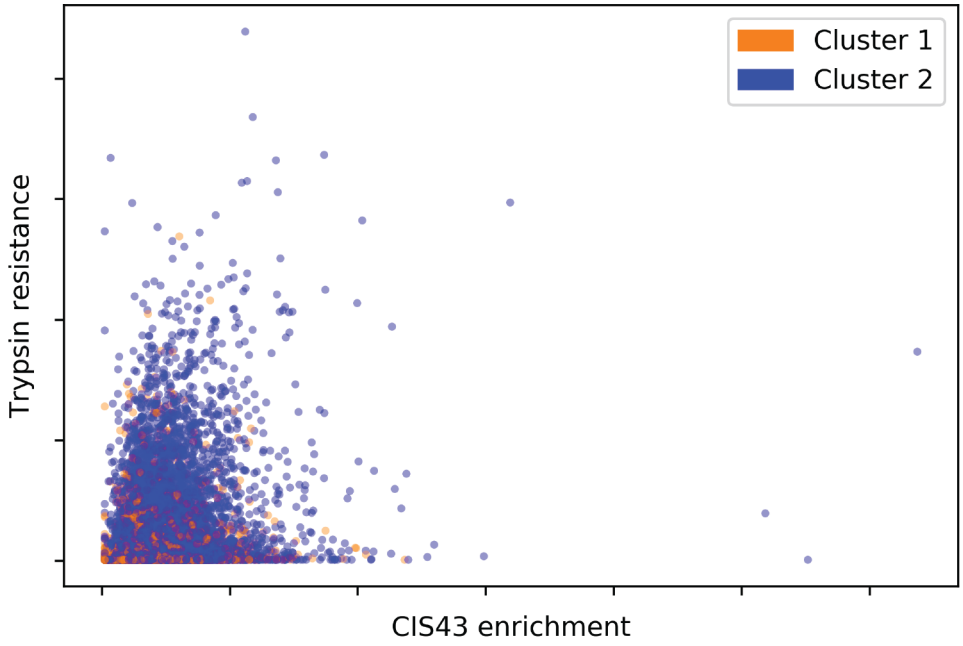
k_D titrations of designed immunogen libraries displayed on yeast. Yeast cells induced to display the design library were incubated with either CIS42 or CIS43 IgG at concentrations of 1,5,20, and 100nM. Exponential decay curves were then fit by regression to the total FITC fluorescence at each concentration, as measured by an Accuri C6 flow cytometer. K_Ds were calculated as the half-maximal value of the curve fit.

Figure 4.2



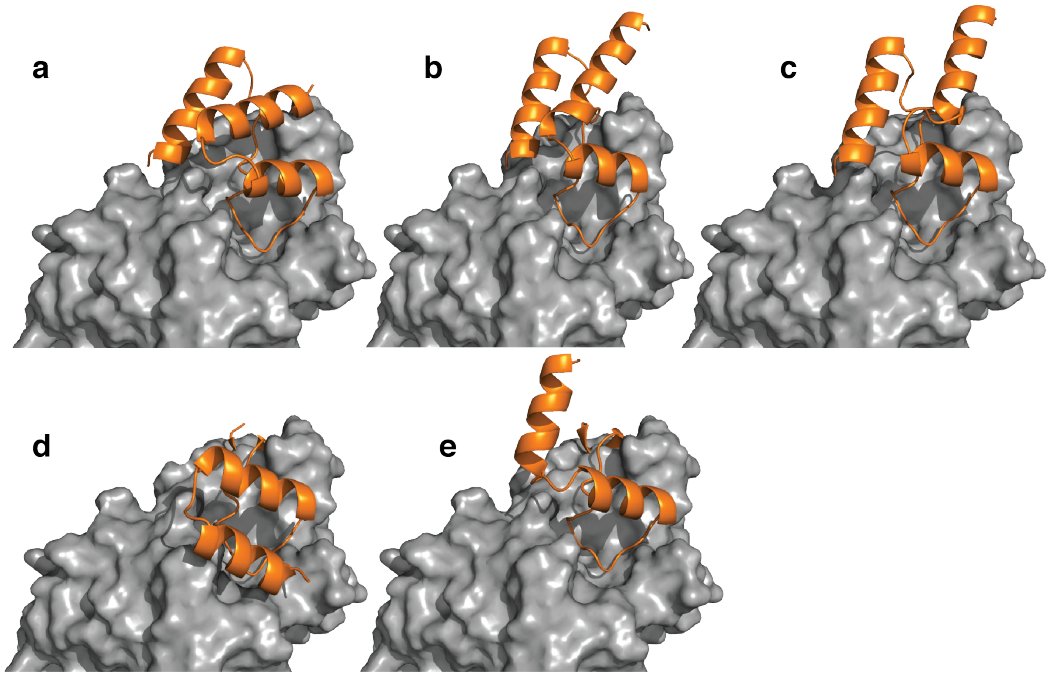
Yeast display FACS sorting strategy. An initial filtering sort at 500pM was performed to remove mis-synthesized immunogen designs and CIS42-specific designs. The results of that first sort were then sorted against three concentrations each of CIS42 and CIS43. All sorts were performed as triplicate biological replicates.

Figure 4.3



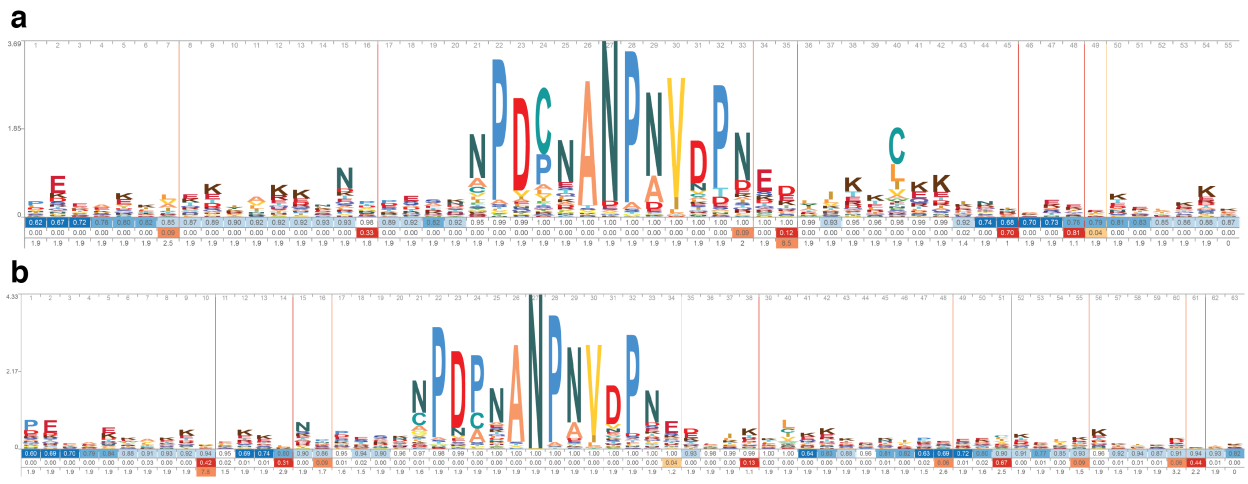
Trypsin resistance of designed immunogen libraries. FACS sorting for C-Myc display at three different concentrations of trypsin was used to estimate immunogen design protein stability. Comparing this stability estimate to CIS43 enrichment shows no clear patterns. Data points are colored by their cluster assignment in the Gaussian mixture clustering model.

Figure 4.4



Design models of the five immunogens selected for individual characterization. a, CSPL01, **b,** CSPL02, **c,** CSPL03, **d,** CSPS01, **d,** CSPS02 (orange) in complex with the crystal structure model of CIS43 (grey).

Figure 4.5



Hidden Markov model sequence lgoos of specific vs. nonspecific binders. a, CIS43-specific binders show a strong tendency towards having a cysteine in the fourth position of the CSP motif sequence, as well as a second cysteine six residues (two helical turns) after the end of the CSP motif. In design models, these two cysteines form a disulfide bond to constrain the CSP motif into the CIS43-binding arrangement. **b,** Non-CIS43-specific binders show no strong sequence patterns other than the peptide 21 motif itself.

Table 4.1: Illumina sequencing barcodes and read numbers for each sorted FACS pool.

FACS pool	Truseq Barcode	# reads
agilent_sept_17_naive_S42	TSBC42	1782905
replicate_1_round_1_IgG43_500pM_S27	TSBC27	924879
replicate_1_round_2_1-243_trypsin_S35	TSBC35	835665
replicate_1_round_2_1-27_trypsin_S34	TSBC34	759088
replicate_1_round_2_1-3_trypsin_S33	TSBC33	870264
replicate_1_round_2_display_S30	TSBC30	860099
replicate_1_round_2_IgG42_500pM_S2	TSBC02	336009
replicate_1_round_2_IgG42_50pM_S1	TSBC01	678867
replicate_1_round_2_IgG42_5nM_S3	TSBC03	373629
replicate_1_round_2_IgG43_500pM_S11	TSBC11	361808
replicate_1_round_2_IgG43_50pM_S10	TSBC10	415313
replicate_1_round_2_IgG43_5nM_S12	TSBC12	429372
replicate_2_round_1_IgG43_500pM_S28	TSBC28	769171
replicate_2_round_2_1-243_trypsin_S38	TSBC38	715099
replicate_2_round_2_1-27_trypsin_S37	TSBC37	759607
replicate_2_round_2_1-3_trypsin_S36	TSBC36	757908
replicate_2_round_2_display_S31	TSBC31	658642
replicate_2_round_2_IgG42_500pM_S5	TSBC05	360113
replicate_2_round_2_IgG42_50pM_S4	TSBC04	342011
replicate_2_round_2_IgG42_5nM_S6	TSBC06	391974
replicate_2_round_2_IgG43_500pM_S14	TSBC14	715132
replicate_2_round_2_IgG43_50pM_S13	TSBC13	646539
replicate_2_round_2_IgG43_5nM_S15	TSBC15	592424
replicate_3_round_1_IgG43_500pM_S29	TSBC29	852233
replicate_3_round_2_1-243_trypsin_S41	TSBC41	868300
replicate_3_round_2_1-27_trypsin_S40	TSBC40	797866
replicate_3_round_2_1-3_trypsin_S39	TSBC39	710543
replicate_3_round_2_display_S32	TSBC32	697513
replicate_3_round_2_IgG42_500pM_S8	TSBC08	440680
replicate_3_round_2_IgG42_50pM_S7	TSBC07	339549

replicate_3_round_2_IgG42_5nM_S9	TSBC09	420382
replicate_3_round_2_IgG43_500pM_S17	TSBC17	892679
replicate_3_round_2_IgG43_50pM_S16	TSBC16	676149
replicate_3_round_2_IgG43_5nM_S18	TSBC18	654421

Table 4.2: DNA adapter and PCR primer sequences

pETCON3 5' adapter
GGGTCGGCTTCGCATATG
pETCON3 3' adapter
CTCGAGGGTGGAGGTTCC
NGS_fwd_extend_adaptor
AGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTTCGCATATG
NGS_rev_extend _adaptor
GATCTCTATTACAAGTCCTCTTCAGAAATAAGCTTTTGTTCGGAACCTCCACCCTCGAG
pETCON_inner_fwd
CCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNGGAGGGTCGGCTAGC
pETCON_inner_rev
GAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNNNNGCTTTTGTTCGGATCCGC
miseq_outer_fwd
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
miseq_outer_rev_TSBC01
CAAGCAGAAGACGGCATACGAGATATCACGGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC02
CAAGCAGAAGACGGCATACGAGATCGATGTGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC03
CAAGCAGAAGACGGCATACGAGATTTAGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC04
CAAGCAGAAGACGGCATACGAGATTGACCAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC05
CAAGCAGAAGACGGCATACGAGATACAGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC06
CAAGCAGAAGACGGCATACGAGATGCCAATGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC07
CAAGCAGAAGACGGCATACGAGATCAGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC08
CAAGCAGAAGACGGCATACGAGATACTTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC09
CAAGCAGAAGACGGCATACGAGATGATCAGGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC10
CAAGCAGAAGACGGCATACGAGATTAGCTTGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC11
CAAGCAGAAGACGGCATACGAGATGGCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC12
CAAGCAGAAGACGGCATACGAGATCTTGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC13

CAAGCAGAAGACGGCATACGAGATAGTCAAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC14
CAAGCAGAAGACGGCATACGAGATAGTTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC15
CAAGCAGAAGACGGCATACGAGATATGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC16
CAAGCAGAAGACGGCATACGAGATCCGTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC17
CAAGCAGAAGACGGCATACGAGATGTAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC18
CAAGCAGAAGACGGCATACGAGATGTCCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC27
CAAGCAGAAGACGGCATACGAGATATTCCTGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC28
CAAGCAGAAGACGGCATACGAGATCAAAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC29
CAAGCAGAAGACGGCATACGAGATCAACTAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC30
CAAGCAGAAGACGGCATACGAGATCACCGGGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC31
CAAGCAGAAGACGGCATACGAGATCACGATGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC32
CAAGCAGAAGACGGCATACGAGATCACTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC33
CAAGCAGAAGACGGCATACGAGATCAGGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC34
CAAGCAGAAGACGGCATACGAGATCATGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC35
CAAGCAGAAGACGGCATACGAGATCATTTTGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC36
CAAGCAGAAGACGGCATACGAGATCCAACAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC37
CAAGCAGAAGACGGCATACGAGATCGGAATGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC38
CAAGCAGAAGACGGCATACGAGATCTAGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC39
CAAGCAGAAGACGGCATACGAGATCTATACGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC40
CAAGCAGAAGACGGCATACGAGATCTCAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC41
CAAGCAGAAGACGGCATACGAGATGACGACGTGACTGGAGTTCAGACGTGTGCTCTTCCG
miseq_outer_rev_TSBC42
CAAGCAGAAGACGGCATACGAGATTAATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCG

Table 4.3: Amino acid sequences of tested designs

CSPL01	PDEAKIAQTLVEEFNDSSSINPDCNANPNVDPNEDIITLCREIKDLKDILTAVL

	KAAK
CSPL02	PEDADIAKKCIKELQVEPGFNPDCNANPNVDPNETWIKLCQKSLDIICKIVKT ALKEK
CSPL03	PERAKLIKTLATELNIPDDCNPDCNANPNVDPNEDVTRLCQTDDVAALKICIT LIKEI
CSPS01	PDGSNPDCNANPNVDPNEKWWQECLERLKVPEEVKTAVQEAIV
CSPS02	PESSSNPDCNANPNVDPNEDLIKKCEKINVPTEEIKKEIEEKK

Protocol 4.1: MotifGraft_2017.xml

```
<ROSETTASCRIPTS>
	<SCOREFXNS>
		<ScoreFunction name="SFXN_beta_nov_15_mod" weights="beta_nov15"
symmetric="0" >
			<Reweight scoretype="p_aa_pp" weight="0.0"/>
			<Reweight scoretype="rama_prepro" weight="0.0"/>
		</ScoreFunction>
		<ScoreFunction name="SFXN_beta15_vanilla_cart_cst"
weights="beta_nov15_cart.wts" symmetric="0" >
			<Reweight scoretype="coordinate_constraint" weight="0.01"/>
		</ScoreFunction>
		<ScoreFunction name="SFXN_beta15_vanilla_cart"
weights="beta_nov15_cart.wts" symmetric="0" />
		<ScoreFunction name="SFXN_beta15_vanilla" weights="beta_nov15.wts"
symmetric="0" />
	</SCOREFXNS>
	<MOVERS>
		//Here is the mover that generates multiple pose results
		<MotifGraft name="motif_grafting"
			context_structure="%%contextpdb%%"
			motif_structure="%%motifpdb%%"
			RMSD_tolerance="1.5"
			NC_points_RMSD_tolerance="1.5"
			clash_score_cutoff="10"
			clash_test_residue="ALA"
			hotspots="3:9"
			combinatory_fragment_size_delta="1:2"
			max_fragment_replacement_size_delta="0:2"
			full_motif_bb_alignment="0"
			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="0"
			allow_repeat_same_graft_output="0" />
		//HERE You add  all the other design movers
```

```
<MultiplePoseMover name="MPM_design" max_input_poses="10">
<SELECT>
</SELECT>
        <ROSETTASCRIPTS>
            <SCOREFXNS>
                <ScoreFunction name="SFXN_beta_nov_15_mod"
weights="beta_nov15" >
                    <Reweight scoretype="p_aa_pp" weight="0.0"/>
                    <Reweight scoretype="rama_prepro"
weight="0.0"/>
                </ScoreFunction>
                <ScoreFunction name="SFXN_beta15_vanilla_cart_cst"
weights="beta_nov15_cart.wts" symmetric="0" >
                <Reweight scoretype="coordinate_constraint"
weight="0.01"/>
            </ScoreFunction>
                <ScoreFunction name="SFXN_beta15_vanilla_cart"
weights="beta_nov15_cart.wts" symmetric="0" />
                <ScoreFunction name="SFXN_beta15_vanilla"
weights="beta_nov15.wts" symmetric="0" />
            </SCOREFXNS>
            <TASKOPERATIONS>
                <ProteinInterfaceDesign name="pack_long"
design_chain1="0" design_chain2="0" jump="1" interface_distance_cutoff="15"/> //
            </TASKOPERATIONS>
            <RESIDUE_SELECTORS>
          <Not name="not_context" >
            <ResiduePDBInfoHasLabel property="CONTEXT"/>
          </Not>
        </RESIDUE_SELECTORS>
            <MOVERS>
                <SwitchChainOrder name="keep_only_chain_A"
chain_order="1"/>
                <SwitchChainOrder name="keep_only_chain_B"
chain_order="2"/>
                <TaskAwareMinMover name="min"
scorefxn="SFXN_beta15_vanilla" bb="0" chi="1" task_operations="pack_long"/>
            </MOVERS>
            <FILTERS>
                <ScoreType name="totalScore"
scorefxn="SFXN_beta15_vanilla" score_type="total_score" threshold="0.0" confidence="0"/>
                <Ddg name="ddg"  scorefxn="SFXN_beta15_vanilla"
threshold="-35.0" jump="1" repeats="5" repack="1" relax_mover="min" confidence="1" />
                <ResidueCount name="nres" confidence="0" />
                <Geometry name="omega" omega="165"
cart_bonded="5000" residue_selector="not_context" confidence="0"/>
```

```
                                        <ShapeComplementarity name="sc_filt" jump="1"
verbose="1" min_sc="0.70" write_int_area="0" confidence="1" />
                                        <FragmentLookupFilter name="faulty_fragments"
lookup_name="source_fragments_4_mer"
store_path="/lab/databases/VALL_clustered/backbone_profiler_database_06032014"
lookup_mode="first" chain="2" threshold="1"  confidence="0" />
                                        <Sasa name="interface_buried_sasa" threshold="600"
confidence="1"/>
                                        <BuriedUnsatHbonds2 name="interface_unsat_hbond2"
cutoff="2" confidence="0" jump_number="1"/>
                                        SSPrediction name="sspred"
cmd="/work/dadriano/PROGRAMS/psipred/runpsipred_single" use_probability="0"
use_svm="0" threshold=0.80 confidence="0"/>
                                        <CalculatorFilter name="ddg_nm2" equation="ddg /
((sasa/100.0)+0.001)" threshold="-2.8" confidence="0" >
                                                <Var name="ddg" filter="ddg"/>
                                                <Var name="sasa" filter="interface_buried_sasa"/>
                                         </CalculatorFilter>
                                        <CalculatorFilter name="score_res"
equation="totalScore/nres" threshold="0.0" confidence="0" >
                                <Var name="totalScore" filter="totalScore"/>
                                <Var name="nres" filter="nres"/>
                        </CalculatorFilter>
                                        <CalculatorFilter name="score_res_active"
equation="totalScore/nres" threshold="-2.0" confidence="1" >
                                                <Var name="totalScore" filter="totalScore"/>
                                                <Var name="nres" filter="nres"/>
                                        </CalculatorFilter>
                                        <MoveBeforeFilter name="score_res_chainB"
mover="keep_only_chain_B" filter="score_res_active" confidence="1" />
                                        MoveBeforeFilter name="sspred_chainB"
mover="keep_only_chain_B" filter="sspred" confidence="0"/> IF YOU USE THIS run only one
design per work directory at a time or your reults will MESS-UP
                                </FILTERS>
                                <TASKOPERATIONS>
                                        <InitializeFromCommandline name="init"/>
                                        <IncludeCurrent name="current" />
                                        <ConsensusLoopDesign
name="disallow_nonnative_loop_sequences"/>
                                        <RestrictIdentities name="notGPH" identities="PRO,HIS"
prevent_repacking="0" />
                                        <RestrictToRepacking name="onlyRepack" />
                                        <OperateOnResidueSubset name="hotspot_onlyrepack">
                                                <ResiduePDBInfoHasLabel
property="HOTSPOT"/>
                                                <RestrictToRepackingRLT/>
```

```
                </OperateOnResidueSubset>
                OperateOnResidueSubset name="scaffold_onlyrepack">
                        ResiduePDBInfoHasLabel
property="SCAFFOLD"/>
                        RestrictToRepackingRLT/>
                /OperateOnResidueSubset>
                OperateOnResidueSubset name="motif_norepack">
            ResiduePDBInfoHasLabel property="MOTIF"/>
            PreventRepackingRLT/>
        /OperateOnResidueSubset>
                <OperateOnResidueSubset name="context_norepack">
                        <ResiduePDBInfoHasLabel
property="CONTEXT"/>
                        <PreventRepackingRLT/>
                </OperateOnResidueSubset>
                DisallowIfNonnative name="only_native_GPH"
disallow_aas="GPH"/>

                <LimitAromaChi2 name="limitaro" />  #avoids extra
rotamers
                <RestrictIdentities name="noCYS" identities="CYS"
prevent_repacking="1" />
                ProteinInterfaceDesign name="PID" repack_chain1="0"
repack_chain2="1" design_chain1="0" design_chain2="1" interface_distance_cutoff="10.0" />
                <OperateOnResidueSubset name="pidAB">
                        <Not>
                                <InterfaceByVector cb_dist_cut="11.0"
nearby_atom_cut="5.5" vector_angle_cut="75.0" vector_dist_cut="9.0">
                        <Chain chains="1"/>
                        <Chain chains="2"/>
                </InterfaceByVector>
                        </Not>
                        <PreventRepackingRLT/>
                </OperateOnResidueSubset>
                <OperateOnResidueSubset name="not_pidAB">
                        <InterfaceByVector cb_dist_cut="11.0"
nearby_atom_cut="5.5" vector_angle_cut="75.0" vector_dist_cut="9.0">
                    <Chain chains="1"/>
                    <Chain chains="2"/>
                </InterfaceByVector>
                        <PreventRepackingRLT/>
                </OperateOnResidueSubset>
                <OperateOnResidueSubset name="pidMotifScaffold">
                <Not>
                    <InterfaceByVector cb_dist_cut="11.0"
nearby_atom_cut="5.5" vector_angle_cut="75.0" vector_dist_cut="9.0">
```

```
                                        <ResiduePDBInfoHasLabel property="MOTIF"/>
                                        <Chain chains="2"/>
                                    </InterfaceByVector>
                                </Not>
                                <PreventRepackingRLT/>
                            </OperateOnResidueSubset>
                                    <LayerDesign name="layer_design_core" layer="core"
core_E="20" core_H="20" core_L="25" surface_E="50" surface_H="50" surface_L="26"
make_pymol_script="0">
                                        <core>
                                            <all append="MAFILVWY"
exclude="CERKHGQ"/> #PDNST left normal
                                        </core>
                                        <boundary>
                                            <all append="ACDEIKLNQRVYWF"
exclude="HMST"/> #PG left alone
                                        </boundary>
                                        <surface>
                                            <all append="ACDEHKNQRSTY"
exclude="MVILWF" /> #PG left alone
                                        </surface>
                                    </LayerDesign>
                                </TASKOPERATIONS>
                                <FILTERS>
                                </FILTERS>
                                <MOVERS>
                                    <RollMover name="random_rb_perturb" chain="2"
random_roll="1" random_roll_angle_mag="3.0" random_roll_trans_mag="0.3" />
                                    <AddConstraintsToCurrentConformationMover
name="constrainCA" CA_only="1" cst_weight="0.01" use_distance_cst="0" />
                                    <ClearConstraintsMover name="clearAllConstraints" />
                                    <PackRotamersMover name="pack_sc_chainAB"
scorefxn="SFXN_beta15_vanilla"
task_operations="init,current,limitaro,onlyRepack,pidAB,hotspot_onlyrepack" />
                                    <PackRotamersMover name="design_interface_hard_mod"
scorefxn="SFXN_beta_nov_15_mod"
task_operations="init,current,limitaro,disallow_nonnative_loop_sequences,noCYS,pidAB,conte
xt_norepack,hotspot_onlyrepack" />
                                    <PackRotamersMover
name="design_interface_with_scaffold_hard_mod" scorefxn="SFXN_beta_nov_15_mod"
task_operations="init,current,disallow_nonnative_loop_sequences,limitaro,noCYS,pidMotifScaf
fold,context_norepack,hotspot_onlyrepack" />
                                    <PackRotamersMover name="design_binder_core_hard"
scorefxn="SFXN_beta15_vanilla"
task_operations="init,current,disallow_nonnative_loop_sequences,limitaro,noCYS,layer_design
_core,not_pidAB,context_norepack,hotspot_onlyrepack" />
```

```
                                        <TaskAwareMinMover name="min_kine_sc_chainB"
scorefxn="SFXN_beta15_vanilla" bb="1" chi="1" jump="0" cartesian="0"
task_operations="init,current,limitaro,onlyRepack,context_norepack,hotspot_onlyrepack" />
                                        <TaskAwareMinMover name="min_kine_sc_chainAB"
scorefxn="SFXN_beta15_vanilla" bb="0" chi="1" jump="0" cartesian="0"
task_operations="init,current,limitaro,onlyRepack,pidAB,hotspot_onlyrepack" />
                                        <TaskAwareMinMover name="min_cart_cst_chainB"
scorefxn="SFXN_beta15_vanilla_cart_cst" bb="1" chi="1" jump="1" cartesian="1"
tolerance="0.001" max_iter="1000"
task_operations="init,current,limitaro,onlyRepack,context_norepack,hotspot_onlyrepack" />
                                        <TaskAwareMinMover name="min_cart_chainB"
scorefxn="SFXN_beta15_vanilla_cart" bb="1" chi="1" jump="1" cartesian="1"
tolerance="0.001" max_iter="1000"
task_operations="init,current,limitaro,onlyRepack,context_norepack,hotspot_onlyrepack" />
                                        <FastRelax name="fast_relax_b_beta"
scorefxn="SFXN_beta15_vanilla"
task_operations="init,current,limitaro,onlyRepack,context_norepack,hotspot_onlyrepack" >
                                                <MoveMap name="mappyfd">
                                                        <Chain number="1" chi="0" bb="0"/>
                                                        <Chain number="2" chi="1" bb="1"/>
                                                        <Jump number="1" setting="1"/>
                                                </MoveMap>
                                        </FastRelax>
                                </MOVERS>
                                <APPLY_TO_POSE>
                                </APPLY_TO_POSE>
                                <PROTOCOLS>
                                        **DESIGN STEP1>
                                        <Add mover_name="random_rb_perturb"  />
                        <Add mover_name="pack_sc_chainAB" />
                                        **pre-Filter By the size of the iface>
                                        <Add  filter="interface_buried_sasa" />
                                        **design >
                                        <Add mover_name="design_interface_hard_mod" />
                                        <Add
mover_name="design_interface_with_scaffold_hard_mod" />
                                        <Add mover_name="design_binder_core_hard" />
                                        **relax >
                                        <Add mover_name="constrainCA" /> #Constraint CAs
                                                <Add mover_name="min_cart_cst_chainB" />
                                        <Add mover_name="clearAllConstraints" /> #Remove
constraints to CAs
                                        <Add mover_name="pack_sc_chainAB" />
                                        <Add mover_name="fast_relax_b_beta" />
                                        **filter Scaffold geometry >
                                        Add  filter="omega" />
```

```
                    Add  filter="faulty_fragments"/>

                                **DESIGN STEP2>
                                <Add mover_name="design_interface_hard_mod" />
                                <Add
mover_name="design_interface_with_scaffold_hard_mod" />
                                <Add mover_name="design_binder_core_hard" />
                                <Add mover_name="constrainCA" /> #Constraint CAs
                                        <Add mover_name="min_cart_cst_chainB" />
                                <Add mover_name="clearAllConstraints" /> #Remove
constraints to CAs

                                **RELAX STEP3>
                                Add mover_name="design_binder_core_hard" />
                    <Add mover_name="constrainCA" /> #Constraint CAs
                        <Add mover_name="min_cart_cst_chainB" />
                    <Add mover_name="clearAllConstraints" /> #Remove constraints to
CAs
                                <Add mover_name="pack_sc_chainAB" />
                                <Add mover_name="fast_relax_b_beta" />

                                **RELAX STEP4>
                                <Add mover_name="min_kine_sc_chainB" />
                                <Add mover_name="min_kine_sc_chainAB" />
                                <Add mover_name="constrainCA" />
                                <Add mover_name="min_cart_chainB" />
                                <Add mover_name="clearAllConstraints" />
                                <Add mover_name="fast_relax_b_beta" />

                                **FILTER STEP5
                                <Add  filter="omega" />
                                <Add  filter="faulty_fragments"/>
                                <Add  filter="interface_unsat_hbond2" />
                                <Add  filter="sc_filt" />
                                <Add  filter="interface_buried_sasa" />
                                <Add  filter="totalScore" />
                                <Add  filter="nres" />
                                <Add  filter="score_res" />
                                <Add  filter="score_res_chainB" />
                                <Add  filter="ddg" />
                                <Add  filter="ddg_nm2" />
                            </PROTOCOLS>
                        </ROSETTASCRIPTS>
                </MultiplePoseMover>

        </MOVERS>
```

```
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>

        <PROTOCOLS >
        //HERE you combine everything together
        <Add mover_name="motif_grafting" />
        <Add mover_name="MPM_design" />
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

Protocol 4.2: blueprintbuilder_2017.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
   <ScoreFunction name="sfxn_std"  weights="beta_nov15"/>
        <ScoreFunction name="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" />
        </ScoreFunction>

        <ScoreFunction name="SFXN2" 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" />
        </ScoreFunction>


</SCOREFXNS>
<FILTERS>

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

  ResidueCount name="cys_count" residue_types="CYD" confidence="0" />
  <ResidueCount name="res_count" confidence="0" />

   <CalculatorFilter name="bb" equation="hbond / rescount" threshold="-0.23" confidence="1">
      <Var name="hbond" filter="hbond_sfn"/>
      <Var name="rescount" filter="res_count"/>
```

```
   </CalculatorFilter>

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

   CalculatorFilter name=mean_dslf equation="dslf / cyscount" threshold="-0.35"
confidence=1>
      VAR name="dslf" filter="dslf_fa13"/>
      VAR name="cyscount" filter="cys_count"/>
   /CalculatorFilter>

         <HelixKink name="hk1" bend="40" blueprint="%%blueprint%%" />
         SheetTopology name="sf1" blueprint="%%blueprint%%" />
         HelixPairing name="hf1" blueprint="%%blueprint%%" />
         <SecondaryStructure name="ss1" blueprint="%%blueprint%%" />
         <CompoundStatement name="cs1">
                  <AND filter_name="ss1" />
                  <AND filter_name="hk1" />
                  AND filter_name="sf1" />
                  AND filter_name="hf1" />
         </CompoundStatement>

      <Holes name="holes" threshold="6.0" confidence="0" />


         <ScoreType name="total_score_cen" score_type="total_score" scorefxn="SFXN2"
confidence="0" threshold="0" />

   <AveragePathLength name="apl" max_path_length="999999" path_tightness="5"/> /change
max_path_length to 8 (or lower)
   <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"/>

</FILTERS>
<TASKOPERATIONS>
   <LimitAromaChi2 name="limitchi2" include_trp="1" />
   <LayerDesign name="layer_all" layer="core_boundary_surface_Nterm_Cterm"
use_sidechain_neighbors="True" pore_radius="2.0" verbose="true" />
   <NoRepackDisulfides name="exemptdisulf" />
  <OperateOnCertainResidues name="keep_pd1">
   <ResidueIndexIs indices="31,32,33,34,35,36"/>   all plus 297
   <PreventRepackingRLT/>
  </OperateOnCertainResidues>
</TASKOPERATIONS>
```

```
<MOVERS>
	<Dssp name="dssp" />

	SheetCstGenerator name="sheet_new1" cacb_dihedral_tolerance="0.6"
blueprint="%%blueprint%%" />
	RemoveCsts name="sheet_rm1" generator="sheet_new1" />
	<SetSecStructEnergies name="set_ssene1" scorefxn="SFXN1"
blueprint="%%blueprint%%" />
	<BluePrintBDR name="bdr1" use_abego_bias="1" scorefxn="SFXN1"
constraints_NtoC="-1.0" blueprint="%%blueprint%%" />
	<DumpPdb name="dump0" fname="%%blueprint%%pass_ss" tag_time="True"/>
	<DumpPdb name="dump" fname="%%blueprint%%pass_cs" tag_time="True"/>
	<DumpPdb name="dump2" fname="%%blueprint%%pass_degree" tag_time="True"/>


	<ParsedProtocol name="build_dssp1" >
		<Add mover_name="bdr1" />
		<Add mover_name="dssp" />
		<Add filter_name="ss1" />
	Add mover_name="dump0" />
		<Add filter_name="hk1" />
		Add filter_name="sf1" />
		Add filter_name="hf1" />
	<Add mover_name="dump" />
		<Add filter_name="degree" />
	<Add mover_name="dump2" />
	</ParsedProtocol>


	<LoopOver name="lover1" mover_name="build_dssp1" iterations="1000" drift="0"
ms_whenfail="FAIL_DO_NOT_RETRY" />
	<ParsedProtocol name="phase1" >
		<Add mover_name="set_ssene1" />
		<Add mover_name="lover1" />
	</ParsedProtocol>

</MOVERS>

<PROTOCOLS>
	<Add mover_name="phase1" />
	<Add mover_name="dssp" />
	<Add filter_name="total_score_cen" />
	<Add filter_name="cs1" />
</PROTOCOLS>
</ROSETTASCRIPTS>
```

Protocol 4.3: seq_design_for_motifgraft.xml

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="nov16_std" weights="beta_nov15.wts" />
        <ScoreFunction name="nov16_cst" weights="beta_nov15_cst.wts" />
        <ScoreFunction name="nov16_soft" weights="beta_nov15_soft.wts" />
    </SCOREFXNS>
      <TASKOPERATIONS>
            <RestrictIdentitiesAtAlignedPositions name="no_design_graft"
prevent_repacking="1" source_pdb="%%graft_pdb%%" resnums="%%graft_nodesign%%"
chain="1" />
      </TASKOPERATIONS>
    <RESIDUE_SELECTORS>

        <ResidueName name="polar_selector"
residue_name3="SER,THR,ASP,GLU,HIS,TYR,ASN,GLN,CYS,LYS,ARG,TRP" />
        <Chain name="chainA_selector" chains="1" />
        <Chain name="chainB_selector" chains="2" />

            <Task name="graft_nodesign" designable="False" packable="False"
fixed="True" task_operations="no_design_graft" />
            <Task name="graft_designable" designable="True" fixed="False"
task_operations="no_design_graft" />

        <InterfaceByVector name="interface_selector" grp1_selector="chainA_selector"
grp2_selector="chainB_selector"/>

            <And name="not_graft" selectors="graft_designable,chainA_selector"/>

    </RESIDUE_SELECTORS>
    <TASKOPERATIONS>
        <IncludeCurrent name="include_curr"/>

        <OperateOnResidueSubset name="repack_chainB" selector="chainB_selector">
<RestrictToRepackingRLT/> </OperateOnResidueSubset>

        <InitializeFromCommandline name="init" />
        <LimitAromaChi2 name="limitchi2" include_trp="1" />
        <IncludeCurrent name="inclcur" />
        <ExtraRotamersGeneric name="exrot" ex1="1" ex2="1" extrachi_cutoff="1" />


    </TASKOPERATIONS>
    <FILTERS>
            <Sasa name="interface_buried_sasa" threshold="1200" confidence="0"/>
    </FILTERS>
```

```
<MOVERS>
        <SwitchChainOrder name="switch_order" chain_order="21"/>

    <RigidBodyPerturbNoCenter name="random_perturb" />

    <DumpPdb name="dumpraw" fname="raw.pdb" scorefxn="nov16_std" />
    DumpPdb name="dumprelax" fname="%%bp_file%%_prelax" tag_time="True"
scorefxn="nov16_std" />

    <VirtualRoot name="AddVRoot" removable="true" />
    <VirtualRoot name="RemVRoot" remove="true" />

    <AddConstraints name="add_constraints" >
        <CoordinateConstraintGenerator name="coordinate_constraints"
sidechain="false" ca_only="true" />
    </AddConstraints>

    <RemoveConstraints name="remove_constraints"
constraint_generators="coordinate_constraints" />

    <Disulfidize name="add_disulfs" scorefxn="nov16_std" set1="not_graft"
set2="not_graft" min_disulfides="1" max_disulfides="2" match_rt_limit="2.0"
max_disulf_score="1.5" min_loop="8" mutate_gly="true" mutate_pro="true" />

    <FastRelax name="scaffold_relax" scorefxn="nov16_cst" repeats="4"
task_operations="init,limitchi2,inclcur,repack_chainB">
        <MoveMap name="chainA_only">
            <Chain number="1" chi="true" bb="true"/>
            <Chain number="2" chi="true" bb="false"/>
        </MoveMap>
    </FastRelax>

    FastDesign name="design1" scorefxn="nov16_soft" repeats="3"
task_operations="init,limitchi2,inclcur,repack_chainB,repack_chainC,core_sheet_res,core_helix
_res,core_loop_res,boundary_sheet_res,boundary_helix_res,boundary_loop_res,surface_sheet_re
s,surface_helix_res,surface_loop_res,no_design_graft" >
        MoveMap name="chainA_only">
            Chain number="1" chi="true" bb="false"/>
            Chain number="2" chi="true" bb="false"/>
            Chain number="3" chi="true" bb="false"/>
        /MoveMap>

    /FastDesign>

    <MultiplePoseMover name="MPM_design" max_input_poses="100">
        <ROSETTASCRIPTS>
```

```
<SCOREFXNS>

    <ScoreFunction name="nov16_std"  weights="beta_nov15.wts" />
    <ScoreFunction name="nov16_soft" weights="beta_nov15_soft.wts" />
    <ScoreFunction name="nov16_cart" weights="beta_nov15_cart.wts" />

        <ScoreFunction name="VDW" weights="empty" >
        <Reweight scoretype="fa_atr" weight="1.0"/>
        </ScoreFunction>

        <ScoreFunction name="TotalHydrophobic"
weights="/work/grocklin/gabe_scripts/total_hydrophobic_weights.wts"/>

</SCOREFXNS>

  <TASKOPERATIONS>
      <RestrictIdentitiesAtAlignedPositions name="no_design_graft"
prevent_repacking="1" source_pdb="%%graft_pdb%%" resnums="%%graft_nodesign%%"
chain="1" />
</TASKOPERATIONS>

<RESIDUE_SELECTORS>

        <Task name="graft_selector" designable="False"
packable="False" fixed="True" task_operations="no_design_graft" />
        <Task name="graft_designable" designable="True" fixed="False"
task_operations="no_design_graft" />

        ResiduePDBInfoHasLabel name="graft_selector"
property="MOTIF" />

    <Chain name="chainA_selector" chains="1" />
    <Chain name="chainB_selector" chains="2" />

        <Neighborhood name="chainA_within_10"
selector="chainA_selector" distance="10" />
        <Neighborhood name="chainB_within_10"
selector="chainB_selector" distance="10" />

        <And name="chainA_interface"
selectors="chainB_within_10,chainA_selector" />
        <And name="chainB_interface"
selectors="chainA_within_10,chainB_selector" />
```

```
                                        <Not name="not_chainA_interface"
selector="chainA_interface" />
                                        <And name="chainA_notinterface"
selectors="not_chainA_interface,chainA_selector" />

                                        <Not name="not_chainB_interface"
selector="chainB_interface" />
                                        <And name="chainB_notinterface"
selectors="not_chainB_interface,chainB_selector" />

                                <Layer name="core_selector" select_core="true"
use_sidechain_neighbors="true"/>
                                <Layer name="boundary_selector" select_boundary="true"
use_sidechain_neighbors="true"/>
                                <Layer name="surface_selector" select_surface="true"
use_sidechain_neighbors="true"/>

                                <SecondaryStructure name="sheet_selector" ss="E" use_dssp="true"/>
                                <SecondaryStructure name="helix_selector" ss="H" use_dssp="true"/>
                                <SecondaryStructure name="loop_selector" ss="L" use_dssp="true"/>

                                <And name="core_sheets"
selectors="sheet_selector,core_selector,chainA_selector"/>
                                <And name="core_helix"
selectors="helix_selector,core_selector,chainA_selector"/>
                                <And name="core_loops"
selectors="loop_selector,core_selector,chainA_selector"/>
                                <And name="boundary_sheets"
selectors="sheet_selector,boundary_selector,chainA_selector"/>
                                <And name="boundary_helix"
selectors="helix_selector,boundary_selector,chainA_selector"/>
                                <And name="boundary_loops"
selectors="loop_selector,boundary_selector,chainA_selector"/>
                                <And name="surface_sheets"
selectors="sheet_selector,surface_selector,chainA_selector"/>
                                <And name="surface_helix"
selectors="helix_selector,surface_selector,chainA_selector"/>
                                <And name="surface_loops"
selectors="loop_selector,surface_selector,chainA_selector"/>

                        </RESIDUE_SELECTORS>

                        <TASKOPERATIONS>

                                ### don't design the target
                                <OperateOnCertainResidues name="restrict_chA">
```

```
                                        <ChainIs chain="A"/>
                                        <RestrictToRepackingRLT/>
                                        </OperateOnCertainResidues>


                                        ### don't design the target
                                        <OperateOnCertainResidues name="restrict_chB">
                                        <ChainIs chain="B"/>
                                        <RestrictToRepackingRLT/>
                                        </OperateOnCertainResidues>

                        <OperateOnResidueSubset name="core_sheet_res"
selector="core_sheets" > <RestrictAbsentCanonicalAASRLT aas="FILVWY"/>
</OperateOnResidueSubset>
                        <OperateOnResidueSubset name="core_helix_res"
selector="core_helix" > <RestrictAbsentCanonicalAASRLT aas="AFILVWY"/>
</OperateOnResidueSubset>
                        <OperateOnResidueSubset name="core_loop_res"
selector="core_loops" > <RestrictAbsentCanonicalAASRLT aas="AFILPVWY"/>
</OperateOnResidueSubset>

                        <OperateOnResidueSubset name="boundary_sheet_res"
selector="boundary_sheets" > <RestrictAbsentCanonicalAASRLT
aas="DEFIKLNQRSTVWY"/> </OperateOnResidueSubset>
                        <OperateOnResidueSubset name="boundary_helix_res"
selector="boundary_helix" > <RestrictAbsentCanonicalAASRLT
aas="ADEIKLNQRSTVWY"/> </OperateOnResidueSubset>
                        <OperateOnResidueSubset name="boundary_loop_res"
selector="boundary_loops" > <RestrictAbsentCanonicalAASRLT
aas="ADEFGIKLNPQRSTVWY"/> </OperateOnResidueSubset>

                        <OperateOnResidueSubset name="surface_sheet_res"
selector="surface_sheets" > <RestrictAbsentCanonicalAASRLT aas="DEHKNQRST"/>
</OperateOnResidueSubset>
                        <OperateOnResidueSubset name="surface_helix_res"
selector="surface_helix" > <RestrictAbsentCanonicalAASRLT aas="DEHKNQRST"/>
</OperateOnResidueSubset>
                        <OperateOnResidueSubset name="surface_loop_res"
selector="surface_loops" > <RestrictAbsentCanonicalAASRLT aas="DEGHKNPQRST"/>
</OperateOnResidueSubset>

                                        <OperateOnResidueSubset
name="repack_chainA_noninterface" selector="chainA_notinterface">
                                                <RestrictToRepackingRLT/>
                                        </OperateOnResidueSubset>
```

```
                        <OperateOnResidueSubset name="repack_chainB_interface"
selector="chainB_interface">
                                <RestrictToRepackingRLT/>
                        </OperateOnResidueSubset>

                                <OperateOnResidueSubset
name="norepack_chainB_noninterface" selector="chainB_notinterface">
                                        <PreventRepackingRLT/>
                                </OperateOnResidueSubset>

                                <OperateOnResidueSubset name="repack_graft"
selector="graft_selector">
                                        <RestrictToRepackingRLT/>
                                </OperateOnResidueSubset>

                        <InitializeFromCommandline name="init" />
                        <LimitAromaChi2 name="limitchi2" include_trp="1" />
                        <IncludeCurrent name="inclcur" />
                        <ExtraRotamersGeneric name="exrot" ex1="1" ex2="1"
extrachi_cutoff="1" />
                        <RestrictIdentities name="dont_design_CPH"
identities="PRO,HIS,CYS" prevent_repacking="0" />

                        <LayerDesign name="layer_all"
layer="core_boundary_surface_Cterm_Nterm" repack_non_design="True" core_H="20"
core_E="20" core_L="25" surface_H="60" surface_E="70" surface_L="40" pore_radius="2.0"
verbose="true" />
                        <SelectBySASA name="surface_interface" mode="mc"
state="monomer" probe_radius="2.0" core_asa="20" surface_asa="40" core="0" boundary="0"
surface="1" verbose="1" />

                        <RestrictToInterfaceVector name="interface" chain1_num="1"
chain2_num="2" />

                                <ProteinInterfaceDesign name="pack_long"
design_chain1="0" design_chain2="0" jump="1" interface_distance_cutoff="15"/>

                                <ProteinInterfaceDesign name="chain1_interface"
repack_chain1="1" repack_chain2="0" design_chain1="0" design_chain2="0" jump="1"
interface_distance_cutoff="8"/> //
                                ### select interface and allow chain 2 to design
                                <ProteinInterfaceDesign name="chain2_interface"
repack_chain1="0" repack_chain2="1" design_chain1="0" design_chain2="0" jump="1"
interface_distance_cutoff="8"/> //
                                ### this taskop defines the standard interface between two
proteins with jump=1
```

```
                                        <RestrictToInterfaceVector name="restrict_to_interface"
jump="1" CB_dist_cutoff="10.0" nearby_atom_cutoff="5.5" vector_angle_cutoff="75.0"
vector_dist_cutoff="9.0"/>

                                        <LayerDesign name="core_SCN" layer="core"
use_sidechain_neighbors="True" verbose="true" pore_radius="2.0" />
make_pymol_script="true"/>
                                        <LayerDesign name="core_SASA" layer="core"
use_sidechain_neighbors="False" verbose="true" pore_radius="2.0" />
                                        <LayerDesign name="core_boundary"
layer="core_boundary" use_sidechain_neighbors="True" verbose="true" pore_radius="2.0" />

                                        <OperateOnCertainResidues name="no_ala_disulf" >
                                                <ResidueName3Is name3="ALA" />
                                                <RestrictToRepackingRLT />
                                        </OperateOnCertainResidues>

                                        <OperateOnCertainResidues name="no_repack_non-
disulf" >
                                                <ResidueName3Is name3="CYS" />
                                                <RestrictToRepackingRLT />
                                        </OperateOnCertainResidues>

                                        <LayerDesign name="boundary" layer="boundary"
use_sidechain_neighbors="True" verbose="true" pore_radius="2.0" />
make_pymol_script="true"/>
                                        <LayerDesign name="surface"
layer="surface_Nterm_Cterm" use_sidechain_neighbors="True" verbose="true"
pore_radius="2.0" /> make_pymol_script="true"/>


                        </TASKOPERATIONS>

                                <MOVERS>
                                        <TaskAwareMinMover name="min"
scorefxn="nov16_std" bb="0" chi="1" task_operations="pack_long"/>
                                        <SwitchChainOrder name="chainA_only"
chain_order="1"/>
                                        <SwitchChainOrder name="chainB_only"
chain_order="2"/>
                                </MOVERS>

                        <FILTERS>
                                        INTERFACE QUALITY FILTERS
                                        ##############################################
                                        ### ddG w/ and w/o repack/min nov16_std score function
```

```
                                        <Ddg name="ddg"  threshold="-38" jump="1" repeats="5"
repack="1" relax_mover="min" confidence="0" scorefxn="nov16_std" />
                                        <Ddg name="ddg_norepack"  threshold="-10" jump="1"
repeats="1" repack="0" confidence="0" scorefxn="nov16_std"/>

                                        ### ddG w/ and w/o repack VDW score function
                                        <Ddg name="ddg_fa_atr"  threshold="-10" jump="1"
repeats="5" repack="1" relax_mover="min" confidence="0" scorefxn="VDW" />
                                        <Ddg name="ddg_fa_atr_norepack"  threshold="-10"
jump="1" repeats="1" repack="0" confidence="0" scorefxn="VDW"/>

                                        ### determine holes across the interface
                                        <InterfaceHoles name="interface_holes" confidence="0"
jump="1"/>

                                        ### Computes the sasa specifically in the interface: total,
hydrophobic, and polar
                                        <Sasa name="interface_buried_sasa" confidence="0"/>
                                        <Sasa name="interface_hydrophobic_sasa" confidence="0"
hydrophobic="True"/>
                                        <Sasa name="interface_polar_sasa" confidence="0"
polar="True"/>


                                        ### two different types of buried unsatisfied H bond
metrics
                                        <BuriedUnsatHbonds2 name="interface_unsat_hbond2"
confidence="0" jump_number="1"/>

                                        ### total sasa takes all solvent-accessible residues into
account, hydrophobic in this case
                                        <TotalSasa name="exposed_hydrophobics"
confidence="0" hydrophobic="True" />
                                        ### calculate the exposed hydropobic by protein chain
                                        <MoveBeforeFilter name="exposed_np_chain1"
mover="chainA_only" filter="exposed_hydrophobics"/>
                                        <MoveBeforeFilter name="exposed_np_chain2"
mover="chainB_only" filter="exposed_hydrophobics"/>

                                        ### approximation of fraction of non-polar buried/ exposed
residues chain 1; assuming buried nps are equally distributed among both chains
                                        <CalculatorFilter name="fxn_np_chain1_buried_approx"
equation="(sasa  / 2) / (np+0.01)" threshold="1" confidence="0">
                                        <Var name="sasa" filter="interface_hydrophobic_sasa"/>
                                        <Var name="np" filter="exposed_np_chain1"/>
                                        </CalculatorFilter>
```

```
                                        ### approximation of fraction of non-polar buried/ exposed
residues chain 2; assuming buried nps are equally distributed among both chains
                                        <CalculatorFilter name="fxn_np_chain2_buried_approx"
equation="(sasa / 2) / (np+0.01)" threshold="1" confidence="0">
                                        <Var name="sasa" filter="interface_hydrophobic_sasa"/>
                                        <Var name="np" filter="exposed_np_chain2"/>
                                        </CalculatorFilter>

                                        ### difference of total exposed hydrophobics (np) of chain
1 and the hydrophobics of chain 1 at the interface (approximate since assuming that the
hydrophobics at the interface are qually contributed by chain 1 and chain 2)
                                        <CalculatorFilter name="np_chain1_exposed_approx"
equation="np - (sasa / 2)" threshold="1" confidence="0">
                                        <Var name="sasa" filter="interface_hydrophobic_sasa"/>
                                        <Var name="np" filter="exposed_np_chain1"/>
                                        </CalculatorFilter>

                                        ### difference of total exposed hydrophobics (np) of chain
2 and the hydrophobics of chain 1 at the interface (approximate since assuming that the
hydrophobics at the interface are qually contributed by chain 1 and chain 2)
                                        <CalculatorFilter name="np_chain2_exposed_approx"
equation="np - (sasa / 2)" threshold="1" confidence="0">
                                        <Var name="sasa" filter="interface_hydrophobic_sasa"/>
                                        <Var name="np" filter="exposed_np_chain2"/>
                                        </CalculatorFilter>

                                        #### min/repack ddG per 1000 square Angstrom
nov16_std score function
                                        <CalculatorFilter name="ddg_per_1000sasa"
equation="1000 * ddg / (sasa+0.01)" threshold="1" confidence="0">
                                        <Var name="ddg" filter="ddg"/>
                                        <Var name="sasa" filter="interface_buried_sasa"/>
                                        </CalculatorFilter>

                                        #### ddG per 1000 square Angstrom nov16_std score
function
                                        <CalculatorFilter name="ddg_norepack_per_1000sasa"
equation="1000 * ddg / (sasa+0.01)" threshold="1" confidence="0">
                                        <Var name="ddg" filter="ddg_norepack"/>
                                        <Var name="sasa" filter="interface_buried_sasa"/>
                                        </CalculatorFilter>

                                        #### min/repack ddG per 1000 square Angstrom VDW
score function
```

```
                                        <CalculatorFilter name="ddg_fa_atr_per_1000sasa"
equation="1000 * ddg / (sasa+0.01)" threshold="1" confidence="0">
                                        <Var name="ddg" filter="ddg_fa_atr"/>
                                        <Var name="sasa" filter="interface_buried_sasa"/>
                                        </CalculatorFilter>

                                        #### ddG per 1000 square Angstrom VDW score function
                                        <CalculatorFilter
name="ddg_fa_atr_norepack_per_1000sasa" equation="1000 * ddg / (sasa+0.01)" threshold="1"
confidence="0">
                                        <Var name="ddg" filter="ddg_fa_atr_norepack"/>
                                        <Var name="sasa" filter="interface_buried_sasa"/>
                                        </CalculatorFilter>

                                        ### fraction of the interface that is hydrophobic
                                        <CalculatorFilter name="interface_fxn_hydrophobic"
equation="hydrophobic / (sasa + 0.01)" threshold="1" confidence="0">
                                        <Var name="hydrophobic"
filter="interface_hydrophobic_sasa"/>
                                        <Var name="sasa" filter="interface_buried_sasa"/>
                                        </CalculatorFilter>

                                        ### interface average desgree for chain2
                                        <AverageDegree name="interface_averageDegree_chA"
threshold="8.3" task_operations="chain1_interface" confidence="0"/>
                                        <AverageDegree name="interface_averageDegree_chB"
threshold="8.3" task_operations="chain2_interface" confidence="0"/>


                                        ### interface shape complementarity
                                        <ShapeComplementarity name="interface_sc" verbose="0"
min_sc="0.55" write_int_area="1" jump="1" confidence="0"/>
                                        ### normal interface atomic contact count
                                        <AtomicContactCount name="interface_contactcount"
task_operations="restrict_to_interface" confidence="0" />

                                        ### NetChargeFilter
                                        <NetCharge name="NetCharge_binder" min="-100"
max="100" chain="2" />
                                        <NetCharge name="NetCharge_target" min="-100"
max="100" chain="1" />

                                        ### filter to obtain interface residues from tracer
                                        <DesignableResidues name="des_res_chA"
task_operations="restrict_to_interface,restrict_chB" />
```

```
                                        <DesignableResidues name="des_res_chB"
task_operations="restrict_to_interface,restrict_chA" />

                                        MONOMER QUALITY FILTERS

        ###################################################
                                        <AverageDegree name="degree_core_SCN"
task_operations="core_SCN" confidence="0" threshold="9.4" />
                                        <AverageDegree name="degree_core_SASA"
task_operations="core_SASA" confidence="0" threshold="9.4" />
                                        <AverageDegree name="degree" confidence="0"
threshold="9.5" />

                                        <ResidueCount name="res_count_all"
max_residue_count="9999" confidence="0" />
                                        <ResidueCount name="res_count_core_SCN"
task_operations="core_SCN,no_ala_disulf" max_residue_count="9999" confidence="0"/>
                                        <ResidueCount name="res_count_core_SASA"
task_operations="core_SASA,no_ala_disulf" max_residue_count="9999" confidence="0"/>
                                        <ResidueCount name="nres_cys" residue_types="CYS"
confidence="0" />
                                        <ResidueCount name="AlaCount" residue_types="ALA"
max_residue_count="9999" confidence="0"/>

                                        <TotalSasa name="total_sasa" threshold="1"
upper_threshold="1000000000000000" report_per_residue_sasa="False" confidence="0" />

                                        <CalculatorFilter name="mean_sasa" equation="Tsasa /
(rescount2 + 0.01)" threshold="-0.30" confidence="0" >
                                                <Var name="Tsasa" filter="total_sasa"/>
                                                <Var name="rescount2" filter="res_count_all"/>
                                        </CalculatorFilter>

                                        <CalculatorFilter name="percent_core_SCN"
equation="rescount_coreSCN / (rescount3 + 0.01)" threshold="-0.35" confidence="0" >
                                                <Var name="rescount3" filter="res_count_all"/>
                                                <Var name="rescount_coreSCN"
filter="res_count_core_SCN"/>
                                        </CalculatorFilter>

                                        <CalculatorFilter name="percent_core_SASA"
equation="rescount_coreSASA / (rescount4 + 0.01)" threshold="-0.35" confidence="0" >
                                                <Var name="rescount4" filter="res_count_all"/>
                                                <Var name="rescount_coreSASA"
filter="res_count_core_SASA"/>
                                        </CalculatorFilter>
```

```
                                        <AtomicContactCount name="contact_all" confidence="0"
/>
                                        <AtomicContactCount name="contact_core_SCN"
task_operations="core_SCN" confidence="0" />
                                        <AtomicContactCount name="contact_core_SASA"
task_operations="core_SASA" confidence="0" />

                                        <ScoreType name="dslf_fa13" scorefxn="nov16_std"
score_type="dslf_fa13" threshold="0"/>

                                        <CalculatorFilter name="mean_dslf" equation="dslf /
(cyscount+0.01)" threshold="-0.35" confidence="0">
                                                <Var name="dslf" filter="dslf_fa13"/>
                                                <Var name="cyscount" filter="nres_cys"/>
                                        </CalculatorFilter>

                                        <DisulfideEntropy name="entropy" lower_bound="0"
tightness="2" confidence="0"/>
                                        <TaskAwareScoreType name="dslf_quality_check"
task_operations="no_repack_non-disulf" score_type="dslf_fa13"  mode="individual"
threshold="-0.27" confidence="0" />
                                        <ScoreType name="hbond_sfn" scorefxn="nov16_std"
score_type="hbond_lr_bb" threshold="0"/>
                                        <ScoreType name="mono_p_aa_pp"
scorefxn="nov16_std" score_type="p_aa_pp" threshold="0"/>
                                        <ScoreType name="total_score" scorefxn="nov16_std"
threshold="0"/>
                                        <SSShapeComplementarity name="ss_sc" verbose="0"
confidence="0" />
                                        <SSShapeComplementarity name="helix_sc" verbose="1"
loops="0" helices="1" confidence="0" />
                                        <SSShapeComplementarity name="loop_sc" verbose="1"
loops="1" helices="0" confidence="0" />
                                        <TotalSasa name="exposed_total" confidence="0"/>
                                        <TotalSasa name="exposed_polars" confidence="0"
polar="True" />

                                        <CalculatorFilter name="fxn_exposed_is_np"
equation="exposed / total" threshold="1" confidence="0">
                                                <Var name="total" filter="exposed_total"/>
                                                <Var name="exposed"
filter="exposed_hydrophobics"/>
                                        </CalculatorFilter>
```

```
                                        <ScoreType name="total_hydrophobic"
scorefxn="TotalHydrophobic" threshold="0"/>

                                        <CalculatorFilter name="buried_np" equation="total -
exposed" threshold="1" confidence="0">
                                                <Var name="total" filter="total_hydrophobic"/>
                                                <Var name="exposed"
filter="exposed_hydrophobics"/>
                                        </CalculatorFilter>

                                        <CalculatorFilter name="buried_over_exposed"
equation="buried / (exposed+0.01)" threshold="1" confidence="0">
                                                <Var name="buried" filter="buried_np"/>
                                                <Var name="exposed"
filter="exposed_hydrophobics"/>
                                        </CalculatorFilter>

                                        <CalculatorFilter name="buried_minus_exposed"
equation="buried - exposed" threshold="1" confidence="0">
                                                <Var name="buried" filter="buried_np"/>
                                                <Var name="exposed"
filter="exposed_hydrophobics"/>
                                        </CalculatorFilter>

                                        <CalculatorFilter name="bb" equation="hbond /
(rescount+0.01)" threshold="-0.23" confidence="0">
                                                <Var name="hbond" filter="hbond_sfn"/>
                                                <Var name="rescount" filter="res_count_all"/>
                                        </CalculatorFilter>

                                        <CalculatorFilter name="p_aa_pp_perres"
equation="p_aa_pp/rescount" threshold="-0.2" confidence="0">
                                                <Var name="p_aa_pp" filter="mono_p_aa_pp" />
                                                <Var name="rescount" filter="res_count_all"/>
                                        </CalculatorFilter>

                                        <CalculatorFilter name="score_perres"
equation="total/rescount" threshold="-2.8" confidence="0">
                                                <Var name="total" filter="total_score" />
                                                <Var name="rescount" filter="res_count_all"/>
                                        </CalculatorFilter>

                                        <CavityVolume name="cavity_volume" />
                                        <PackStat name="pack" confidence="0" repeats="3" />
                                        <Holes name="holes" confidence="0"/>
```

<SSPrediction name="mismatch_probability" confidence="0" cmd="/work/dadriano/PROGRAMS/psipred/runpsipred_single" use_probability="0" use_svm="0" />

<SecondaryStructureHasResidue name="one_core_each" secstruct_fraction_threshold="1.0" res_check_task_operations="core_SCN" required_restypes="VILMFYW" nres_required_per_secstruct="1" filter_helix="1" filter_sheet="1" filter_loop="0" min_helix_length="4" min_sheet_length="3" min_loop_length="1" confidence="0" />
<SecondaryStructureHasResidue name="two_core_each" secstruct_fraction_threshold="1.0" res_check_task_operations="core_SCN" required_restypes="VILMFYW" nres_required_per_secstruct="2" filter_helix="1" filter_sheet="1" filter_loop="0" min_helix_length="4" min_sheet_length="3" min_loop_length="1" confidence="0" />
<SecondaryStructureHasResidue name="ss_contributes_core" secstruct_fraction_threshold="1.0" res_check_task_operations="core_boundary" required_restypes="VILMFYW" nres_required_per_secstruct="1" filter_helix="1" filter_sheet="1" filter_loop="0" min_helix_length="4" min_sheet_length="3" min_loop_length="1" confidence="0" />
<FragmentLookupFilter name="faulty_fragments_tolerant" lookup_name="source_fragments_4_mer_tolerant" store_path="/lab/databases/VALL_clustered/backbone_profiler_database_06032014" lookup_mode="first" chain="2" threshold="999999" confidence="0" />
<FragmentLookupFilter name="faulty_fragments" lookup_name="source_fragments_4_mer" store_path="/lab/databases/VALL_clustered/backbone_profiler_database_06032014" lookup_mode="first" chain="2" threshold="999999" confidence="0" />

REQUIRED FOR SILLY REASONS
##
<MoveBeforeFilter name="ch1_score_perres" mover="chainA_only" filter="score_perres" confidence="0" />
<MoveBeforeFilter name="ch1_p_aa_pp_perres" mover="chainA_only" filter="p_aa_pp_perres" confidence="0" />
<MoveBeforeFilter name="ch1_res_count_core_SCN" mover="chainA_only" filter="res_count_core_SCN" confidence="0" />
<MoveBeforeFilter name="ch1_res_count_core_SASA" mover="chainA_only" filter="res_count_core_SASA" confidence="0" />
<MoveBeforeFilter name="ch1_percent_core_SCN" mover="chainA_only" filter="percent_core_SCN" confidence="0" />
<MoveBeforeFilter name="ch1_percent_core_SASA" mover="chainA_only" filter="percent_core_SASA" confidence="0" />
<MoveBeforeFilter name="ch1_contact_all" mover="chainA_only" filter="contact_all" confidence="0" />
<MoveBeforeFilter name="ch1_contact_core_SCN" mover="chainA_only" filter="contact_core_SCN" confidence="0" />

```
                                        <MoveBeforeFilter name="ch1_contact_core_SASA"
mover="chainA_only" filter="contact_core_SASA" confidence="0" />
                                        <MoveBeforeFilter name="ch1_degree"
mover="chainA_only" filter="degree" confidence="0" />
                                        <MoveBeforeFilter name="ch1_entropy"
mover="chainA_only" filter="entropy" confidence="0" />
                                        <MoveBeforeFilter name="ch1_dslf_quality_check"
mover="chainA_only" filter="dslf_quality_check" confidence="0" />
                                        <MoveBeforeFilter name="ch1_mean_dslf"
mover="chainA_only" filter="mean_dslf" confidence="0" />
                                        <MoveBeforeFilter name="ch1_cavity_volume"
mover="chainA_only" filter="cavity_volume" confidence="0" />
                                        <MoveBeforeFilter name="ch1_ss_sc"
mover="chainA_only" filter="ss_sc" confidence="0" />
                                        <MoveBeforeFilter name="ch1_helix_sc"
mover="chainA_only" filter="helix_sc" confidence="0" />
                                        <MoveBeforeFilter name="ch1_loop_sc"
mover="chainA_only" filter="loop_sc" confidence="0" />
                                        <MoveBeforeFilter name="ch1_exposed_total"
mover="chainA_only" filter="exposed_total" confidence="0" />
                                        <MoveBeforeFilter name="ch1_exposed_hydrophobics"
mover="chainA_only" filter="exposed_hydrophobics" confidence="0" />
                                        <MoveBeforeFilter name="ch1_exposed_polars"
mover="chainA_only" filter="exposed_polars" confidence="0" />
                                        <MoveBeforeFilter name="ch1_fxn_exposed_is_np"
mover="chainA_only" filter="fxn_exposed_is_np" confidence="0" />
                                        <MoveBeforeFilter name="ch1_holes"
mover="chainA_only" filter="holes" confidence="0" />
                                        <MoveBeforeFilter name="ch1_bb" mover="chainA_only"
filter="bb" confidence="0" />
                                        <MoveBeforeFilter name="ch1_buried_np"
mover="chainA_only" filter="buried_np" confidence="0" />
                                        <MoveBeforeFilter name="ch1_buried_over_exposed"
mover="chainA_only" filter="buried_over_exposed" confidence="0" />
                                        <MoveBeforeFilter name="ch1_buried_minus_exposed"
mover="chainA_only" filter="buried_minus_exposed" confidence="0" />
                                        <MoveBeforeFilter name="ch1_pack"
mover="chainA_only" filter="pack" confidence="0" />
                                        <MoveBeforeFilter name="ch1_mismatch_probability"
mover="chainA_only" filter="mismatch_probability" confidence="0" />
                                        <MoveBeforeFilter name="ch1_one_core_each"
mover="chainA_only" filter="one_core_each" confidence="0" />
                                        <MoveBeforeFilter name="ch1_two_core_each"
mover="chainA_only" filter="two_core_each" confidence="0" />
                                        <MoveBeforeFilter name="ch1_ss_contributes_core"
mover="chainA_only" filter="ss_contributes_core" confidence="0" />
```

```
                                    <MoveBeforeFilter name="ch1_AlaCount"
mover="chainA_only" filter="AlaCount" confidence="0" />
                                    <MoveBeforeFilter name="ch1_faulty_fragments_tolerant"
mover="chainA_only" filter="faulty_fragments_tolerant" confidence="0" />
                                    <MoveBeforeFilter name="ch1_faulty_fragments"
mover="chainA_only" filter="faulty_fragments" confidence="0" />
                                    MoveBeforeFilter name="ch1_worst_ninemer"
mover="chainA_only" filter="worst_ninemer" confidence="0" />

                </FILTERS>

                <MOVERS>
                    <AtomTree name="atree" simple_ft="1" />

                    <Dssp name="dssp" />

                    <FastDesign name="design1" scorefxn="nov16_soft" repeats="3"
task_operations="init,limitchi2,inclcur,core_sheet_res,core_helix_res,core_loop_res,boundary_s
heet_res,boundary_helix_res,boundary_loop_res,surface_sheet_res,surface_helix_res,surface_lo
op_res,repack_graft,repack_chainB_interface,norepack_chainB_noninterface" >
                        <MoveMap name="chainA_only">
                            <Chain number="1" chi="true" bb="false"/>
                            <Chain number="2" chi="true" bb="false"/>
                        </MoveMap>
                    </FastDesign>

                    <FastDesign name="design2" scorefxn="nov16_std" repeats="3"
task_operations="init,limitchi2,inclcur,core_sheet_res,core_helix_res,core_loop_res,boundary_s
heet_res,boundary_helix_res,boundary_loop_res,surface_sheet_res,surface_helix_res,surface_lo
op_res,repack_graft,repack_chainB_interface,norepack_chainB_noninterface,repack_chainA_no
ninterface" >
                        <MoveMap name="chainA_only">
                            <Chain number="1" chi="true" bb="false"/>
                            <Chain number="2" chi="true" bb="false"/>
                        </MoveMap>
                    </FastDesign>

                    <TaskAwareMinMover name="kinematic_min_noBB"
scorefxn="nov16_std" bb="0" chi="1" jump="1" task_operations="init,limitchi2,inclcur,exrot"
/>

                    <TaskAwareMinMover name="chain1_cart_min"
scorefxn="nov16_cart" bb="1" chi="1" jump="1" cartesian="1"
task_operations="init,repack_chainB_interface,norepack_chainB_noninterface,exrot" />
                    <TaskAwareMinMover name="chain1_kinematic_min_noBB"
scorefxn="nov16_std" bb="0" chi="1" jump="1"
```

```
task_operations="init,limitchi2,inclcur,repack_chainB_interface,norepack_chainB_noninterface,
exrot" />

                        <TaskAwareMinMover name="chain1_kinematic_min_wBB"
scorefxn="nov16_std" bb="1" chi="1" jump="1"
task_operations="init,limitchi2,inclcur,repack_chainB_interface,norepack_chainB_noninterface,
exrot"/>

                </MOVERS>

                <PROTOCOLS>

                    <Add mover_name="atree" />
                    <Add mover_name="dssp" />

                    <Add mover_name="design1"/>

                    <Add mover_name="kinematic_min_noBB" />

                    <Add mover_name="design2"/>

                    <Add mover_name="chain1_kinematic_min_noBB"/>
                    <Add mover_name="chain1_kinematic_min_wBB"/>
                    <Add mover_name="chain1_cart_min"/>

                            <Add mover_name="dssp" />

                    <Add filter_name="ch1_score_perres" />
                    <Add filter_name="ch1_p_aa_pp_perres" />
                    <Add filter_name="ddg" />

                            <Add filter_name="des_res_chA" />
                            <Add filter_name="des_res_chB" />

                            <Add filter_name="interface_holes"/>
                            <Add filter_name="interface_unsat_hbond2" />
                            <Add filter_name="interface_buried_sasa" />
                            <Add filter_name="interface_hydrophobic_sasa" />
                            <Add filter_name="interface_polar_sasa" />
                            <Add filter_name="interface_fxn_hydrophobic" />

                            <Add filter_name="ddg_norepack" />
                            <Add filter_name="ddg_per_1000sasa" />
                            <Add filter_name="ddg_norepack_per_1000sasa" />

                            <Add filter_name="ddg_fa_atr" />
```

```
                    <Add filter_name="ddg_fa_atr_norepack" />
                    <Add filter_name="ddg_fa_atr_per_1000sasa" />
                    <Add filter_name="ddg_fa_atr_norepack_per_1000sasa"
/>

                    <Add filter_name="fxn_np_chain1_buried_approx" />
                    <Add filter_name="fxn_np_chain2_buried_approx" />
                    <Add filter_name="np_chain1_exposed_approx" />
                    <Add filter_name="np_chain2_exposed_approx" />

                    <Add filter_name="interface_averageDegree_chA" />
                    <Add filter_name="interface_averageDegree_chB" />
                    <Add filter_name="interface_contactcount" />
                    <Add filter_name="interface_sc" />

                    <Add filter_name="NetCharge_binder" />
                    <Add filter_name="NetCharge_target" />

                    <Add filter_name="ch1_res_count_core_SCN" />
                    <Add filter_name="ch1_res_count_core_SASA" />
                    <Add filter_name="ch1_percent_core_SCN" />
                    <Add filter_name="ch1_percent_core_SASA" />
                    <Add filter_name="ch1_contact_all" />
                    <Add filter_name="ch1_contact_core_SCN" />
                    <Add filter_name="ch1_contact_core_SASA" />
                    <Add filter_name="ch1_degree" />
                    <Add filter_name="ch1_entropy" />
                    <Add filter_name="ch1_dslf_quality_check" />
                    <Add filter_name="ch1_mean_dslf" />
                    <Add filter_name="ch1_cavity_volume" />
                    <Add filter_name="ch1_ss_sc" />
                    <Add filter_name="ch1_helix_sc" />
                    <Add filter_name="ch1_loop_sc" />
                    <Add filter_name="ch1_exposed_total" />
                    <Add filter_name="ch1_exposed_hydrophobics" />
                    <Add filter_name="ch1_exposed_polars" />
                    <Add filter_name="ch1_fxn_exposed_is_np" />
                    <Add filter_name="ch1_holes" />
                    <Add filter_name="ch1_bb" />
                    <Add filter_name="ch1_buried_np" />
                    <Add filter_name="ch1_buried_over_exposed" />
                    <Add filter_name="ch1_buried_minus_exposed" />
                    <Add filter_name="ch1_pack" />
                    <Add filter_name="ch1_mismatch_probability" />
                    <Add filter_name="ch1_one_core_each" />
                    <Add filter_name="ch1_two_core_each" />
```

```
                                        <Add filter_name="ch1_ss_contributes_core" />
                                        <Add filter_name="ch1_AlaCount" />
                                        <Add filter_name="ch1_faulty_fragments_tolerant" />
                                        <Add filter_name="ch1_faulty_fragments" />

                    </PROTOCOLS>

                </ROSETTASCRIPTS>
            </MultiplePoseMover>
        </MOVERS>
        <PROTOCOLS>
            Add mover_name="dumpraw"/>
                <Add mover_name="switch_order"/>
            <Add mover_name="AddVRoot"/>
                <Add mover_name="add_constraints"/>
                    Add mover_name="scaffold_relax"/>
                            Add filter_name="interface_buried_sasa"/>
                    Add mover_name="design1"/>
                <Add mover_name="remove_constraints"/>
                Add mover_name="design1"/>
                Add mover_name="random_perturb"/>
            <Add mover_name="RemVRoot"/>
            Add mover_name="dumprelax"/>
            <Add mover_name="add_disulfs"/>
            <Add mover_name="MPM_design"/>
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

References

1. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302,** 1364–1368 (2003).
2. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357,** 168–175 (2017).
3. Kortemme, T. *et al.* Computational redesign of protein-protein interaction specificity. *Nature Structural & Molecular Biology* **11,** 371–379 (2004).
4. Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550,** 74–79 (2017).
5. Klein, J. C. *et al.* Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Research* (2015). doi:10.1093/nar/gkv1177
6. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat Meth* **7,** 741–746 (2010).
7. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov* **11,** 191–200 (2012).
8. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56,** 152–179 (1992).
9. Lambert, L. C. & Fauci, A. S. Influenza vaccines for the future. *N. Engl. J. Med.* **363,** 2036–2044 (2010).
10. Girard, M. P., Tam, J. S., Assossou, O. M. & Kieny, M. P. The 2009 A (H1N1) influenza virus pandemic: A review. *Vaccine* **28,** 4895–4902 (2010).
11. Jefferson, T. *et al.* Oseltamivir for influenza in adults and children: systematic review of clinical study reports and summary of regulatory comments. *BMJ* **348,** g2545–g2545 (2014).
12. 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).
13. Corti, D. *et al.* A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333,** 850–856 (2011).
14. Fleishman, S. J. *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332,** 816–821 (2011).
15. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol* **30,** 543–548 (2012).
16. Frokjaer, S. & Otzen, D. E. Protein drug stability: a formulation challenge. *Nat Rev Drug Discov* **4,** 298–306 (2005).
17. Ekiert, D. C. *et al.* Antibody recognition of a highly conserved influenza virus epitope. *Science* **324,** 246–251 (2009).
18. Shusta, E. V., Holler, P. D., Kieke, M. C., Kranz, D. M. & Wittrup, K. D. Directed evolution of a stable scaffold for T-cell receptor engineering. *Nat Biotechnol* **18,** 754–759 (2000).
19. Traxlmayr, M. W. *et al.* Directed evolution of stabilized IgG1-Fc scaffolds by application of strong heat shock to libraries displayed on yeast. *Biochim. Biophys. Acta* **1824,** 542–549 (2012).
20. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10,** 866–876 (2009).

21. Araya, C. L. & Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends in Biotechnology* **29,** 435–442 (2011).
22. Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* (2013). doi:10.1038/nature12443
23. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences* **110,** E1263–72 (2013).
24. Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences* **109,** 16858–16863 (2012).
25. Traxlmayr, M. W. *et al.* Construction of a Stability Landscape of the CH3 Domain of Human IgG1 by Combining Directed Evolution with High Throughput Sequencing. *Journal of Molecular Biology* **423,** 397–412 (2012).
26. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci. U.S.A.* **82,** 488–492 (1985).
27. Patrick, W. M., Firth, A. E. & Blackburn, J. M. User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng.* **16,** 451–457 (2003).
28. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* **1,** 755–768 (2006).
29. Hoover, D. M. & Lubkowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Research* **30,** e43 (2002).
30. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Meth* **6,** 343–345 (2009).
31. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification* **41,** 207–234 (2005).
32. Choi, R. *et al.* Immobilized metal-affinity chromatography protein-recovery screening is predictive of crystallographic structure success. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **67,** 998–1005 (2011).
33. Bryan, C. M. *et al.* High-throughput protein production and purification at the Seattle Structural Genomics Center for Infectious Disease. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **67,** 1010–1014 (2011).
34. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
35. Lander, G. C. *et al.* Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166,** 95–102 (2009).
36. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128,** 82–97 (1999).
37. Azoitei, M. L. *et al.* Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science* **334,** 373–376 (2011).
38. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–227 (2012).
39. Lin, Y.-R. *et al.* Control over overall shape and size in de novo designed proteins. *Proceedings of the National Academy of Sciences* **112,** E5478–85 (2015).
40. Havelund, S. *et al.* The Mechanism of Protraction of Insulin Detemir, a Long-

Acting, Acylated Analog of Human Insulin. *Pharm Res* **21,** 1498–1504
41. Berger, S. *et al.* Computationally designed high specificity inhibitors delineate the roles of BCL2 family proteins in cancer. *eLife* **5,** 1422 (2016).
42. Breman, J. G. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *The American Journal of Tropical Medicine and Hygiene* **64,** 1–11 (2001).
43. Ding, X., Liu, D., Booth, G., Gao, W. & Lu, Y. Virus-Like Particle Engineering: From Rational Design to Versatile Applications. *Biotechnol J* **105,** 1700324 (2018).
44. RTS,S Clinical Trials Partnership. Efficacy and safety of RTS, S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually …. *europepmc.org*
doi:10.1016/S0140-6736(15)60721-8
45. Haynes, B. F. & Bradley, T. Broadly Neutralizing Antibodies and the Development of Vaccines. *JAMA* **313,** 2419–2420 (2015).
46. Impagliazzo, A. *et al.* A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. *Science* **349,** 1301–1306 (2015).
47. Pancera, M. *et al.* Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* **514,** 455–461 (2014).
48. Satterthwait, A. C. *et al.* Conformational restriction of peptidyl immunogens with covalent replacements for the hydrogen bond. *Vaccine* **6,** 99–103 (1988).
49. Zavala, F. & Chai, S. Protective anti-sporozoite antibodies induced by a chemically defined synthetic vaccine. *Immunol. Lett.* **25,** 271–274 (1990).
50. Kisalu, N. K. *et al.* A human monoclonal antibody prevents malaria infection by targeting a new site of vulnerability on the parasite. *Nature Medicine* **385,** 336 (2018).
51. Tan, J. *et al.* A public antibody lineage that potently inhibits malaria infection through dual binding to the circumsporozoite protein. *Nature Medicine* **237,** 639 (2018).
52. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS ONE* **6,** e24109 (2011).
53. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27,** 135–145 (2018).
54. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Research* **14,** 1188–1190 (2004).
55. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353,** 389–394 (2016).
56. Nardin, E. H. *et al.* Circumsporozoite proteins of human malaria parasites Plasmodium falciparum and Plasmodium vivax. *J Exp Med* **156,** 20–30 (1982).
57. Satterthwait, A. C. *et al.* The conformational restriction of synthetic vaccines for malaria. *Bull. World Health Organ.* **68 Suppl,** 17–25 (1990).
58. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C.-M. An improved yeast transformation method for the generation of very large human antibody libraries. *peds.oxfordjournals.org*
59. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–620 (2014).

Appendix 1: Rejected haiku dissertation titles

Computer design
But also evolution
My work lacks focus

Computer designed
De novo antibodies
Malaria sucks

Thousands of proteins
How do we know which is best?
Biotechnology

Churning through proteins
Possibilities endless
Difficult problem

Protein therapeutics
How to make and improve them
Almost nothing worked