Constrained Diversification Enhances Protein Ligand Discovery and Evolution

A DISSERTATION
SUBMITTED TO THE FACULY OF THE
UNIVERSITY OF MINNESOTA
BY

Daniel Ray Woldring

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Benjamin J. Hackel

April 2017

Acknowledgements

Even more important than the projects that you pursue or the laboratory techniques that you learn throughout graduate school, the people you work with have the greatest impact on both your success and happiness.

The office. A place for science, debate, and sarcasm, rich with great conversation and supportive colleagues. Brett, Larry, Max, and Sadie, we've become so close over the past few years. I'm eager to see where life takes each of us in the future. Patrick and Aakash, even as undergrads, you were always one of us. It was a pleasure to brainstorm with you while nice tunes played in the background.

In the professional realm, I have been very fortunate regarding the variety of amazing discussions that I've shared with experts in the field such as Amy Keating, Dan Tawfik, and Frances Arnold. I'm incredibly grateful for the career advice that has been bestowed upon me by Tim Whitehead, Brian Kobilka, Nigel Morrison, David Shonnard, and Sarah Green. Thank you all for your insightful perspective and encouraging words.

My great friends in West Michigan have been a cornerstone in my life for as long as I can remember. Chris, Keith, Jared, Joel, Zach, Travis, Charlie – I love you guys. Thanks for putting up with me for so long.

A special thanks to the brothers of Beta Sigma Theta. Michigan Tech would have been considerably less exciting if not for the ridiculous adventures that we had at 1210 College Ave. Jake Edick, in particular, has been a tremendous friend and confidant – for years and years!

Carolyn, my smart and athletic wife. I cannot say enough great things about you. You keep me humble and well fed. Your partnership means the world to me. I would not have survived graduate school without you.

And who can forget FAMML (Frank, Andrew, Mark, Megan, Lauren). Such a loyal group of charming people. You have always given me a great excuse to take a break from work.

I would, of course, like to thank my parents and brother (Josh) for providing me with a solid foundation from which to grow. I'm proud to have been raised in a Christian household with conservative values. I intend to keep these lessons with me throughout my life.

Finally, there is Ben Hackel. I have never met a man more hard working, intelligent, honest, and caring – nor have I ever held such sincere respect for an individual. I am so thankful that he was willing to train me as a scientist over the past several years. The ethical and safety standards that he instilled in the culture of our lab will have a lasting effect. Thank you for the intellectually stimulating conversations. Thank you for never stifling my creativity. Thank you for demanding that we think critically about every experiment we conduct and every paper we read. Thank you for keeping me on the right path through discouraging times. Thank you for your unwavering support.

Proudly dedicated to my grandparents, Don Woldring, Sandy Nowak, and Ken and Diane Hitchcock – whose faith and actions are a constant source of inspiration.

Abstract

Engineered proteins have strongly benefited the effectiveness and variety of precision drugs, molecular diagnostic agents, and fundamental research reagents. A growing demand for new therapeutics motivates the innovative use of natural proteins – improving upon their native properties – as well as discovering proteins with entirely new functionality. Importantly, these are fundamentally separate goals. While evolving improved function can result from making a few carefully chosen mutations, discovering novel function often requires giant leaps to be taken in protein sequence space. Discovering novel function is a notoriously challenging task. The immensity of sequence space (e.g. proteins of length *n* have 20^n unique options) makes it essentially impossible to experimentally or computationally test all possible protein sequences. Within this space, the landscape is incredibly barren and rugged (i.e. most sequences lack function entirely and making small changes to a protein often damage the activity). Rather than randomly mutating a protein, combinatorial protein libraries provide a systematic and efficient approach for searching sequence space. This method offers precise control over which protein sites are mutated and which amino acids are allowed at the diversified sites. To improve the likelihood of sampling useful sequences, numerous techniques can elucidate the structure-function relationships in proteins. Generally, these techniques have not been applied to combinatorial library design; however, we propose that some, or all, could be greatly beneficial in this area. In this thesis work, protein libraries are designed for the purpose of discovering high affinity, specific binders to a collection of interesting targets. High-throughput sequencing of evolved binders, natural protein-protein interface

composition, structural assessment, and computational analysis of stability upon mutation collectively informed sitewise library designs – residues predicted to support function were allowed but destabilizing residues or those not likely to benefit function were avoided. We use multiple small protein scaffolds (affibody and fibronectin) as model systems to test the hypothesis that constrained sitewise diversity will improve the efficiency of novel protein discovery. This hypothesis was experimentally supported by a direct comparison of high-affinity ligand discovery between the sitewise constrained library and a uniformly diversified library (i.e. allowing all 20 residues at each diversified site). The constrained library showed a 13-fold improved likelihood of binder discovery. Moreover, the constrained library variants demonstrated superior thermal stability (T_m 15 °C higher than unbiased variants). This work provides further evidence that sitewise diversification of protein scaffolds can improve the overall quality of combinatorial libraries by offering broad coverage of sequence space without sacrificing stability.

Table of Contents

List of Tables

List of Figures

Chapter 1 – Introduction

1.1. Discovering and evolving novel protein function is important for clinical utility, fundamental research, and industrial applications.

The proper administration of effective medicines can broadly improve the efficiency of our health care system as well as improve the quality of life for individual patients battling cancer, diabetes, immune diseases, and more. Molecularly 'targeted' therapies and diagnostics are playing an increasingly beneficial role in this endeavor.[1,2] There are >100 protein-based molecular therapeutics approved by the United States Food and Drug Administration including monoclonal antibodies, insulin, human growth hormone, and granulocyte-colony stimulating factor.[2–4] These molecular medicines enable precision therapy to modulate disease biochemistry in a focused manner.[5] Similarly, molecular diagnostics empower scientists and clinicians to detect biomarkers and problematic cell types with high spatiotemporal resolution. Molecular imaging with engineered proteins enables detection and quantitative characterization of disease via molecular ultrasound, positron emission tomography (PET), and single-photon-emission computed tomography (SPECT).[6] Even individual cells can be studied using intravital microscopy techniques.[7,8] Outside of the clinic, a nearly endless variety of engineered proteins can be found. Industrial and research scientists rely on the binding and enzymatic functionality found within core techniques such as immunoaffinity bioseparation, Western blots, and flow cytometry.[9]

1.2. The path to novel protein function is wrought with challenges and pitfalls.

Nature provides many proteins with an array of functionality (e.g. immune system antibodies, photosynthetic enzymes, etc.). However, the scientific community has repeatedly found that natural proteins rarely exist with optimal functionality and biophysical characteristics (thermal stability, chemical stability, solubility, etc.), leaving room for improvement.[10–12] Moreover, to go beyond what nature offers, new applications could be established by taking an existing natural protein, then modifying the protein such that it takes on functionality distinctly different than its native role. Alternatively, one can imagine designing or developing proteins that are not found in nature to achieve new functions. Scenarios such as these demonstrate the utility of either improving protein function or discovering entirely new function. Importantly, the process of evolving improved function with an existing protein is fundamentally separate from discovering entirely new function. While evolving improved function can result from making a few strategically chosen mutations, discovering novel function typically requires huge leaps to be taken in protein sequence space. Thus, protein engineering is a worthy goal, yet a challenging goal because of the complexity of protein interactions both inter- and intra-molecularly.

Such complexity is echoed by our tenuous grasp of protein sequence-function relationships. Unfortunately, sequence space (i.e. all possible polypeptides of the 20 natural amino acids in a protein of length n, which is 20^n sequences) is far too large to either experimentally or computationally test each unique amino acid sequence. The inaccuracy by which we predict mutations that will render a functional protein and which may improve biophysical characteristics is not a minor inconvenience.[13–16] The majority of sequence

space lacks function entirely, making it exceedingly unlikely to find a useful protein simply by chance.[17] Also, even if you have a stable and functional starting point, an average mutation imposed on it will both destabilize and reduce the functional activity of the protein.[18–21] Owning to nature's preferential selection of superior function rather than thermal stability, most natural proteins are only marginally stable ($\Delta G_{Folding}$ = -5 – -10 kcal/mol) to begin with.[22–25] However, proteins, with the exception of intrinsically disordered proteins, require sufficient stability to fold properly and carry out a particular function.[26] Thus, unwisely chosen mutations can yield variants that are unable to fold, thus preventing any potential functionality from being realized. *De novo* engineering of protein function, therefore, requires careful balancing of the requisite high number of mutations and their propensity for detriment.

1.3. Protein evolution benefits from a more stable starting point as well as less destabilized variants

Indeed, the likelihood of destabilizing mutations portrays naïve protein design as an arduous and inefficient task to the point of ineffectiveness. Discovering distinctly unique function from some protein starting point requires changing amino acids at several sites[27], yet identifying a suitable combination of mutations is hindered by destabilization. To increase the fraction of stable variants that result from mutation, multiple approaches should be considered. One field of thought asserts that proteins offering a more thermodynamically stable fold are able to accept a higher fraction of random mutations while still maintaining their native fold.[28,29] Bloom, et al. leveraged this approach to engineer new enzymatic functionality in P450. In their study, novel function was achieved through several functionally beneficial, but destabilizing, mutations. The critical mutations

were supported within a hyperstable P450 variant, but not the less stable parent.[30] In this thesis, we pursue a related, but distinct approach in which strongly destabilizing mutations are avoided in favor of those that tend to be less destabilizing. Even in the absence of a hyperstable starting point, by improving the quality of mutations, an increased fraction of stable variants will follow.

1.4. Navigating sequence space benefits from efficient search methods

In the ideal case of protein engineering, it would be possible to precisely predict which positions within a protein should be modified to particular amino acids in order to elicit the function or biophysical property of interest. This type of rational design requires very low throughput and is well suited for scenarios where there exists rich structural and phylogenetic datasets related to both the protein and function of interest.[31] This is not a common situation. When limited data exists for a target protein or function, it is likely to require screening a much greater variety of mutations to identify the desired properties. The vast and rugged nature of sequence space motivates that we probe the functional landscape (i.e. search for new function in the context of a given protein) in a systematic and efficient manner. This can be accomplished using a wide variety of techniques employed by the protein engineering community using protein libraries.[32]

Protein libraries are composed of numerous unique members (variants) based on a common starting point (parent). The composition of the library (i.e. which sites are diversified and which amino acids are allowed at diverse sites) will significantly impact the ability to discover or evolve protein function. When very little is known about a parent protein, a scouting library can be generated that consists of point mutations scattered randomly throughout the entire protein. This approach, error-prone PCR[33], has the benefit

of broadly searching while allowing for synergistic or epistatic mutations between distant sites; however, only shallow diversity is reached at any particular site and the likelihood of a specific site receiving a point mutation is largely stochastic. Alternatively, synthesized combinatorial libraries allow for much greater control over which sites are diversified and which amino acids are implemented at each site. One approach involves fully sampling all possible amino acids at a select number of protein sites (saturation mutagenesis).[34] This results in a much greater depth of information, but at fewer sites and fails to broadly inform epistatic relationships. Consensus designs take advantage of available phylogenetic or homology information, implementing amino acid diversity that mimics the most frequently observed natural mutations.[35] This approach is primarily limited by the quality and extent of phylogenetic data available and is restricted to functions previously observed in nature.

Building combinatorial protein libraries is straightforward. The amino acid sequences are designed on the DNA level. Standard codons are used for specifying particular amino acids at each position, while degenerate codons allow for combinations of amino acids to be incorporated at a position of choice. Synthetically designed DNA oligos encoding for the protein libraries are amplified using PCR, then shuttled into a high-throughput display system of choice – bacterial, mRNA, phage, ribosome, yeast, etc. In vitro and cell surface display systems provide a link between individual protein variants and the DNA which encodes for the variant. Thus, library member proteins can be easily identified by their DNA sequence after undergoing high-throughput selection methods such as magnetic bead sorting, fluorescence activated cell sorting, or cell panning.[36,37]

Display technologies, while limited by transformation efficiency (bacterial, phage, and yeast) or volume throughput (mRNA and ribosome), allow for libraries that contain

10^9 – 10^{15} unique members.[38,39] These library sizes (i.e. the actual number of unique members being experimentally tested) are significantly smaller than that of sequence space, even for small peptides.

1.5. Experimental data and computational models inform protein library design.

This disparity in searchable sequence space motivates protein libraries to incorporate only the most important variants, those that have the highest likelihood of demonstrating the function or biophysical parameter of interest. Efficient combinatorial libraries should avoid amino acid diversity that evokes extensive destabilization or detracts from functionality. Importantly, numerous techniques can provide insight into sequence-structure-function relationships in proteins. While these methods have generally not been applied to combinatorial library design, we propose that some, or all, could provide utility in this area. Consensus design with natural protein homologs uses the identification of conserved sites to predict stabilizing mutations. Structural data inform computer simulations (e.g. Rosetta[40]) modelling the interface of protein binding events to predict mutations that promote strong affinity interactions.[41] Estimating changes in stability upon mutation have also been successful.[42–45] Even in the absence of computational tools, structural information can be probed to inform library design of a specific protein[46] as well as highlighting the natural prevalence of amino acids within different secondary structures (e.g. glycine is common within loops, but is underrepresented in helices).

Naturally observed sequences can guide protein library design. The amino acids most often observed in protein-protein interfaces suggest the most beneficial sidechains that facilitate binding events. For instance, tyrosine is particularly impactful at driving binding affinity and specificity, whereas serine can be a neutral interactor to offer an option

when tyrosine is not optimal for the particular site.[47–51] Glycine contributes to conformational flexibility.[52–54] Arginine content drives non-specific interactions at binding interfaces.[54,55] Thus, for designing protein libraries where binding affinity is the property of interest, these amino acid preferences should be considered.[56]

Beyond the prevalence of amino acids generally found at binding interfaces, nature can also provide insight to the arrangement of amino acid diversity from site-to-site. Analysis of the natural antibody repertoire shows that amino acid diversity varies at each site. Even within the complementarity determining regions (CDRs) of antibodies, amino acid composition is not uniformly distributed.[57] Allowing variable levels of amino acid diversity has been successfully used in the context of natural[58] and synthetic[53] antibody repertoires as well as, to a limited extent, designed ankyrin repeat proteins[46,59,60] and synthetic fibronectin domain libraries[51,61]. In the body of work presented here, high-throughput evolution, deep sequencing, and sitewise diversification are used to guide efficient library design.[61,62]

Chapter 2 – High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains

Adapted from "Daniel R. Woldring, Patrick V. Holec, Hong Zhou, and Benjamin J. Hackel. 'High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains.' *PLoS One* **2015**, *10* (9), e0138956."

2.1. Synopsis

Discovering new binding function via a combinatorial library in small protein scaffolds requires balance between appropriate mutations to introduce favorable intermolecular interactions while maintaining intramolecular integrity. Sitewise constraints exist in a non-spatial gradient from diverse to conserved in evolved antibody repertoires; yet non-antibody scaffolds generally do not implement this strategy in combinatorial libraries. Despite the fact that biased amino acid distributions, typically elevated in tyrosine, serine, and glycine, have gained wider use in synthetic scaffolds, these distributions are still predominantly applied uniformly to diversified sites. While select sites in fibronectin domains and DARPins have shown benefit from sitewise designs, they have not been deeply evaluated. Inspired by this disparity between diversity distributions in natural libraries and synthetic scaffold libraries, we hypothesized that binders resulting from discovery and evolution would exhibit a non-spatial, sitewise gradient of amino acid diversity. To identify sitewise diversities consistent with efficient evolution in the context of a hydrophilic fibronectin domain, $>10^5$ binders to six targets were evolved and sequenced. Evolutionarily favorable amino acid distributions at 25 sites reveal Shannon entropies (range: 0.3-3.9; median: 2.1; standard deviation: 1.1) supporting the diversity

gradient hypothesis. Sitewise constraints in evolved sequences are consistent with complementarity, stability, and consensus biases. Implementation of sitewise constrained diversity enables direct selection of nanomolar affinity binders validating an efficient strategy to balance inter- and intra-molecular interaction demands at each site.

2.2. Introduction

Protein sequence space is immense, and the protein/function landscape is rugged and barren: very similar sequences often have greatly different function with the majority of sequences lacking any utility[17]. Protein complexity and our naivety of sequence/structure/function interplay[63] hinder robust de novo design, although several designs have been successfully realized[64–66]. Thus, naïve identification of protein sequences with novel functions, or even mutants with improved function, benefits from combinatorial analysis of many proteins. The efficacy of this approach is directly dependent on combinatorial library quality and the phenotype selection process. The essence of discovery and evolutionary efficiency is to intelligently search sequence space by identifying the effective extent and amino acid distribution of diversity (if any) at each site. Protein discovery and evolution must balance [67] variance sufficient for generation of novel function (dominated by intermolecular interactions) versus conservation sufficient to maintain a high probability of foldable stability (intramolecular interactions)[21]. This challenge is heightened in small domains[68] that have limited area for a binder interface and require mutation of a larger fraction of the molecule[69].

Antibody repertoires have evolved sitewise amino acid distributions across a range of diversities (Fig. 1A), which are used in natural and synthetic antibody libraries[70–74]. Yet most synthetic scaffold libraries – including affibodies[75], affitins[76,77], knottins[78,79],

anticalins[80], Fynomers[81], Sso7d[82], and OBodies[83] – are binary with a fully conserved framework and uniformly diversified paratope (Fig. 1B). Note that different scaffolds use different uniform distributions including NNK codons[84] or complementarity biases[47,55] but generally lack sitewise variation. DARPin domain libraries have six sites with a uniform broad distribution and one site with N/H/Y diversity[59]. A hydrophilic DARPin library includes two additional sitewise diversities[46]. The most sitewise design in non-antibody scaffolds has been introduced in the type III fibronectin domain. Diversification of one, two, or three loops,[85,86] or the sheet surface,[87,88] of this 10 kDa beta sandwich has enabled evolution of binding to a host of molecular targets [86]. Antibody-inspired amino acid bias in putative hot spots has proven effective within fibronectin libraries[87,89–91]. Diversification of two loops is evolutionarily superior to one-loop mutation[92], and although diversification of the third loop (DE loop: G52-T56) is not requisite for high-affinity binding, [86,92–95] it can aid stability [93]. Current library designs randomize G52 with G/S/Y, S53-S55 with Y/S, and 12-22 sites in two other loops with a consistent distribution (30% Y, 15% S, 10% G, 5% each F and W, and 2.5% others except C). Sitewise design was extended beyond the DE loop using accessibility, stability, and homology data yielding nine different diversifications at 11 sites in addition to 12 sites with consistent complementarity-biased diversity [91]. Also, in an alternative paratope approach to engineering fibronectin domains, five sites were identified for three different varieties of constrained diversities (4-8 amino acids) in addition to 12-19 sites with the complementarity biased diversity [87]. While a variety of sitewise diversities have been implemented in the fibronectin domain, the evolved repertoires resulting from these libraries have not been broadly and deeply analyzed.

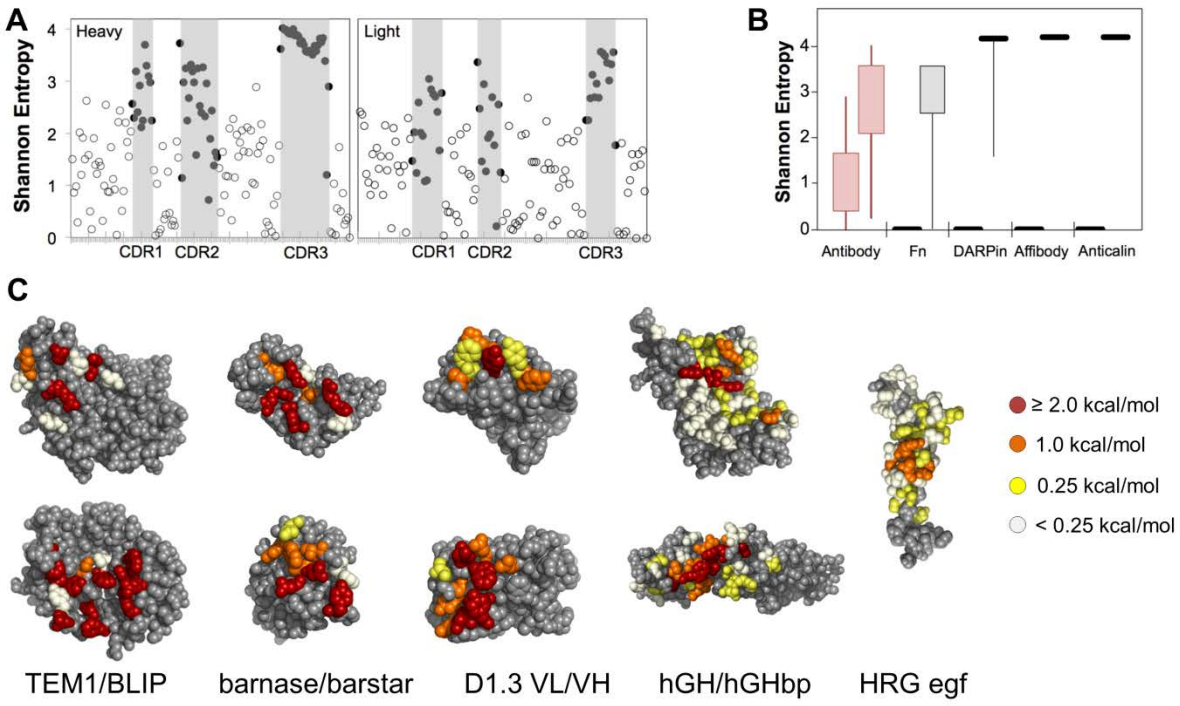


Figure 2-1. ***Diversity gradients in binding molecules*****. (A) The amino acid diversity, measured as Shannon entropy of antibody sequences (from the abysis database at http://www.bioinf.org.uk/abysis/) in the framework (open circles, white background) and CDRs (solid circles, gray background). (B) The Shannon entropy of combinatorial library designs for antibody (data from *A*), fibronectin (Fn) [96], designed ankyrin repeat (DARPin) [97], affibody [98], and anticalin [99], domains. (C) The relative impact of alanine mutation on binding is shown for several protein interfaces: TEM1-β–lactamase (TEM1) and β-lactamase inhibitor protein (BLIP) (PDB: 1JTG) [100]; extracellular RNase (barnase) and its intracellular inhibitory binding partner (barstar) (PDB: 1BRS) [101]; light and heavy chain variable regions of anti-hen egg white lysozyme antibody D1.3 in the context of binding anti-D1.3 antibody E5.2 (not shown) (PDB: 1DVF) [102]; human growth hormone (hGH) and extracellular domain binding partner (hGHpb) (PDB: 1A22) [103,104]; heregulinβ (HRG) egf domain in the context of binding ErbB3 receptor-IgG fusion (not shown) (PDB: 1HAE) [105].**

The current study aims to quantitatively evaluate the broad extents of diversification and sitewise amino acid distributions that evolve in hydrophilic fibronectin domains (Fn3HP) developed as binding ligands. The Fn3HP mutant was previously evolved for hydrophilicity to improve processing and in vivo biodistribution [106]. We posit that a broad repertoire evolved from combinatorial libraries for de novo discovery will exhibit sitewise complementarity-biased amino acid diversity in the binding hot spot

[104,107,108], conserved wild-type sequence in the distal framework, and a gradient of diversification at intermediate sites including bias for conservation or interactive neutrality in proximal regions. This gradient is not purely spatial as protein structure and protein-protein interfaces are complex (Fig. 1C). Moreover, for novel ligand discovery, the exact paratope is not known ahead of time, which blurs designed localization of a hot spot [107].

The approach used was high-throughput discovery and directed evolution of thousands of binding ligands to various targets from a diverse combinatorial library followed by thorough sequencing of the library and binder populations to identify diversities and amino acids consistent with functional hydrophilic fibronectin domains (S1 Fig.). Deep sequencing of evolved protein populations has proven effective for analysis of functionality landscapes for maturation of single protein clones [109], protein families [110,111], and antibody [112] repertoires. Here we apply deep sequencing to several high-throughput discovery and evolutionary campaigns to identify repertoires of evolved hydrophilic fibronectin domains. The results demonstrate a range of diversities and sitewise amino acid preferences consistent with a benefit of a gradient of sitewise constraint. A constrained library based on the observed evolutionary repertoire provides stable, high affinity binders directly without maturation, and the sequence analysis provides a metric to evaluate the balance of inter- and intra-molecular considerations in library design, which are quantitatively assessed.

2.3. Materials and Methods

2.3.1 Library Construction

Oligonucleotides, including amino acid diversity and loop length variation, were synthesized by IDT DNA Technologies (Tables 1-2 and S1-S2 Tables). Full-length Fn3HP amplicons were assembled by overlap extension PCR. The library of pooled diversified DNA was homologously recombined into a pCT yeast surface display vector [92] within yeast strain EBY100 [113] during electroporation transformation. The protocol was similar to that described by Benatuil et al. [114]. Yeast at an optical density at 600 nM of 1.3-1.5 were washed twice with cold water and once with buffer E (1 M sorbitol, 1 mM $CaCl_2$) and resuspended in 0.1 M lithium acetate, 10 mM Tris, 1 mM ethylenediaminetetraacetic acid, pH 7.5. Fresh dithiothreitol was added to 10 mM. Cells were incubated at 30º, 250 rpm for 30 minutes. Cells were washed thrice with cold buffer E and resuspended to 1.4 billion cells per 0.3 mL buffer E. Six µg of linearized pCT vector and 200 pmol of ethanol precipitated gene insert were added and transferred to a 2 mm cuvette. Cells were electroporated at 1.2 kV and 25 µF, diluted in YPD (10 g/L yeast extract, 20 g/L peptone, 20 g/L dextrose), and incubated at 30º for 1 hour. Cells were pelleted and resuspended in 100 mL SD-CAA (16.8 g/L sodium citrate dihydrate, 3.9 g/L citric acid, 20.0 g/L dextrose, 6.7 g/L yeast nitrogen base, 5.0 g/L casamino acids). Plasmid-containing yeast were quantified by dilution plating on SD-CAA agar plates.

Table 2-1. Amino acid diversity encoded in first generation library. Each row was constructed as a separate sublibrary. CDR' refers to a degenerate codon with the following nucleotide frequencies: 20% A, 15% C, 25% G, and 40% T at site 1, 50% A, 25% C, 15% G, and 10% T at site 2, and 0% A, 45% C, 10% G, and 45% T at site 3. Loop length diversity was afforded at mid-loop positions by inserting CDR' diversity sites between sites P25-V29 and G79-S85 of the BC and FG loop, respectively, as denoted by subscripts within table. Length diversity at the DE loop occurred between S53-S55, consisting of diversity matching that of K54 as shown below. Note that throughout the manuscript, a series of unseparated capital amino acid abbreviations refer to equally possible amino acids; for example, SYNT indicates 25% each of serine, tyrosine, asparagine, and threonine at that site.

	BC Loop							DE Loop					FG Loop					
Site	D23	A24	P25	AVT 26-28	V29	R30	Y31	G52	S53	K54	S55	T56	T76	GR 77-78	G79	DSPAS 80-84	S85	K86
Sublib. 1	D	A	P	CDR'$_{1-4}$	AST	CDR'	G	G	S	N_{0-2}	S	T	T	CDR'	GDNSYC	CDR'$_{1-5}$	S	N
Sublib. 2	DSYA	AS	PS	CDR'$_{1-4}$	AST	CDR'	GS	GS	S	$(NS)_{0-2}$	S	TS	TS	CDR'	GDNSYC	CDR'$_{1-5}$	S	NS
Sublib. 3	DSYA	ASYD	PSYH	CDR'$_{1-4}$	AST	CDR'	GSYCDN	GSYCDN	SYNT	$(NSYT)_{0-2}$	SYNT	TSYN	TSYN	CDR'	GDNSYC	CDR'$_{1-5}$	SYNT	NSYT
Sublib. 4	ACDGNSTY	ACDGNSTY	ACDGNSTY	CDR'$_{1-4}$	AST	CDR'	ACDGNSTY	ACDGNSTY	ACDGNSTY	$(ACDGNSTY)_{0-2}$	ACDGNSTY	ACDGNSTY	ACDGNSTY	CDR'	GDNSYC	CDR'$_{1-5}$	ACDGNSTY	ACDGNSTY
Sublib. 5	CDR'	CDR'	CDR'	CDR'$_{1-4}$	AST	CDR'	CDR'	CDR'	CDR'	CDR'$_{0-2}$	CDR'	CDR'	CDR'	CDR'	GDNSYC	CDR'$_{1-5}$	CDR'	CDR'

Table 2-2. Amino acid diversity encoded in second generation library. CDR' refers to a degenerate codon with the following nucleotide frequencies : 20% A, 15% C, 25% G, and 40% T at site 1, 50% A, 25% C, 15% G, and 10% T at site 2, and 0% A, 45% C, 10% G, and 45% T at site 3. Loop length diversity was afforded at mid-loop positions by inserting CDR' diversity sites between sites P25-V29 and G79-S85 of the BC and FG loop, respectively, as denoted by subscripts within table. Length diversity at the DE loop occurred between S53-S55, consisting of either wild-type length or the exclusion of K54.

	BC Loop							**DE Loop**					**FG Loop**						
Site	**D23**	**A24**	**P25**	**AVT 26-28**	**V29**	**R30**	**Y31**	**G52**	**S53**	**K54**	**S55**	**T56**	**T76**	**G77**	**R78**	**G79**	**DSPAS 80-84**	**S85**	**K86**
Generation 2	**D**	**A/ASYDNT**	**P/PSYH**	**CDR'$_{2-4}$**	**AST**	**CDR'**	**GY**	**G**	**SYNT**	**$(NSYT)_{0-1}$**	**SYNT**	**TSYN**	**TSGA**	**GSYADCNT**	**CDR'**	**GSDN**	**CDR'$_{1-5}$**	**S**	**N**

Each resulting Fn3HP naïve yeast library was evaluated for proper library construction by Sanger sequencing clonal plasmids harvested from the transformed yeast (57 clones from generation one and 15 from generation two naïve libraries) and later via Illumina sequencing. The yeast libraries were also labeled with biotinylated anti-HA antibody (goat pAb, Genscript) and anti-c-MYC antibody (9E10, Covance Antibody Products;) to detect the presence of N- and C-terminal epitopes present on either side of the Fn3HP clones, respectively, via flow cytometry. The fractional detection of cells displaying both HA and c-MYC, compared to those displaying HA alone, is indicative of full-length, stop codon-free clones.

2.3.2 Binder Maturation and Evolution

The Fn3HP yeast library was grown in SD-CAA selection media for several doublings (about 20 h) in an incubator shaker at 30°C until an optical density value of 6.0 was reached, at which time the yeast were centrifuged and resuspended in SG-CAA induction media (10.2 g/L sodium phosphate dibasic heptahydrate, 8.6 g/L sodium phosphate monobasic monohydrate, 19.0 g/L galactose, 1.0 g/L dextrose, 6.7 g/L yeast nitrogen base, 5.0 g/L casamino acids) and grown overnight. The induced library was sorted twice via multivalent magnetic bead selections [115] via depletion of non-specific binders on avidin-coated beads and control protein-coated beads followed by enrichment of specific binders on target-coated beads. The pair of magnetic sorts was followed by a flow cytometry selection for full-length clones using the 9E10 antibody against the C-terminal c-MYC epitope tag. Genes were mutated via error-prone PCR with loop shuffling [93], then electroporated into yeast (EBY100) as previously described. Target binding

populations were isolated at two levels of stringency, mid- and high-affinity, for each of the four campaigns. A mid-affinity population included all clones that demonstrated either (i) magnetic bead sorting enrichment at least ten-fold greater for target protein than both avidin binding and non-specific control binding or (ii) binding to 50 nM multivalent target (3:1 stoichiometry of target preloaded on streptavidin-fluorophore) assayed via flow cytometry (S2 Fig.). A high-affinity population included all clones exhibiting binding to 50 nM monovalent target assayed via flow cytometry. Herein, clones meeting these criteria are referred to as mid- and high- affinity binders, respectively. Flow cytometry was performed as previously described [116].

2.3.3 Illumina MiSeq Sample Preparation and Sequence Analysis

Plasmid DNA was isolated from yeast using Zymoprep Yeast Plasmid Miniprep II. DNA samples were divided into separate groups based on library generation of origin and binding affinity. Three categories were included for each generation: naïve clones from the initial libraries, mid-affinity binders collected via magnetic bead sorting, and high-affinity binders collected using FACS. In total, six pools of DNA were isolated and uniquely analyzed in association with generations one and two. Following plasmid DNA extraction, two rounds of PCR were completed to assemble the Fn3HP gene fragment with Illumina primers, index tags, multiplexing bar codes, and TruSeq universal adapter (S4 Table). For all PCR conducted during amplicon library preparation, KAPA HiFi polymerase was used as it has been shown to reduce clonal amplification bias due to GC content [117] as well as fragment length bias [118]. Compatible multiplexing and adapter primers were designed according to TruSeq sample preparation guidelines. Amplicons were pooled and

supplemented with 25% PhiX control library to increase MiSeq read accuracy. Illumina MiSeq paired-end sequencing with 2 x 250 read length was conducted (University of Minnesota Genomics Center) to obtain $7.2x10^6$ pass filter (PF) reads from the populations of interest, of which 90% of all pass filter bases were above Q30 quality metric (99.9% read accuracy).

2.3.4 Sequence Analysis

Raw data generated through MiSeq consisted of forward and reverse read files (FASTQ) for each of the six multiplexed sublibraries. Assembly of paired end reads was done using PANDAseq [119]. Assembled reads were analyzed using in-house Python [120] code. Analysis work flow for each of the six subgroups (e.g. naïve, mid-affinity, high-affinity populations originating from first and second generation libraries) consisted of first identifying full-length fibronectin DNA sequences, isolating each of the three diversified loop regions, and, lastly, calculating the amino acid frequency at each site. Additional calculations were necessary for the mid-affinity and high-affinity populations to both remove statistically rare events and avoid overcounting dominant clones. The removal of background (i.e. the rarest 2% of sequences, as determined by the rarity of nonspecific binders) was a precaution taken when analyzing the mid-affinity populations to account for the rare non-binding clones inherently collected via magnetic bead sorting [115]. To address the potential detriment of overcounting within all binding populations, the sequences for each loop region were clustered based on 80% or greater sequence homology. For each cluster of similar sequences, the summation of the amino acids at each site were weighted by a power of one-half, then aggregated across all clusters. The resulting weighted sitewise

amino acid values were used for frequency calculations. Statistical analysis was performed using two sample Student's t-test. Statistical significance was assessed while adjusting for familywise error rate using Bonferroni method, denoted at level $\alpha = 0.005$.

2.3.5 Stability Assessment

High-affinity clones from three separate target binding campaigns of the current study were individually evaluated for stability using thermal denaturation midpoint, T_m, in the context of yeast surface display, as previously described [93]. Wild-type Fn3HP and seven random clones from the second generation initial library were produced with a C-terminal six-histidine tag in BL21(DE3) and purified by immobilized metal affinity chromatography and reverse phase high performance liquid chromatography. Purified proteins (1 mg/mL in 2 mM 4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid, 50 mM NaCl, 2 mM ethylenediaminetetraacetic acid, 1 mM dithiothreitol) were analyzed via circular dichroism [121] using a JASCO J815 instrument. Measurements of molar ellipticity were taken at 218 nm while heating from 20-98°C at a rate of 1 °C per minute. Stability measurements of 15 engineered fibronectin clones were retrieved from previously published studies wherein library design was implemented through a binary approach: broadly diversifying the anticipated paratope, using NNS [122] and NNB [93] codons, and fully conserving all other positions.

2.3.6 Correlative Parametric Analysis of Amino Acid Distributions

To evaluate the correlation of evolved sitewise amino acid frequencies with several computed parameter matrices (sitewise computational stability, sitewise amino acid

frequency in natural homologs, complementarity, and exposure), a sitewise amino acid frequency matrix (F) is calculated from Equation 1:

$$F = \sum_{k}^{a,b,c}(\alpha_k + \beta_k \cdot \varepsilon) \cdot f_k \quad \text{(Eq. 1)}$$

where α_k and β_k are tunable weights to scale the primary parameter data (fk) as a function of exposure score, ε (see below). Parameter weights that are most consistent with experimental data were calculated using a least-square method to minimize error between the calculated matrix, F, and objective matrix, defined as the sitewise amino acid frequencies observed in binder sequences evolved from the second generation library. Constraints are placed such that each set of α values sum to 1.0 and each set of β values sum to zero.

2.3.7 Amino Acid Frequencies in Natural Homologs

The Pfam database offers an extensive collection of protein families compiled through hidden Markov models [123]. The Fn3 protein family homologs (PF00041) were aligned to the 101 amino acids in Fn3HP. To reduce the impact of dominant replicate sequences, while still accounting for their repeated observation, the frequency of replicate sequences were counted as the square root of the total number of occurrences. The amino acid frequency at each site was computed (S3B Table).

2.3.8 Stability Matrices

FoldX [124] was used to determine the mutability of sites 23-31, 52-56, and 76-86 within the tenth type III domain of human fibronectin in the context of several structures cataloged in the Protein Data Bank [125] (PDBs: 1FNA, 1TTG, 2OBG, 2OCF, 2QBW, 3CSB, 3CSG, 3K2M, 3QHT, 3RZW, 3UYO). After performing FoldX repair, random mutants were generated for each of the eleven structures by randomizing the BC, DE, and FG loop regions in accordance with the second generation diversity design scheme. At this point, baseline stabilities were individually calculated for each mutant. To analyze the stability impact upon residue substitution for each position in the diversified regions, all 19 natural residue substitutions were individually introduced to the random mutants. The change in stability ($\Delta\Delta G_{folding}$) upon mutation was then calculated for each PDB structure's collection of mutants. This process was iterated ($n > 50$) until the $\Delta\Delta G_{folding}$ associated with each position and residue converged to within 0.2 kcal/mol for at least five consecutive mutants. At each site, the stability impact upon substitution to each amino acid was calculated, creating stability matrices for each starting PDB. The sequences corresponding to the wild-type structures were aligned to account for loop length diversity. Average $\Delta\Delta G_{folding}$ values were calculated for all 20 amino acids at each diversified site (S3A Table).

2.3.9 Exposure

The likelihood of a loop position to be proximal to or directly involved with a target binding interface is influenced both by exterior exposure of the side chain (i.e. solvent accessible surface area) as well as its proximity to a region offering sufficient diversified surface area to enable the required enthalpic interactions. The site-specific exposure score

is calculated as the product of the solvent accessible surface area [91] and an estimated likelihood of residing at the target binding interface. The latter was quantified on a sitewise basis, averaged across eleven Fn3 crystal structures, using a geometric algorithm. Using Python, the fibronectin orientation that presents maximal diversified surface is identified as follows. BC, DE, and FG loop residues are mutated to alanine to remove wild-type residue size bias. The area of accessible (*i.e.*, visible in a two-dimensional projection as a planar approximation of the interface) diversified surface is calculated for each rotational orientation of fibronectin. The orientation that maximizes accessible view of the paratope, as well as any orientations within 5% of this projected area, is identified starting with a coarse-grained search and optimizing with a fine-grained search. For each site, the maximum area of accessible side chain surface within this set of optimized orientations is calculated. This calculation is repeated for all sites. Sitewise values are averaged across all Fn3 PDB models.

2.4. Results

2.4.1 Library design and construction

As a collection of starting points for the evolution of diverse ligands, a combinatorial library was created with various levels of diversity throughout the potential paratope of the hydrophilic fibronectin loops (Table 1). Each loop varied in length as guided by natural sequence frequency [93]. The core of the BC and FG loops – 2-5 sites and 2-6 sites, respectively, depending on loop length – had full amino acid diversity biased to mimic the third heavy chain complementarity-determining region (CDR) of antibodies. Two sites spatially within the BC and FG loop cores were constrained based on previous

experimental results. V29, which benefits as a small, reasonably hydrophobic amino acid [91,126], was constrained as A, S, or T. G79, which benefits from glycine bias [91], was mildly constrained to G, S, Y, D, N, or C to increase glycine frequency while mimicking CDRs. Twelve sites adjacent to the core of the BC and FG loops were afforded five levels of diversity: i) wild-type, ii) wild-type or serine (as a small, mid-hydrophilicity neutral interactor [127,128]), iii) wild-type, serine, or tyrosine (the most generally effective side chain for complementarity [49]), iv) moderate chemical diversity (A, C, D, G, N, S, T, or Y), or v) full antibody-mimicking amino acid diversity. All framework sites are conserved as the sequence of the tenth type III domain of human fibronectin with the hydrophilic mutations V1S, V4S, V11T, A12N, T16N, L19T, V45S, and V66Q [106] as well as the stabilizing D7N [129]. Five sub-libraries were constructed using separate DNA oligonucleotides with degenerate codons for each level of diversity. The five sub-libraries for each of the three loops were pooled at equimolar levels.

The gene libraries were transformed into a yeast surface display system [113], which yielded 2.0×10^8 transformants. DNA sequencing of 57 random naïve clones indicated 61% had full-length sequences, 16% contained stop codons naturally arising from the CDR' diversity, and 21% contained frameshifts. This finding was supported by flow cytometry analysis that revealed 64% of proteins were full-length as evaluated by the presence of a C-terminal c-myc epitope. Thus, the library contained 1.2×10^8 unique, full-length Fn3HP clones.

2.4.2 Selection and analysis of binding populations from first generation library

To identify a diverse set of selective ligands for a range of protein epitopes, the pooled library was sorted, using magnetic beads with immobilized protein targets and fluorescence-activated cell sorting (FACS), to yield binders to goat immunoglobulin G (IgG), rabbit IgG, lysozyme, or transferrin. These targets were selected to provide a diverse set of epitopes for targeting. Following a single round of mutagenesis, then two rounds of magnetic bead sorting, an enriched population of mutants was isolated that demonstrated mid-affinity, selective binding to transferrin. Selectivity was evidenced by a 30:1 ratio of fibronectin-displaying yeast selected for binding transferrin relative to binding negative control proteins (avidin and lysozyme). This population was then sorted for high-affinity binders via FACS. Similarly, though with one additional round of mutagenesis, mid- and high-affinity, selective binders for goat IgG, rabbit IgG, and lysozyme were identified (S2 Fig.).

The binding populations were sequenced via Illumina MiSeq resulting in $4.2x10^5$ sequences, including $1.1x10^5$ unique sequences. Analysis of similar sequences, identified and clustered using an in-house algorithm, revealed 3,590 unique families of unrelated sequences. Thus, a broad set of evolutionary solutions was identified for Fn3HP-based ligands.

Amino acid frequencies were measured at each site in the naïve and functionally evolved populations to reveal evolutionary impact. The amino acid distribution in evolved ligands exhibits substantial sitewise preferences in both the broadly diversified paratope core (Fig. 2C), in which the original library provided complementarity-biased diversity, and the adjacent sites (Fig. 2A) in which wild-type and neutral bias were implemented

along with lesser complementarity bias. For example, at site 30, N is depleted from 10% in the original library to 1% in evolved clones; conversely, K is enriched from 2% to 11%. At site 85, Y is depleted from 10% to 3% whereas P is enriched from 1% to 12%, suggesting a substantial evolutionary benefit. At site 24, S is depleted from 16% to 6%, and enrichment is broadly distributed by several amino acids. The sitewise amino acid distributions in the population of evolved fibronectin domains, and their deviation from the original unsorted library distributions, can be evaluated in a variety of ways to reveal evolutionary insight.

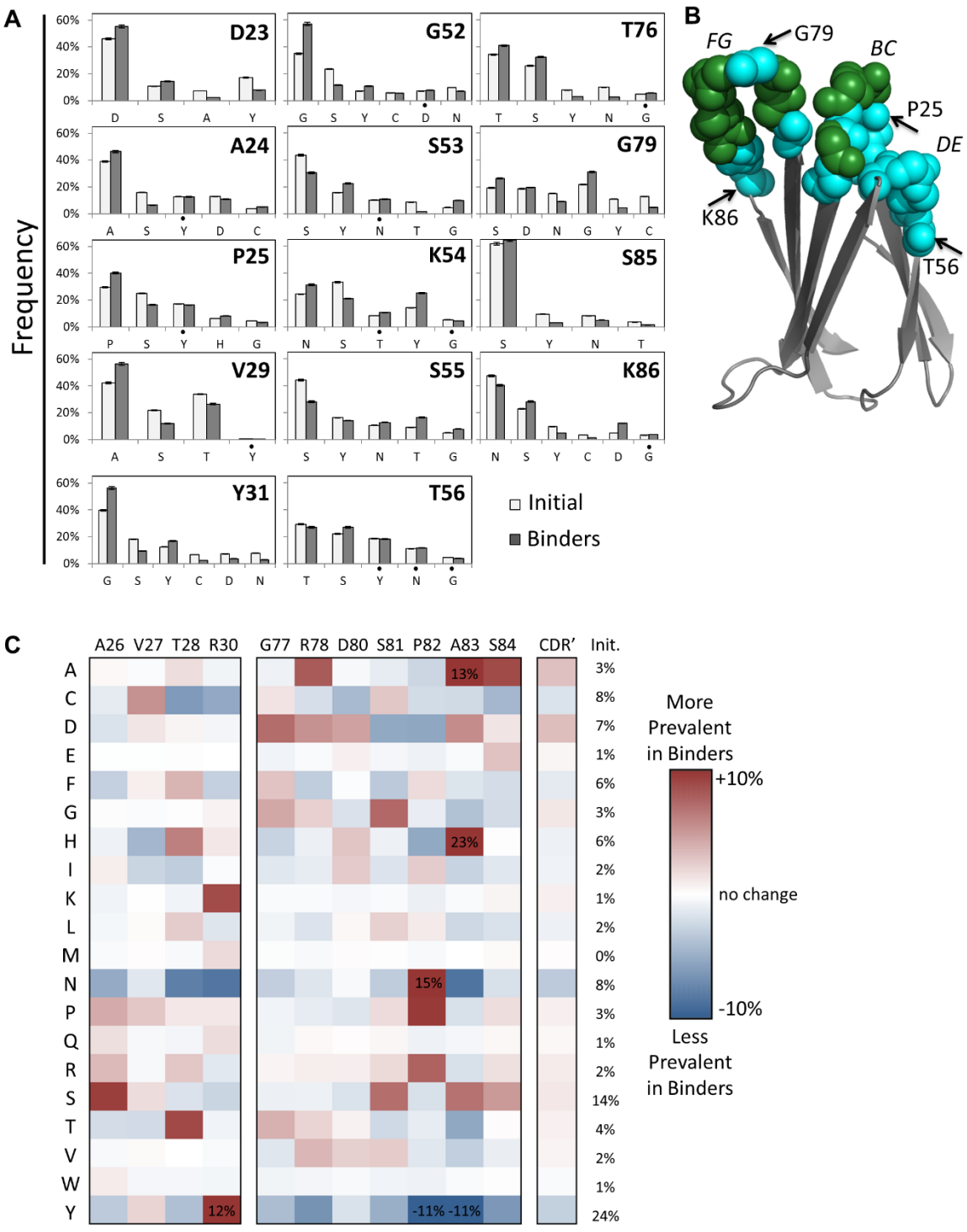


Figure 2-2. ***First generation sitewise amino acid distribution.*** **(A) The amino acid frequencies at the indicated sites for the first generation library (white) and the binding populations (gray). Bars and error bars are mean ± standard deviation. Statistical significance, while adjusting for family wise error rate using Bonferroni method, was achieved at level $\alpha = 0.005$ for all data with the exception of those denoted**

(•). (B) Solution structure (PDB: 1TTG) of wild-type fibronectin domain with backbone residues of diversified loop sites denoted by spheres. BC, DE, and FG loops are labeled as are several sites for reference. (C) The difference between the amino acid frequencies in the binding populations and the first generation library are shown for each amino acid at each site. The average across all fully diversified sites is also presented. *Init.* indicates the average amino acid frequency in the initial library across all CDR' sites.

2.4.3 Shannon Entropy

Shannon entropy ($H = -\sum_{i=1\text{-}20} p_i \ln p_i$, where p_i is the fraction of amino acid i at a particular site), which describes relative diversity ranging from 0 (fully conserved) to 4.3 (5% of each amino acid) [130], was calculated at each site for the naïve and evolved populations to measure constraint within functional ligands. Diversity at 21 of 25 sites is more constrained in the evolved repertoire than in the unsorted library as evidenced by reductions in the Shannon entropies (average: -0.2 units; S3 Fig.). Twelve sites have reduction of at least 0.15 units. Notably, G52 (2.8 to 2.0) and G31 (2.7 to 2.1) are largely driven by 22% and 17% increases in constraint of wild-type G. D23 (2.6 to 2.2), N54 (2.8 to 2.4), and T76 (2.8 to 2.4) are driven by broader constraint.

2.4.4 Wild-Type Constraint

Even with wild-type constraint averaging 39% in the edges of the BC loop in the original library, wild-type was further enriched in evolved sequences by an average of 11%: D23 (46% to 55%), A24 (39% to 46%), P25 (29% to 40%), and G31 (40% to 56%). Conversely, at the sites in the middle of the loop, fully diversified with complementarity bias, wild-type enrichment was less frequent: absent at A26 (2% to 2%), V27 (<1% to <1%), and R30 (2% to <1%) but present at T28 (6% to 15%).

The FG loop exhibited increases in wild-type constraint throughout: T76, G77, R78, G79, D80, S81, P82, S84, and S85 all elevated wild-type with an average increase of 6%. A83 also increased but with very few sequences available for analysis because of loop length variability. Also, wild-type was not considered at site 86 to eliminate the large, charged side chain K.

In the peripheral DE loop, wild-type constraint was diminished at three of four sites: S53 (44% to 31%), S55 (44% to 28%), and T56 (29% to 27%). Yet at the other site, G52, wild-type was substantially enriched in evolved binders (35% to 57%).

2.4.5 Enrichments Predicted by Proteome or Fibronectin Homologs

The extent to which sitewise biases were correlative with wild-type or homologous residues was evaluated. Of 39 residues with an enrichment of at least 5% (in magnitude, not relative increase), 13 are wild-type, 13 are wild-type homologs (A26S, R30K, S55T, T76S, S84A, N86D, N86S, V29A, S53G, S53D, S55D, G79S, and S81G), and 13 are non-homologous as defined by proteome-wide amino acid homology (BLOSUM62 [131]).

Of the 26 residues with enrichment of non-wild-type amino acids, 14 (54%) are also enriched in fibronectin homologs (S3B Table). While these homolog enrichments support the concept of applying consensus design [35] to combinatorial evolution, the other 12 residues (46%) highlight the limitations of such an approach (*i.e.*, the functional capacity of mutations not observed in natural evolution).

2.4.6 Serine

Serine was frequent in the naïve library adjacent to the fully diversified sites because of its purported ability to act as a neutral interactor, providing neither substantial detriment nor benefit. It was mildly reduced (from 29% to 24%) in the evolved repertoire.

2.4.7 Sites with Complementarity Bias

On average across all sites with full, complementarity-biased diversity, the biased distribution – based on the distribution observed in natural antibody repertoires – was generally preserved. Modest exceptions were that A and D were enriched by 3% each whereas N and Y were depleted by 3% each.

2.4.8 Comparison of Evolutionary Probabilities from the Original and Evolved Repertoires

The original and evolved sitewise amino acid frequencies were compared for their likelihood to yield functional ligands. The repertoire evident in the evolved population is a more efficient starting point for ligand discovery. Using leave-one-out cross-validation – with family clusters partitioned to ensure unbiased training sets (see Materials and Methods for details) – 6.6-fold more clones were more likely to be identified from the constrained repertoire versus the original library (Fig. 3A). The median increase in likelihood was 795%; *i.e.*, an evolved clone was 8-fold more likely to be identified from a combinatorial library based on the new repertoire than the original library. Notably, this original library already had constraint built into it from previous studies, detailed above (particularly constraint to a small hydrophobic residue at 29 and glycine bias at 79).

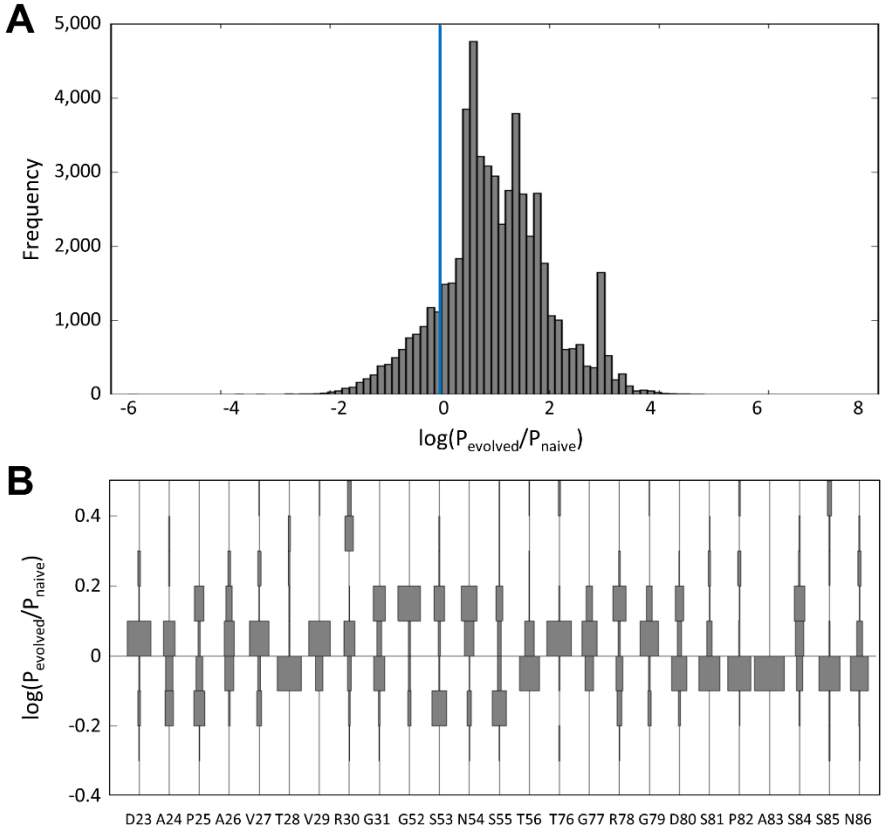


Figure 2-3. ***Distribution of evolutionary probabilities from the original and evolved repertoires*****. To estimate the likelihood of an evolved clone to have been identified from either the first generation repertoire or from that which is observed within the evolved population of binders, an exhaustive leave-one-out cross-validation (LOOCV) was conducted. The training set consisted of n-1 sequence clusters derived from the first generation mid-and high-affinity evolved binder data. Using the training set sequence clusters, a sitewise frequency matrix was calculated. The cross-validation test set consisted of a single sequence cluster that had been excluded from the training set analysis. Each sequence within the test set was assigned a probability for being observed within the original library (P_{naive}) and the training set ($P_{evolved}$). Log-odds score, $\log(P_{evolved}/P_{naïve})$, histograms are shown for test set sequences (A) and further evaluated on a sitewise basis (B).**

The probability of evolution from the constrained population versus the original population was also evaluated on a sitewise basis (Fig. 3B). Sites exhibit a range of evolutionary enhancements. For example, sites 52 and 54 have 107% and 213% increased likelihood of ligand discovery from the constrained repertoire whereas sites 78 and 81 are essentially neutral (1% and 2% increase towards constrained repertoire).

2.4.9 Loop Length Variation

The BC and DE loops frequently evolved loops one amino acid shorter than wild-type length but also exhibited frequent wild-type lengths (Fig. 4). Extended loops were very rare in both BC (2%) and DE (0.1%). Also, a two amino acid deletion in the BC loop, while present in 33% of the original library, only appeared in 2% of evolved domains. Loop lengths were more broadly distributed in the FG loop but with a notable evolutionary preference towards shorter loops.

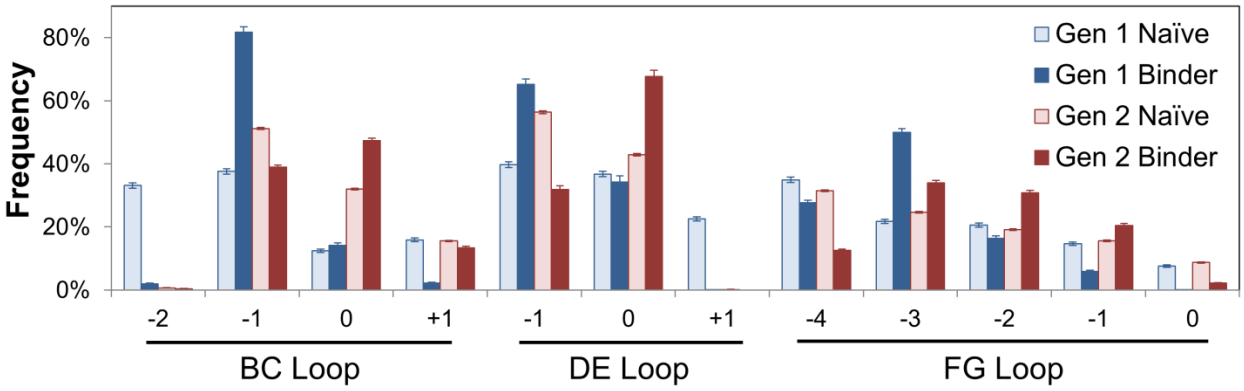


Figure 2-4. *Loop length variation.* Loop length frequencies in naïve and binder populations as identified by Illumina sequencing of first and second generation (details in Table 3). Bars and error bars represent mean ± standard deviation.

Table 2-4. High-throughput sequencing. Illumina sequencing statistics for 251 basepair paired end run on MiSeq for six uniquely barcoded libraries. Collectively, 7.2 million pass filter reads were obtained, including 25% PhiX control library. Within all pass filter read bases, 90% were above Q30 quality metric (99.9% read accuracy). Thorough sampling was observed across all six libraries with a coefficient of variance of 30%.

	Generation 1		Generation 2	
	Bead	Flow	Bead	Flow
Total Sequences	181,716	239,934	259,519	225,205
Unique Sequences	65,971	47,957	128,485	103,615
Unique Families	3,334	256	3,733	1,234

2.4.10 Design, construction, selection, and analysis of binding populations from second generation library

A. Library Design Based on First Generation Evolved Repertoire

The amino acid distributions observed in binding ligands evolved from the first generation library were used to design a second generation library, which enables further study of evolutionary repertoires from a more constrained library as well as evolution of useful ligands. Within the BC loop at site 23, wild-type D was enriched (46% in the naïve library to 55% in binders) whereas Y, introduced at 17% because of complementarity bias, was substantially depleted in binders to 8±1%. Small, mildly hydrophilic residues were also enriched – G (5% to 11%) and S (11% to 14%) – but not to levels comparable with wild-type bias. Thus, D23 was conserved in the second generation library. At site 24, wild-type A was elevated (39% to 46%), Y was maintained (13% to 13%), and S was

substantially depleted (16% to 6%). Thus, the options of wild-type conservation and mild diversity were further explored in the second generation. At site 25, wild-type P was enriched (29% to 40%), Y was maintained (17% to 16%), S slightly declined (25% to 17%), and H was enriched (6% to 8%). Wild-type conservation appears beneficial yet Y, S, and H warrant further consideration. At site 29, the more hydrophobic A was enriched (42% to 56%) whereas the more hydrophilic S and T were depleted (22% to 12% and 34% to 26%, respectively) but still frequent. Thus, second generation design maintained a distribution of AST. Y31 was enriched from 12% to 17% in binders. Glycine, which occurs with 31% frequency at this site within natural sequences of homologous proteins, increased in prevalence from 40% to 56%. Substantial decreases in S (18% to 9%), N (8% to 3%), and C (7% to 2%) were observed. The second generation library contained GY diversity. Wild-type G52 was enriched (35% to 57%) whereas S was depleted (23% to 11%). Wild-type conservation appears strongly beneficial at site 52. At sites 53-55, Y and N were enriched or maintained whereas S was depleted, but still present at reasonable levels. Thus, the second generation library implemented YNST diversity at these sites. T56 exhibited similar results without S depletion leading to TYSN design. At FG loop site 76, enrichment was observed for both wild-type T (34% to 41%) and S (26% to 32%) whereas N (10% to 3%) and Y (8% to 3%) were depleted in binders. Thus, the next design included T and S as well as the other small mid-hydrophilic G and A. At site 79, wild-type G was enriched (22% to 31%) as was S (19% to 27%). D (19% to 20%) was maintained but C (13% to 5%), Y (11% to 5%), and N (15% to 9%) decreased. Thus, GSDN diversity was used in the second library. Wild-type S85 maintained (62% to 65%), which prompts future conservation. At site 86, N was mildly decreased (48% to 40%) while S increased from

23% to 28% and Y decreased from 10% to 5%. The second generation design was intended to be NS but was erroneously synthesized as conserved N.

At G77, wild-type was enriched from 5% to 9% in binders. Y remained frequent in binders (20% to 16%), and D (11% to 18%) and T (4% to 8%) were enriched. Thus, G77 was set to GSYADTNC in generation two. Though several enrichments and depletions are evident elsewhere, all other CDR' sites will be maintained as CDR'.

The second generation library (Table 2) was constructed from degenerate oligonucleotides. $4.2x10^9$ yeast transformants were obtained. 71% were full-length as assessed by cytometry and corroborated by Sanger sequencing (67% full-length).

B. Selection and Analysis

Mid- and high-affinity binders to MET, lysozyme, and rabbit IgG, as well as mid-affinity binders for tumor necrosis factor receptor superfamily member 10b, were evolved and sequenced using Illumina MiSeq with barcodes to identify mid- and high-affinity binders. Sequences were aligned, clustered, and counted, with accommodations to reduce overcounting of highly similar sequences, using an in-house algorithm. $4.8x10^5$ sequences were collected with $2.3x10^5$ identified as unique (Table 3). The sitewise differences between amino acid frequencies in the naïve library and selected binders were calculated at uniquely constrained sites (Fig. 5A) and CDR' sites (Fig. 5B).

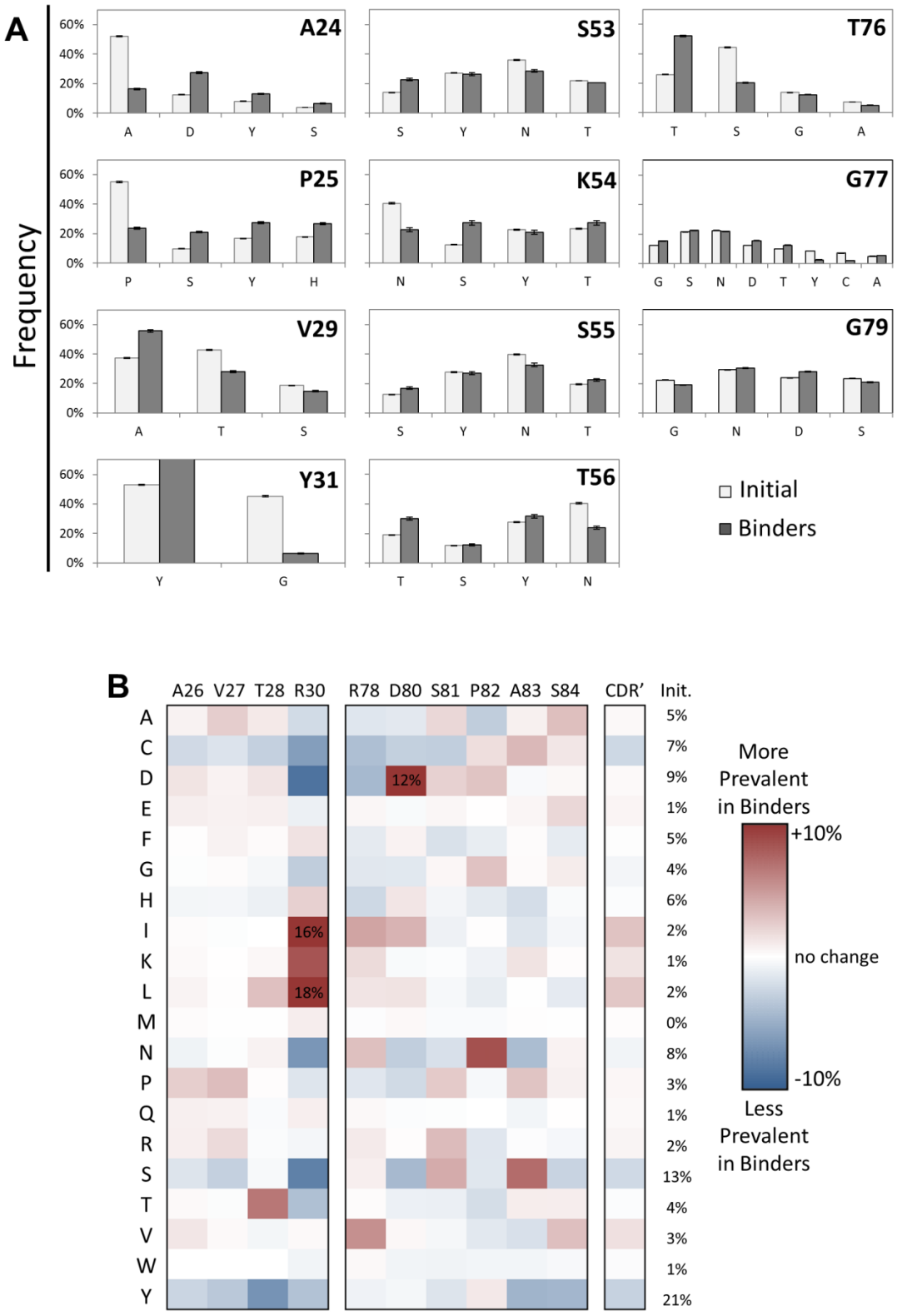


Figure 2-5. ***Second generation sitewise amino acid distribution.*** **(A) The amino acid frequencies at the indicated sites for the second generation library (white) and the binding populations (gray). Bars and error bars are mean ± standard deviation. (B)**

The difference between the amino acid frequencies in the binding population and the second generation library are shown for each amino acid at each site. The average across all fully diversified sites is also presented. *Init.* indicates the average amino acid frequency in the initial library across all CDR' sites.

Sites 24 and 25 exhibit similar results in which significant wild-type conservation was not maintained in binders (52% to 16% of A24 and 55% to 24% of P25) and the other amino acid options were elevated fairly uniformly. At site 29, alanine was increased (37% to 56%) while threonine was depleted (43% to 28%). At site 31, wild-type tyrosine is substantially enriched (53% to 92%) at the expense of glycine (45% to 7%).

In the middle of the DE loop, sites 53-55, asparagines were depleted from their overly high starting points (36% to 29%, 41% to 23%, and 40% to 33%) while serines, which were more rare than designed in the naïve library, were increased (14 to 23%, 13% to 27%, and 13 to 17%). Y and T were essentially maintained thereby supporting the SYNT diversity when equally implemented. Asparagine was also decreased at site 56 (41% to 24%), but wild-type threonine was preferentially increased (19% to 30%).

At the edge of the FG loop, wild-type T76 was increased from 26% to 52% in binders while serine was decreased from 44% to 20%. At site 77, wild-type G (12% to 15%) and aspartic acid (12% to 16%) were enriched, serine (21% to 22%) and asparagine (22% to 22%) were maintained, and tyrosine (9% to 3%) and cysteine (7% to 2%) were depleted. At site 79, the GDSN diversity was consistently maintained in binders.

C. Sites with Complementarity Bias

In the fully diversified sites, the antibody-inspired diversity was maintained for many amino acids. Sitewise exceptions include wild-type conservation at D80 (9% to

22%), T28 (4% to 11%), and S81 (14% to 18%) and enrichment of isoleucine and leucine at site 30 (4% to 38%). Slight overall exceptions – decrease in cysteine (7% to 4%) and increases in hydrophobics isoleucine, leucine, and valine (sum 8% to 15%) – all compensate for imperfections in the degenerate codons, yielding frequencies more in line with natural antibody repertoires. The decrease of cysteine residues is driven by a lack of enrichment of single-cysteine clones more than depletion of dual-cysteine clones, which is perhaps suggestive of beneficial disulfide bond formation (Fig. 6). Evaluation of cysteine pairs in dual-cysteine clones indicates a strong enrichment of clones with a cysteine in sites 26-28, especially 27, of the BC loop and 80-84 of the FG loop. Simultaneous cysteines at sites 76 and 84 are also frequently selected in binders.

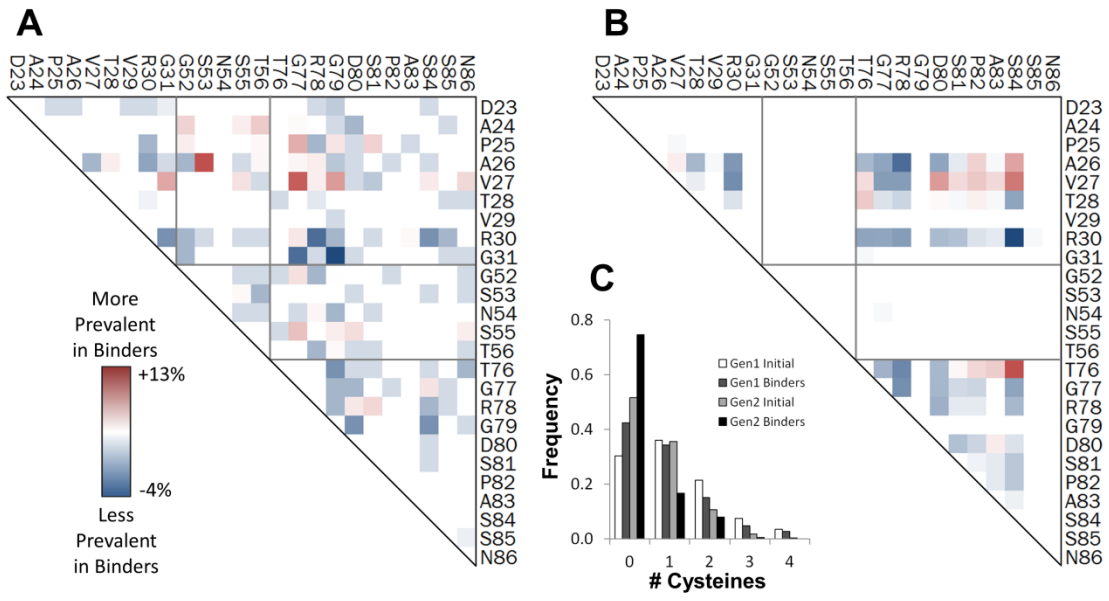


Figure 2-6. ***Cysteine frequency analysis.*** **(A, B) Change in pairwise frequency of clones containing exactly two cysteines (high-affinity populations minus naïve library) for the first generation (A) and second generation (B) libraries. (C) Frequencies of clones containing the indicated number of cysteines in initial libraries and binder populations.**

D. Loop Length Variation

Loop length analysis indicated diverse lengths were used in binders with a preference for wild-type lengths, or one amino acid decreases, in the BC and DE loops and a broader distribution in the FG loop with preference for 1-3 residues less than wild-type (Fig. 4). The longest FG loop, which is only observed in the tenth type III domain of human fibronectin but not the other fourteen human type III domains, is rarely observed (2%) in binders. The shortest FG loop was also less frequently observed in binders (13%) than in the unsorted library (31%).

E. Framework Mutations

The framework sites that were intended for conservation were also analyzed within the naïve and binder sequences to identify mutations, occurring during oligonucleotide synthesis, gene assembly, or directed evolution, that were preferentially present in binding clones. Four mutations were enriched (Table 4). Notably, the P44S mutation, and to a lesser extent S43F, are likely introduced by polyadenylation of the 3' tail by Taq polymerase during error-prone PCR [132] and then amplified by evolutionary selection.

Table 2-5. Enrichment of framework mutations. Full length fibronectin sequences from first and second generation naïve and binder populations were analyzed. Four framework positions demonstrated enrichment for non-wild type residues. Prevalence of these amino acids in natural homologs is shown in the two right most columns. Bottom row indicates median values of wild-type and any single mutant across all sites.

Mutation	Frequency in this work				Natural Frequency	
	First Generation		Second Generation			
	Initial	Binders	Initial	Binders	Wild-type	Mutant
I20V	0.5%	8%	0.5%	2%	16%	30%
S43F	0.6%	10%	1%	12%	19%	1%
P44S	17%	27%	11%	34%	23%	7%
I88T	0.1%	1%	0.2%	6%	13%	3%
median (all sites)	<0.01%	<0.01%	<0.01%	0.01%	21%	2%

F. Diversity

The binders generated from the second generation library exhibit a range of sitewise amino acid frequency distributions that are not purely spatial (Fig. 7A). Four sites exhibit Shannon entropies [130,133] in excess of 3.5. Three additional sites are in excess of 3.0. Nine sites have entropies from 2.0 – 3.0. Six diversified sites exhibit Shannon entropies below 2.0. Four sites were conserved, as designed, based on first generation library analysis.

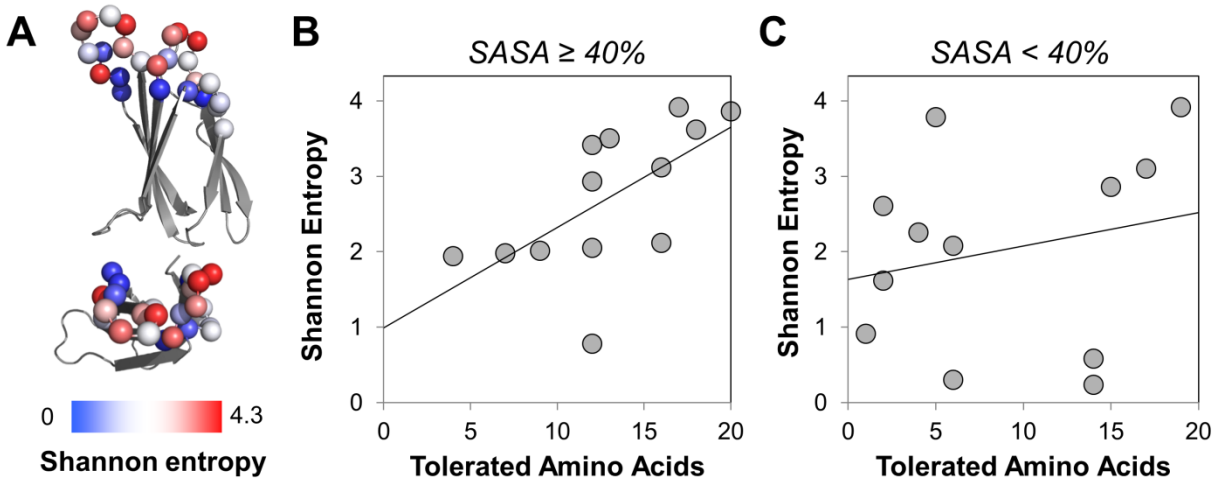


Figure 2-7. *Shannon entropy landscape among binding population.* (A) The α carbons of evaluated sites are shown in spheres colored based on Shannon entropy of binding sequences from the second generation campaigns. (B, C) The Shannon entropies at each site of second generation binders are plotted versus the number of amino acids that are tolerated at that site based on computational stability predictions. (B) Sites with solvent accessible surface area (in other fibronectin domain mutants) ≥40%. Pearson coefficient = 0.63. Slope = 0.13. (C) Sites with solvent accessible surface area <40%. Pearson coefficient = 0.22. Slope = 0.04.

2.4.11 Binder phenotypic characterization

By constraining diversity at select sites, we aim to improve the balance of inter- and intra-molecular interaction evolution and reduce destabilization upon mutation. Thus, we evaluated the stability of several fibronectin mutant populations: binders from both library generations in this work and binders evolved from binary (fully conserved framework, fully diversified loops) libraries from previous literature as well as the naïve second generation population and the parental fibronectin domains (human and hydrophilic mutant) (Fig. 8A). The first, second, and third quartile stabilities are higher for the first and second generation libraries relative to binders from less biased libraries. Further still, the median stability of the less biased library binders is less than even that of the naïve

members of the second generation library ($p < 0.05$). Note that Fn3HP is of essentially equivalent stability as Fn3 (84 ºC vs. 85 ºC).

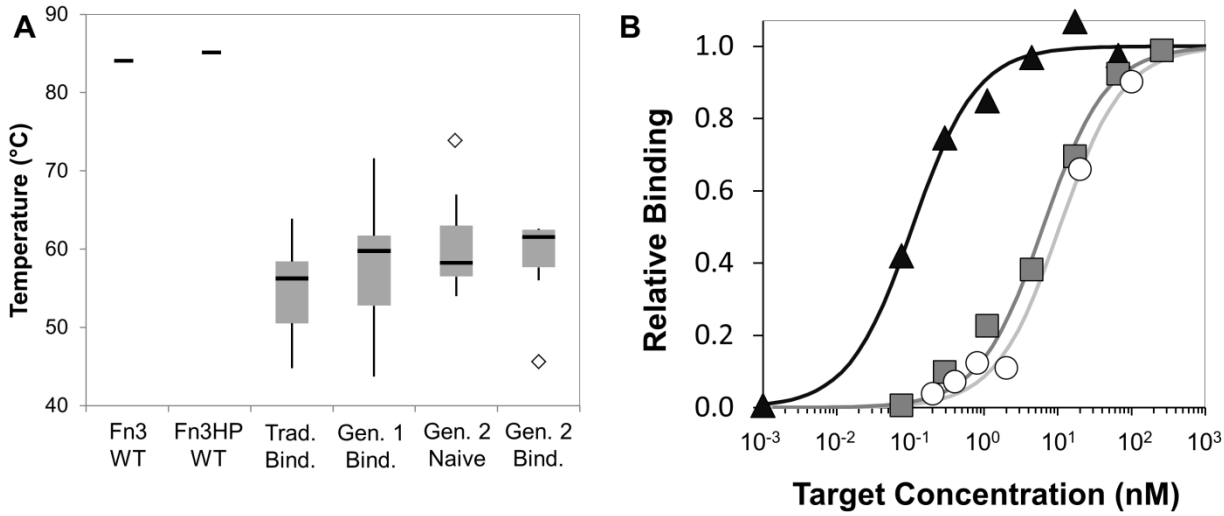


Figure 2-8. ***Clonal characterizations.*** **(A) Thermal stabilities of wild-type (WT) fibronectin [134] and Fn3HP are shown. Median (black line), second and third quartiles (gray box), upper and lower inner fences (vertical lines), and outliers (diamonds) are shown for a sampling of clones from binarily diversified traditional libraries [93,122,134] (n=15, *Trad. Bind.*) and three populations of the current study: first generation library binders, second generation naïve library, and second generation library binders. (B) Affinity titrations. Yeast displaying Fn3HP variants were incubated with the indicated concentration of biotinylated target molecule. Binding was detected by streptavidin-fluorophore and flow cytometry. Data points are from a single representative experiment. Affinities were calculated as 150 ±60 pM (rabbit IgG, clone 0.6.2, black triangles), 4 ±3 nM (lysozyme, clone 0.6.3, gray squares), and 11 nM (MET, clone 3.4.3, white circles).**

In addition to yielding stable binders, the second generation library yields high-affinity binders with little to no evolution. Three binder campaigns continued with additional sorts to identify the strongest binders in the population. Rabbit IgG and lysozyme binders were characterized following two rounds of magnetic bead selection, one round of cytometry sorting at target concentrations of 50 nM and a final round of cytometry sorting at 1 nM, wherein the top 1% of binding events were isolated. Titrations curves of

representative clones from the most stringently sorted, non-evolved rabbit IgG and lysozyme populations (Fig. 8B) revealed affinities of 150±60 pM and 4±3 nM, respectively. High-affinity MET binders were isolated following three iterations of evolution, each iteration consisting of two magnetic bead selections, one cytometry sort for full-length clones, and one round of error-prone PCR. A representative clone from the evolved MET binding population yielded an affinity of 11 nM (Fig. 8B).

2.4.12 Library Design Principles

The high-throughput binder engineering and analysis described herein provides one means of identifying the extents of diversification, as well as the relevant amino acids, at each site. To further explore the broad sequence data set of highly functional clones resulting from this study, we examined if any computational means could have guided this library refinement; *i.e.* if any computable parameters correlate with the evolved repertoire. The FoldX algorithm [124,135] was used to predict each site's tolerance to mutation. The change in stability ($\Delta\Delta G_{folding}$) upon mutation across >500 theoretical library variants (see Materials and Methods) was predicted for each of the twenty amino acids at each site. The mutational tolerance was assessed in terms of the number of minimally destabilizing ($\Delta\Delta G_{folding} < 0.75$ kcal/mol) amino acid substitutions allowed at each site. Observed amino acid diversity (Shannon entropy) in evolved binders correlated with computational mutational tolerance at exposed sites (Fig. 7B). While no correlation was observed for less exposed sites (Fig. 7C), it should be noted that the key outliers are the sites most distant from the paratope center (D23 and S85).

More broadly, four elements were evaluated for their correlation with evolved repertoires: sitewise amino acid frequencies from natural fibronectin homologs, sitewise computational stabilities of each amino acid at each site within the context of diverse fibronectin clones, complementarity-biased amino acid distributions observed in antibody CDRs, and sitewise sidechain exposure. The first two elements – frequency in natural homologs and computational stability – provide sitewise amino acid distributions; complementarity bias provides a single site-independent amino acid distribution; and the fourth element – residue exposure – provides a sitewise weight. We examined the ability of these four elements to combine to generate sitewise amino acid distributions that matched the experimentally observed frequencies. The relative weights of the first three elements were allowed to vary for each site based purely on the fourth element: solvent and target accessibility of that site (Eq. 1). Sitewise frequencies in natural homologs were calculated from 58,058 homologs from the Pfam database [136]. Sitewise computational stabilities were calculated using FoldX as described above. Complementarity bias was calculated as the amino acid distribution observed in expressed human and mouse antibody CDR-H3 sequences [58]. Solvent accessibility is the relative solvent accessible surface area [137] of each side chain averaged over 11 fibronectin structures. Target accessibility was scored based on the orientation of the amino acid side chain relative to the rest of the diversified paratope (detailed in Materials and Methods).

The relative weights of homolog frequency, computational stability, and complementarity bias were computed that, when linearly combined, yield a sitewise amino acid distribution that is most consistent with the evolved distribution. Weights were calculated for both an exposure-dependent and exposure-independent term, which were

summed. The weights provide a relative comparison of the predictive value of each parameter. To evaluate the predictive value overall, the parameters were compared to an unbiased control input matrix (uniformly 5% of each amino acid). Weighted inclusion of all elements yields a library design that matches the experimentally evolved distribution 22 standard deviations superior to designs based on unbiased input matrices (S3 Fig.). For well-exposed sites, amino acid diversity is effectively mimicked by 62% complementarity bias, 30% stability computation, and 8% natural frequency (Fig. 9). At sites with less exposure, natural amino acid frequency should be more heavily weighted at the expense of stability and, less so, complementarity. Design based on a single element is inferior to randomness for stability, marginally effective (2 standard deviations above random) for natural frequency, and strongly effective (16 standard deviations above random) for complementarity. In short, the evolved repertoire correlates strongly with complementarity and moderately with computed stability and frequency in natural homologs, varying dependent upon sidechain exposure.

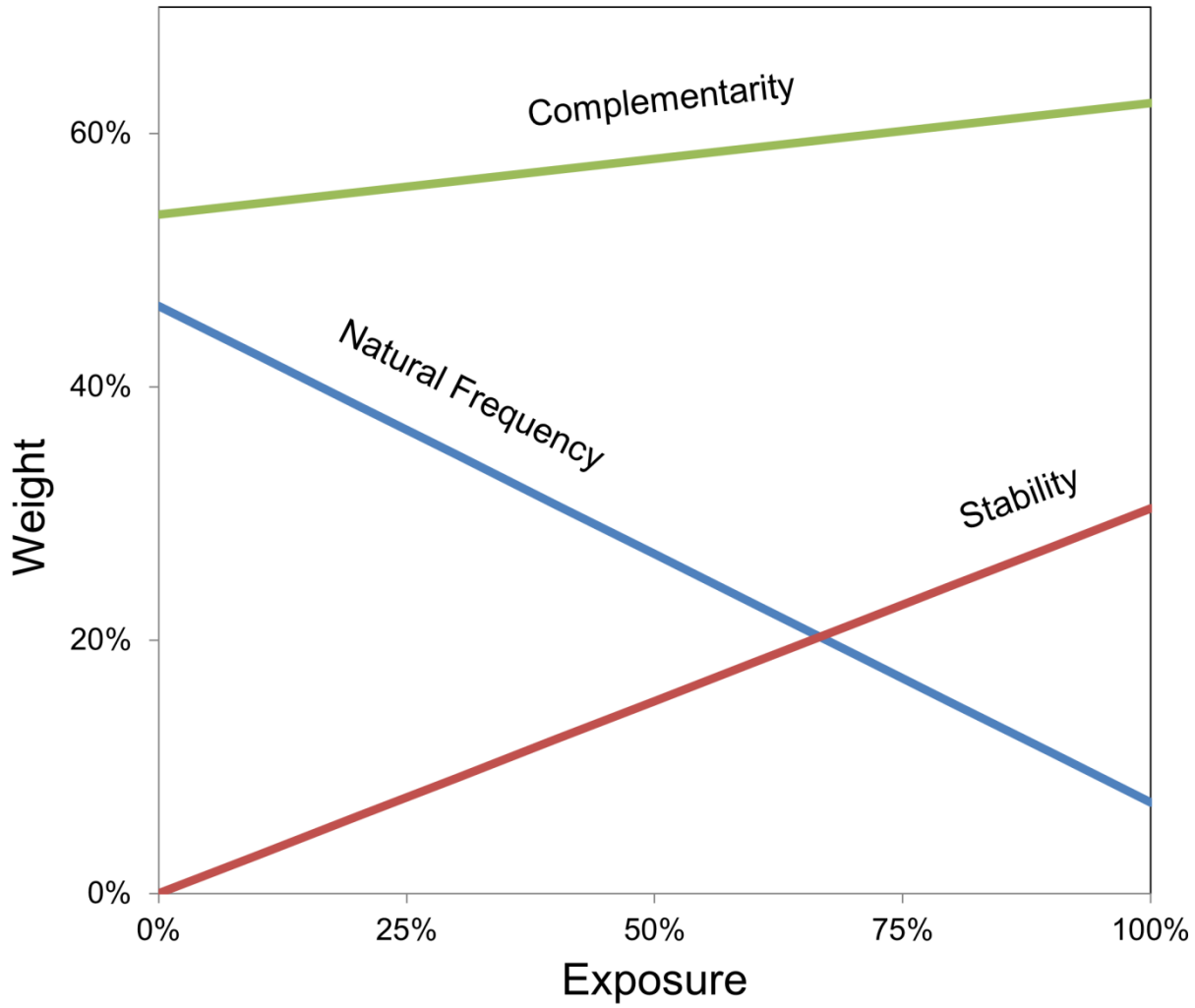


Figure 2-9. ***Correlative parametric analysis of amino acid distributions.*** **Sitewise evaluation of theoretical stability upon mutation and natural sequence frequency, as well as overall amino acid prevalence at binding interfaces of antibodies (i.e. complementarity), generate sitewise amino acid frequencies. The ability of these frequencies – scaled linearly based on solvent exposure and target exposure (Equation 1 and S3 Fig.) – to collectively mimic the observed sitewise amino acid distributions in binding populations is evaluated. The optimal weights for each contributing data set as a function of exposure are shown.**

2.5. Discussion

In the pursuit of a broadly functional combinatorial library capable of yielding binders to numerous targets, the benefit of diversification is unclear for sites peripheral

[108,138,139] to a 'hot spot' that enthalpically drives high-affinity binding [107,108,140]. Moreover, the location of the hot spot can differ across different epitopes being targeted by a scaffold. These peripheral sites can (a) directly contact target, (b,c) impact neighboring residue orientation to improve interfacial enthalpy or reduce entropic penalty upon binding, and/or (d) stabilize the protein. Yet these potential benefits can be offset by the inverse impacts: make unfavorable interfacial contact, worsen neighboring residue orientation, and/or destabilize the protein. If sufficient 'hot spot' interfacial area is not yet present for high-affinity binding, then additional sites must be diversified to enable favorable interaction. At some point, this expanded paratope provides sufficient interface for strong, specific affinity. Similar tradeoffs can be considered for peripheral sites. Given the typical detriment of random mutation, [19,141,142] the average peripheral mutation will hinder all four elements thereby suggesting against diversity. Though as a corollary, on average, mutations in the 'hot spot' will negatively impact the last two elements by worsening the entropic penalty upon binding and destabilizing the protein because of imperfect interactions with the conserved peripherals. Thus, peripherals need to be chosen to make neutral to good contact with: (a) intermolecular target; (b,c) intramolecular neighbors involved in binding; and (d) all intramolecular neighbors. For *a*-*c*, since beneficial interactions will be unlikely, amino acids – such as serine [48,55,128] – that yield relatively neutral interactions may be wise. For *d*, beneficial interactions are likely for the wild-type residue and conserved neighbors based on their coevolution so conservation should be the aim. Since the precise locations of the hot spots and these transitions will vary for each new ligand-target interface (Fig. 1C), we hypothesized the evolved repertoires will exhibit a gradient of diversity from extensive diversity in the potential paratope hot spot to full

conservation in the framework. Importantly, this gradient includes moderate diversity, with structural bias, within the paratope interfacing with target yet peripheral to the hot spot. Moreover, more mild diversity is included adjacent to the interfacial residues to yield optimal intramolecular contacts with the newly identified paratope. The range of Shannon entropies (Fig. 7) and amino acid frequency distributions (Figs. 2 and 5) across many binding sequences against several targets support the benefit of a gradient of diversities within combinatorial libraries. The particular amino acid distributions support the hypothesized benefits of wild-type conservation, serine bias, and complementarity bias, at appropriate sites.

Sitewise optimization of this gradient between intra- and inter-molecular interaction biases can be achieved with broad, high-throughput binder generation and deep sequencing as demonstrated here. Yet this requires a sufficiently effective library to generate numerous binders, which may be difficult for new scaffolds or paratopes. Analysis conducted in the current study (Fig. 9) provides further evidence that initial combinatorial library design can be guided by complementarity-determining residues and, when available, natural homolog frequencies, stability data (theoretical or experimental), and side chain exposure to solvent and target. Ongoing studies can quantify the values of complementarity, stability, natural sequence frequency, exposure – and potentially other metrics – in the context of other protein topologies and other functions, such as catalysis.

Sitewise optimization of amino acid frequency, with a range of diversities, can be implemented in numerous ways. Trimer phosphoramidite codons can be used in oligonucleotide synthesis, [74] which enables precise control over each distribution but elevates synthesis complexity and cost. Independent oligonucleotides can be synthesized

for each loop sequence, which further elevates control by enabling pairwise (and higher order) site design albeit at an elevated synthesis scope. Simpler, less expensive single-nucleotide mixed degenerate oligonucleotide synthesis can approximate many amino acid distributions, especially with the inclusion of unbalanced nucleotide frequencies as used in this study. The amino acid distribution within antibody CDR-H3 can be closely approximated by unbalanced single-nucleotide methods [91], but it must compromise on the genetic code connectivity of glycine, tyrosine, and cysteine. Achieving the desired high frequencies of tyrosine (20%) and glycine (16%) yields much more cysteine than desired (10%). In these libraries, we opted to maintain high tyrosine (17%) while limiting cysteine (5%) at the expense of low glycine (4%). Wild-type glycine bias at sites G52 and G79, as well as G77 in the second generation library, enabled this successful compromise as glycine at fully diversified sites was only marginally enriched in binders relative to the original library (2.7% to 4.0% in the first generation; 3.8% to 4.7% in the second generation). Thus, selective sitewise bias is able to effectively provide the evolutionary benefit of the presence of glycine within the DE and FG loops.

Note that the sublibrary synthesis approach in generation one (Table 1) yields coupling between sites within each loop. For example, wild-type D23 conservation pulls wild-type conservation in other BC sites during generation one analysis. In the absence of this coupling in generation two analysis, wild-type conservation at other BC sites (A24, P25, and Y31) is reduced. In the DE loop, wild-type G52 conservation pulls N54 conservation, which converts to N54 depletion in generation two in the absence of G52 coupling. Thus, when evaluating a new scaffold or paratope design, sublibrary construction

enables analysis of numerous diversification strategies, but care must be taken to consider coupled sites.

While cysteines were overall depleted from binding sequences relative to the naïve library, select inter- and intra-loop cysteine pairs were enriched. These occurred at proximal locations that are structurally sensible for disulfide bond formation, but further validation is needed to confirm the existence of disulfide bonding. Enhanced evolutionary efficiency of this class of clones warrants consideration of biased design to drive the conformational restriction beneficial to numerous topologies including stapled helical peptides [143], shark new antigen receptors [144], camelid antibody domains [145], and previous fibronectin clones [92]. Yet, while entropically beneficial, this conformational restriction may limit the diversity of paratopes that a library can present. Moreover, it eliminates the benefits of cysteine-free ligands: intracellular use, efficient cytoplasmic production in bacteria, and genetically introduced cysteines for site-specific thiol chemistry.

In conclusion, the extent of diversity and particular amino acid distributions consistent with a broad capacity for evolution of new binding activity were determined for a combinatorial library of a hydrophilic fibronectin domain. A gradient of diversity including sitewise constraints was revealed in evolved clones, which is consistent with natural antibody repertoires but converse to current synthetic scaffold combinatorial library designs. Importantly, the extensive dataset allowed for initial characterization of a broadly applicable data driven library design model that guides the most beneficial distribution of amino acids at each position.

2.6. Acknowledgements

We are grateful to Aaron Becker at the University of Minnesota Genomics Center for assistance with Illumina sequencing, the Masonic Cancer Center Flow Cytometry Core Facility, Lauren Mills at the Minnesota Supercomputing Institute for guidance on data analysis, Brett Case for homolog analysis, and Max Kruziki for helpful comments on the manuscript.

2.7. Supplemental Information

Figure S2-1. Schematic of Study. Ligands to six different targets are generated using yeast display of Fn3HP mutants with magnetic and fluorescence selections. Deep sequencing reveals functional ligand sequences. Multiple informatics analyses indicate sitewise amino acid frequencies and their implications, relative evolutionary fitness of constrained library design, and the correlation of constrained designs with computable parameters.

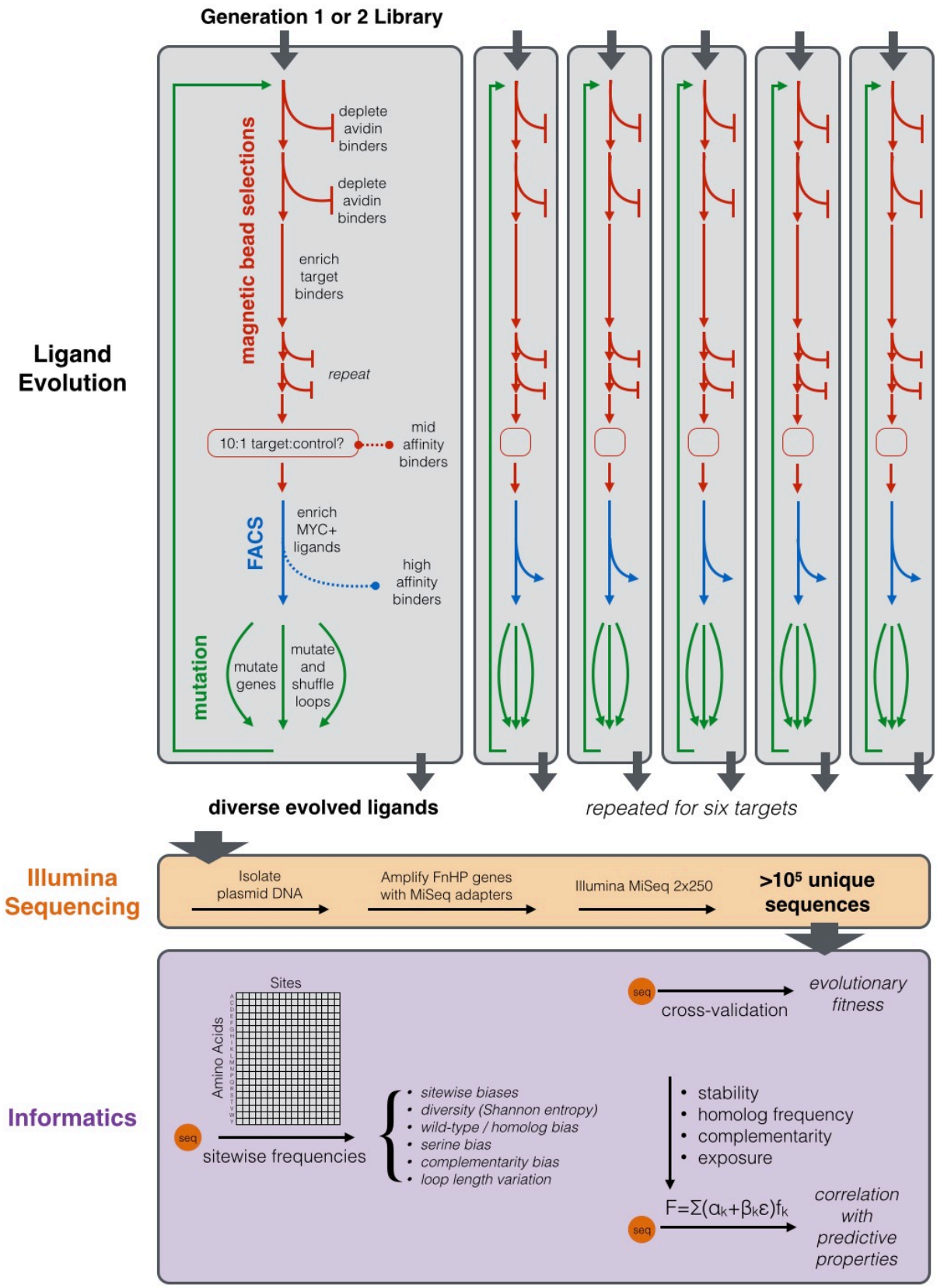

Generation 1 or 2 Library
Ligand Evolution
magnetic bead selections
deplete avidin binders
deplete avidin binders
enrich target binders
repeat
10:1 target:control?
mid affinity binders
FACS
enrich MYC+ ligands
high affinity binders
mutation
mutate genes
mutate and shuffle loops
diverse evolved ligands
repeated for six targets
Illumina Sequencing
Isolate plasmid DNA
Amplify FnHP genes with MiSeq adapters
Illumina MiSeq 2x250
>10⁵ unique sequences
Informatics
Sites
Amino Acids
seq
sitewise frequencies
sitewise biases
diversity (Shannon entropy)
wild-type / homolog bias
serine bias
complementarity bias
loop length variation
seq
cross-validation
evolutionary fitness
stability
homolog frequency
complementarity
exposure
F=Σ(αk+βkε)fk
seq
correlation with predictive properties

Figure S2-2. Binder selection via fluorescence activated cell sorting (FACS). Diverse populations of evolved clones were isolated via cytometry. Two representative campaigns, goat IgG and lysozyme are shown (left column) with 50 nM multivalent target (3:1 stoichiometry of target preloaded on streptavidin-AlexaFluor488). Specificity controls (right column) were conducted under identical multivalent labeling conditions where non-cognate proteins lysozyme and rabbit IgG were incubated with the evolved goat IgG and lysozyme populations, respectively. Binding clones within the gated region (solid line) were isolated from each target sample for further analysis. The analogous gated regions (dashed lines) within the control samples are shown for comparison.

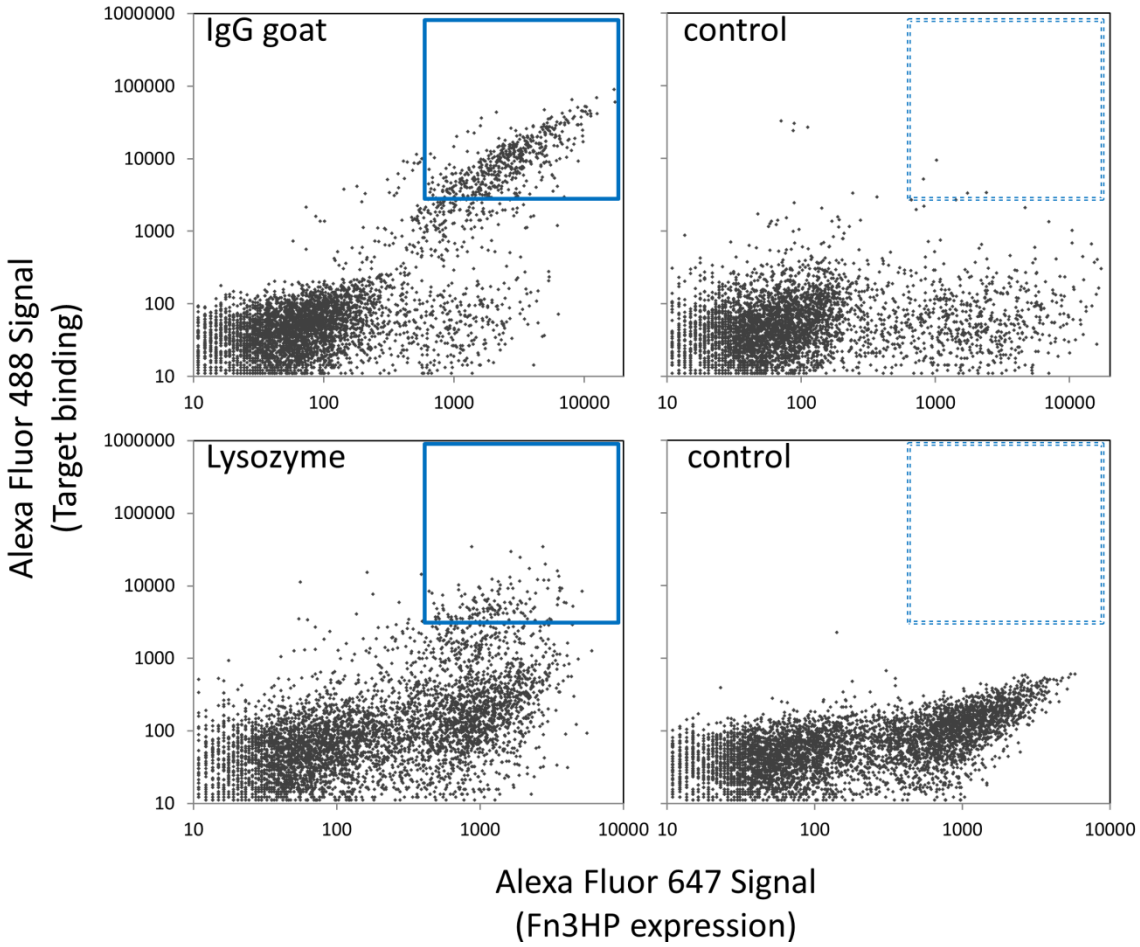
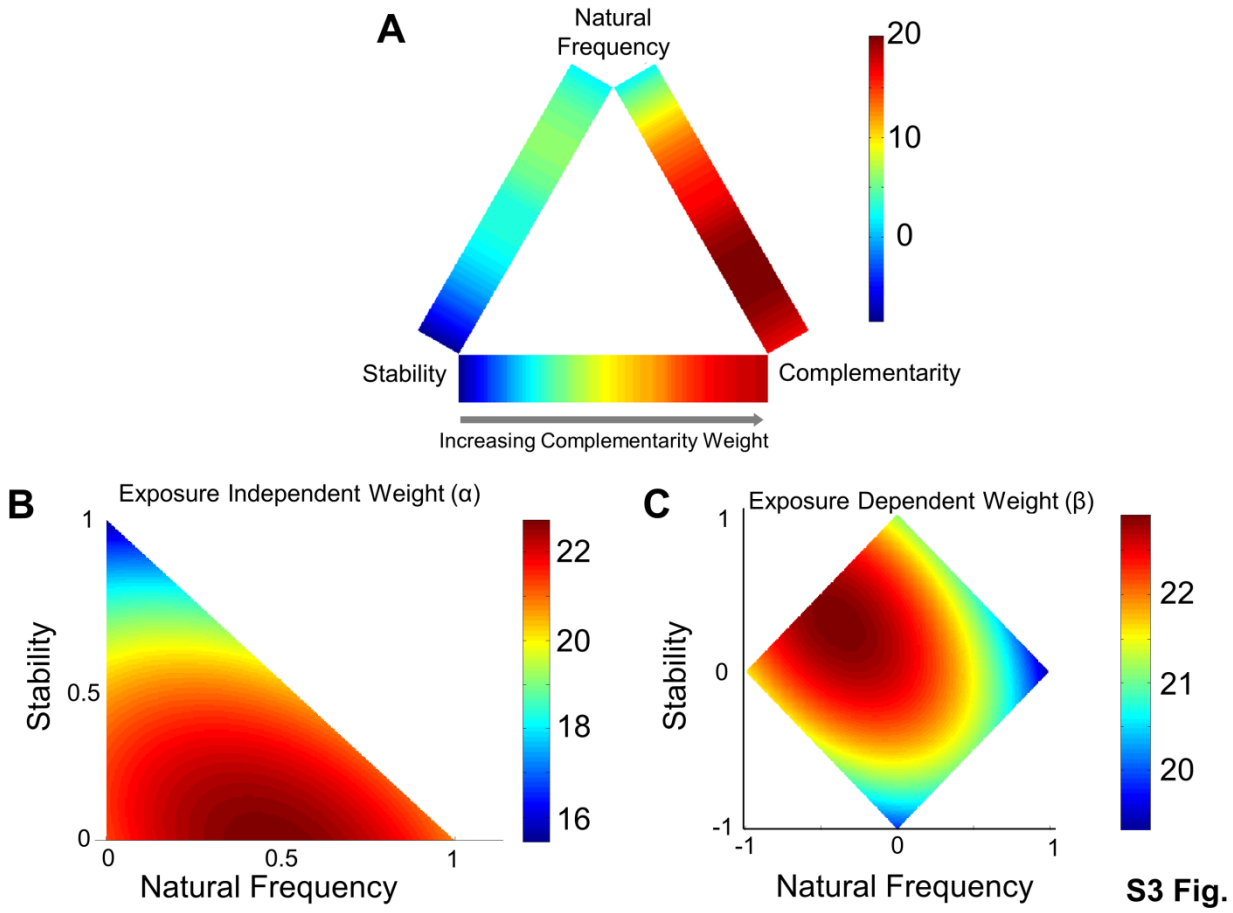


Figure S2-3. Correlative parametric analysis of amino acid distributions. Sitewise evaluation of theoretical stability upon mutation and natural sequence frequency, as well as overall amino acid prevalence at binding interfaces of antibodies (i.e. complementarity), generate sitewise amino acid frequencies. The ability of these frequencies – scaled linearly based on solvent exposure and target exposure (Equation 1) – to collectively mimic the observed sitewise amino acid distributions in binding populations is evaluated. The optimal weights for each contributing data set as a function of exposure are shown. (A-C) For the indicated weights of each metric, the other free parameters were varied to optimize the match between modeled sitewise amino acid distributions and experimentally observed sequences. The qualities of the fits are presented as the number of standard deviations above the fit obtained if unbiased data are used (*i.e.* uniformly 5% amino acid diversity rather than stability, homology, and complementarity bias). (A) Relative success when limited to two data inputs. Exposure independent (α) and dependent (β) weights are varied, subject to the indicated average weight, to maximize fit. (B) Sensitivity of exposure independent weights (α). All α values are fixed as indicated (note that all α's sum to 1 so complementarity weight is implicit). Exposure dependent weights are varied to maximize fit. 55% complementarity, 45% natural sequence frequency, and 0% theoretical stability optimize fit. (C) As in (B) but with set β values and varied α values.


A
Natural Frequency
20
10
0
Stability
Complementarity
Increasing Complementarity Weight
B
Exposure Independent Weight (α)
Stability
1
0.5
0
0
0.5
1
Natural Frequency
22
20
18
16
C
Exposure Dependent Weight (β)
Stability
1
0
-1
-1
0
1
Natural Frequency
22
21
20

S3 Fig.

Table S2-1. Hydrophilic fibronectin (Fn3HP) sequence information and library oligonucleotides. (A) Fn3HP framework amino acid and DNA sequence. All framework sites are conserved as the sequence of the tenth type III domain of human fibronectin with the hydrophilic mutations V1S, V4S, V11T, A12N, T16N, L19T, V45S, and V66Q [106], underlined, as well as the stabilizing D7N [129], shown with overbar.

SSDSPR$\bar{N}$LEV TNATPNSLTI SWxxxxxxxx xYRITYGETG GNSPSQEFTV PxxxxxATIS GLKPGQDYTI TVYAVxxxxx xxxxxxPISI NYRTEIDKPS Q

TCC TCC GAC TCT CCG CGT AAC CTG GAG GTT ACC AAC GCA ACT CCG AAC TCT CTG ACT ATT TCT TGG
NNN NNN NNN NNN NNN NNN NNN NNN NNN TAC CGT ATC ACC TAC GGC GAA ACT GGT GGT AAC TCC CCG
AGC CAG GAA TTC ACT GTT CCG NNN NNN NNN NNN NNN GCG ACC ATC AGC GGT CTG AAA CCG GGC CAG
GAT TAT ACC ATT ACC GTG TAC GCT GTA NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN CCA ATC
AGC ATC AAT TAT CGC ACC GAA ATC GAC AAA CCG TCT CAG

(B) Oligonucleotide DNA sequences used for constructing generation one library. Sequences are composed of standard nucleotides (ACGT), degenerate nucleotides (RYMKSWHBVDN), and a specialty codon mix (xyz) which uses the following nucleotide frequencies : 20% A, 15% C, 25% G, and 40% T at site 1, 50% A, 25% C, 15% G, and 10% T at site 2, and 0% A, 45% C, 10% G, and 45% T at site 3. Oligos are arranged by loop (BC, DE, FG), sublibraries a-e, and amino acid length of the diversified region within the loop**.**

Loop/Sublib/Len (AA)	Oligo Sequence
BC/a/10	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/b/10	ACTCTCTGACTATTTCTTGGKMTKCCYCCxyzxyzxyzxyzDCTxyzRGCTACCGTATCACCTACGGCGAAAC
BC/c/10	ACTCTCTGACTATTTCTTGGKMCKMTYMTxyzxyzxyzxyzDCTxyzDRTTACCGTATCACCTACGGCGAAAC
BC/d/10	ACTCTCTGACTATTTCTTGGDVTDVTDVTxyzxyzxyzxyzDCTxyzDVTTACCGTATCACCTACGGCGAAAC
BC/e/10	ACTCTCTGACTATTTCTTGGxyzxyzxyzxyzxyzxyzxyzDCTxyzxyzTACCGTATCACCTACGGCGAAAC
BC/a/9	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/b/9	ACTCTCTGACTATTTCTTGGKMTKCCYCCxyzxyzxyzDCTxyzRGCTACCGTATCACCTACGGCGAAAC
BC/c/9	ACTCTCTGACTATTTCTTGGKMCKMTYMTxyzxyzxyzDCTxyzDRTTACCGTATCACCTACGGCGAAAC
BC/d/9	ACTCTCTGACTATTTCTTGGDVTDVTDVTxyzxyzxyzDCTxyzDVTTACCGTATCACCTACGGCGAAAC
BC/e/9	ACTCTCTGACTATTTCTTGGxyzxyzxyzxyzxyzxyzDCTxyzxyzTACCGTATCACCTACGGCGAAAC
BC/a/8	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/b/8	ACTCTCTGACTATTTCTTGGKMTKCCYCCxyzxyzDCTxyzRGCTACCGTATCACCTACGGCGAAAC
BC/c/8	ACTCTCTGACTATTTCTTGGKMCKMTYMTxyzxyzDCTxyzDRTTACCGTATCACCTACGGCGAAAC
BC/d/8	ACTCTCTGACTATTTCTTGGDVTDVTDVTxyzxyzDCTxyzDVTTACCGTATCACCTACGGCGAAAC
BC/e/8	ACTCTCTGACTATTTCTTGGxyzxyzxyzxyzxyzDCTxyzxyzTACCGTATCACCTACGGCGAAAC

BC/a/7	ACTCTCTGACTATTTCTTGGGACGCACCAxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/b/7	ACTCTCTGACTATTTCTTGGKMTKCCYCCxyzDCTxyzRGCTACCGTATCACCTACGGCGAAAC
BC/c/7	ACTCTCTGACTATTTCTTGGKMCKMTYMTxyzDCTxyzDRTTACCGTATCACCTACGGCGAAAC
BC/d/7	ACTCTCTGACTATTTCTTGGDVTDVTDVTxyzDCTxyzDVTTACCGTATCACCTACGGCGAAAC
BC/e/7	ACTCTCTGACTATTTCTTGGxyzxyzxyzxyzDCTxyzxyzTACCGTATCACCTACGGCGAAAC
DE/a/6	CGAGCCAGGAATTCACTGTTCCGGGATCATCAAACTCAACAGCGACCATCAGCGGTCTGAAAC
DE/b/6	CGAGCCAGGAATTCACTGTTCCGRGCTCCTCCARTTCCASCGCGACCATCAGCGGTCTGAAAC
DE/c/6	CGAGCCAGGAATTCACTGTTCCGDRTWMTWMTWMTWMTWMTGCGACCATCAGCGGTCTGAAAC
DE/d/6	CGAGCCAGGAATTCACTGTTCCGDVTDVTDVTDVTDVTDVTGCGACCATCAGCGGTCTGAAAC
DE/e/6	CGAGCCAGGAATTCACTGTTCCGxyzxyzxyzxyzxyzxyzGCGACCATCAGCGGTCTGAAAC
DE/a/5	CGAGCCAGGAATTCACTGTTCCGGGATCAAACTCAACAGCGACCATCAGCGGTCTGAAAC
DE/b/5	CGAGCCAGGAATTCACTGTTCCGRGCTCCARTTCCASCGCGACCATCAGCGGTCTGAAAC
DE/c/5	CGAGCCAGGAATTCACTGTTCCGDRTWMTWMTWMTWMTGCGACCATCAGCGGTCTGAAAC
DE/d/5	CGAGCCAGGAATTCACTGTTCCGDVTDVTDVTDVTDVTGCGACCATCAGCGGTCTGAAAC
DE/e/5	CGAGCCAGGAATTCACTGTTCCGxyzxyzxyzxyzxyzGCGACCATCAGCGGTCTGAAAC
DE/a/4	CGAGCCAGGAATTCACTGTTCCGGGATCATCAACAGCGACCATCAGCGGTCTGAAAC
DE/b/4	CGAGCCAGGAATTCACTGTTCCGRGCTCCTCCASCGCGACCATCAGCGGTCTGAAAC
DE/c/4	CGAGCCAGGAATTCACTGTTCCGDRTWMTWMTWMTGCGACCATCAGCGGTCTGAAAC
DE/d/4	CGAGCCAGGAATTCACTGTTCCGDVTDVTDVTDVTGCGACCATCAGCGGTCTGAAAC
DE/e/4	CGAGCCAGGAATTCACTGTTCCGxyzxyzxyzxyzGCGACCATCAGCGGTCTGAAAC
FG/a/11	CATTACCGTGTACGCTGTAACAxyzxyzDRTxyzxyzxyzxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/b/11	CATTACCGTGTACGCTGTAASCxyzxyzDRTxyzxyzxyzxyzxyzTCCARTCCAATCAGCATCAATTATCGCAC
FG/c/11	CATTACCGTGTACGCTGTAWMTxyzxyzDRTxyzxyzxyzxyzxyzWMTWMTCCAATCAGCATCAATTATCGCAC
FG/d/11	CATTACCGTGTACGCTGTADVTxyzxyzDRTxyzxyzxyzxyzxyzDVTDVTCCAATCAGCATCAATTATCGCAC

FG/e/11	CATTACCGTGTACGCTGTAxyzxyzxyzDRTxyzxyzxyzxyzxyzxyzxyzCCAATCAGCATCAATTATCGCAC
FG/a/10	CATTACCGTGTACGCTGTAACAxyzxyzDRTxyzxyzxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/b/10	CATTACCGTGTACGCTGTAASCxyzxyzDRTxyzxyzxyzxyzTCCARTCCAATCAGCATCAATTATCGCAC
FG/c/10	CATTACCGTGTACGCTGTAWMTxyzxyzDRTxyzxyzxyzxyzWMTWMTCCAATCAGCATCAATTATCGCAC
FG/d/10	CATTACCGTGTACGCTGTADVTxyzxyzDRTxyzxyzxyzxyzDVTDVTCCAATCAGCATCAATTATCGCAC
FG/e/10	CATTACCGTGTACGCTGTAxyzxyzxyzDRTxyzxyzxyzxyzxyzxyzxyzCCAATCAGCATCAATTATCGCAC
FG/a/9	CATTACCGTGTACGCTGTAACAxyzxyzDRTxyzxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/b/9	CATTACCGTGTACGCTGTAASCxyzxyzDRTxyzxyzxyzTCCARTCCAATCAGCATCAATTATCGCAC
FG/c/9	CATTACCGTGTACGCTGTAWMTxyzxyzDRTxyzxyzxyzWMTWMTCCAATCAGCATCAATTATCGCAC
FG/d/9	CATTACCGTGTACGCTGTADVTxyzxyzDRTxyzxyzxyzDVTDVTCCAATCAGCATCAATTATCGCAC
FG/e/9	CATTACCGTGTACGCTGTAxyzxyzxyzDRTxyzxyzxyzxyzxyzxyzCCAATCAGCATCAATTATCGCAC
FG/a/8	CATTACCGTGTACGCTGTAACAxyzxyzDRTxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/b/8	CATTACCGTGTACGCTGTAASCxyzxyzDRTxyzxyzTCCARTCCAATCAGCATCAATTATCGCAC
FG/c/8	CATTACCGTGTACGCTGTAWMTxyzxyzDRTxyzxyzWMTWMTCCAATCAGCATCAATTATCGCAC
FG/d/8	CATTACCGTGTACGCTGTADVTxyzxyzDRTxyzxyzDVTDVTCCAATCAGCATCAATTATCGCAC
FG/e/8	CATTACCGTGTACGCTGTAxyzxyzxyzDRTxyzxyzxyzxyzCCAATCAGCATCAATTATCGCAC
FG/a/7	CATTACCGTGTACGCTGTAACAxyzxyzDRTxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/b/7	CATTACCGTGTACGCTGTAASCxyzxyzDRTxyzTCCARTCCAATCAGCATCAATTATCGCAC
FG/c/7	CATTACCGTGTACGCTGTAWMTxyzxyzDRTxyzWMTWMTCCAATCAGCATCAATTATCGCAC
FG/d/7	CATTACCGTGTACGCTGTADVTxyzxyzDRTxyzDVTDVTCCAATCAGCATCAATTATCGCAC
FG/e/7	CATTACCGTGTACGCTGTAxyzxyzxyzDRTxyzxyzxyzCCAATCAGCATCAATTATCGCAC

Table S2-2. Oligonucleotide DNA sequences used for constructing generation two library. Sequences are composed of standard nucleotides (ACGT), degenerate nucleotides (RYMKSWHBVDN), and a specialty codon mix (xyz) which uses the following nucleotide frequencies : 20% A, 15% C, 25% G, and 40% T at site 1, 50% A, 25% C, 15% G, and 10% T at site 2, and 0% A, 45% C, 10% G, and 45% T at site 3. Oligos are arranged by loop (BC, DE, FG), loop specific sublibraries, and amino acid length of the diversified region within the loop**.**

Loop/Sublib/Len (AA)	Oligo Sequence
BC/a/10	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/b/10	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/c/10	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/d/10	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/e/10	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/f/10	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/g/10	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/h/10	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/a/9	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/b/9	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/c/9	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/d/9	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/e/9	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/f/9	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/g/9	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC

BC/h/9	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/a/8	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/b/8	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/c/8	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/d/8	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/e/8	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/f/8	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
BC/g/8	ACTCTCTGACTATTTCTTGGGACGCACCAxyzxyzDCTxyzGGATACCGTATCACCTACGGCGAAAC
BC/h/8	ACTCTCTGACTATTTCTTGGGACDMTYMTxyzxyzDCTxyzTATTACCGTATCACCTACGGCGAAAC
DE/a/5	CGAGCCAGGAATTCACTGTTCCGGGAWMTWMTWMTWMTGCGACCATCAGCGGTCTGAAAC
DE/a/4	CGAGCCAGGAATTCACTGTTCCGGGAWMTWMTWMTGCGACCATCAGCGGTCTGAAAC
FG/a/11	CATTACCGTGTACGCTGTARSCDVTxyzRRCxyzxyzxyzxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/a/10	CATTACCGTGTACGCTGTARSCDVTxyzRRCxyzxyzxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/a/9	CATTACCGTGTACGCTGTARSCDVTxyzRRCxyzxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/a/8	CATTACCGTGTACGCTGTARSCDVTxyzRRCxyzxyzTCAAACCCAATCAGCATCAATTATCGCAC
FG/a/7	CATTACCGTGTACGCTGTARSCDVTxyzRRCxyzTCAAACCCAATCAGCATCAATTATCGCAC

Table S2-3. Correlative parametric analysis of amino acid distributions - input matrices. Library design can be guided by information regarding each position's mutational tolerance and naturally evolved sequence to reduce the prevalence of overly destabilizing mutations as well as identifying structurally stabilizing mutations. Additionally, the chemical diversity found at the interfaces of well characterized natural binders, such as the complementarity determining regions (CDR) of antibodies, can be applied to protein scaffolds to accommodate for strong binding interactions. Here, a model for library design was built based on a linear combination of (A) computational stability, (B) natural homolog sequence frequency, and (C) CDR diversity input matrices. These three elements were weighted based on the (D) target exposure (i.e. proximity to the binding interface) and solvent exposed surface area (i.e. orientation and packing) of each site.

A. Computational sitewise assessment of changes in stability (ΔΔG, kcal/mol) upon mutation. Conducted in FoldX.

	D23	A24	P25	A26	V27	T28	V29	R30	Y31	G52	S53	K54	S55	T56
A	0.34	-	0.28	-	(0.31)	(0.30)	1.67	0.23	0.06	1.70	(0.09)	0.73	(0.23)	0.26
C	0.37	0.29	0.42	0.01	(0.24)	(0.21)	1.28	0.86	(0.15)	3.38	(0.22)	0.86	0.07	0.03
D	-	0.73	0.72	(0.08)	(0.03)	(0.34)	3.28	0.65	1.51	5.39	0.00	0.93	(0.00)	0.97
E	0.18	0.64	0.56	(0.25)	(0.25)	(0.48)	3.24	0.39	0.29	5.24	(0.48)	0.55	(0.02)	(0.13)
F	(0.22)	0.51	0.75	(0.50)	(0.16)	(0.44)	4.17	0.18	(0.33)	6.16	(0.28)	(0.06)	(0.36)	(0.16)
G	1.11	0.35	0.15	0.21	0.02	(0.33)	2.54	0.76	0.34	-	0.20	1.70	0.05	0.75
H	0.59	0.73	1.50	0.10	0.33	(0.15)	4.09	0.91	0.75	6.14	0.14	0.96	0.03	0.54
I	0.30	0.98	0.64	(0.07)	(0.07)	(0.01)	(0.31)	0.81	0.40	11.73	(0.40)	0.95	0.16	(0.03)
K	(0.04)	0.13	0.49	(0.37)	(0.38)	(0.57)	3.01	(0.00)	(0.46)	4.76	(0.73)	-	(0.56)	(0.56)
L	(0.24)	0.38	0.02	(0.47)	(0.56)	(0.48)	0.82	(0.06)	(0.47)	6.31	(0.71)	(0.31)	(0.38)	(0.25)
M	(0.19)	0.04	(0.47)	(0.57)	(0.58)	(0.53)	0.81	0.07	(0.84)	4.41	(0.78)	(0.32)	(0.45)	(0.91)
N	0.35	0.32	0.52	(0.07)	(0.24)	(0.30)	2.13	0.64	0.65	4.25	(0.04)	0.75	(0.54)	0.63
P	2.35	(0.59)	-	(0.08)	1.28	(0.06)	3.89	2.73	2.37	3.52	0.28	3.58	(0.08)	1.71
Q	0.18	0.40	0.49	(0.26)	(0.25)	(0.35)	3.04	0.11	0.06	5.13	(0.39)	0.49	(0.29)	(0.10)
R	0.12	0.16	0.74	(0.32)	(0.22)	(0.51)	4.14	-	(0.40)	5.01	(0.53)	(0.12)	(0.83)	(0.51)
S	0.62	0.24	0.19	0.19	(0.27)	(0.23)	2.43	0.79	0.03	3.21	-	0.84	-	(0.17)
T	0.71	0.73	0.61	0.37	(0.18)	-	1.16	1.12	1.16	8.01	0.32	1.23	0.70	-
V	0.62	0.82	0.71	0.20	-	0.12	-	0.90	0.86	8.15	0.03	1.25	0.40	0.07
W	0.17	1.03	1.87	(0.50)	0.69	(0.19)	7.81	0.49	0.45	7.10	(0.17)	0.03	(0.08)	(0.14)
Y	(0.12)	0.45	1.36	(0.45)	0.01	(0.39)	6.36	0.29	-	6.31	(0.27)	0.00	(0.29)	(0.10)

	T76	G77	R78	G79	D80	S81	P82	A83	S84	S85	K86
A	(0.07)	0.46	0.45	0.37	0.02	(0.04)	(0.70)	(0.15)	(0.34)	(0.06)	0.34
C	(0.07)	0.38	0.68	0.44	0.04	0.06	(0.65)	0.04	(0.20)	(0.13)	0.32
D	1.17	0.68	0.60	0.15	-	(0.03)	(0.61)	(0.01)	0.05	(0.15)	0.51
E	0.58	0.72	0.17	0.27	(0.15)	(0.12)	(0.79)	(0.04)	0.02	(0.12)	0.23
F	(0.69)	(0.14)	0.03	0.11	(0.24)	(0.30)	(1.13)	(0.30)	(0.36)	(0.23)	(0.36)
G	0.45	-	0.70	-	(0.39)	(0.51)	(0.60)	0.14	0.17	0.41	0.19
H	0.97	0.91	0.44	0.40	0.13	0.27	(0.58)	0.04	0.31	0.49	0.31
I	(0.98)	0.68	0.47	1.00	0.62	0.30	(0.92)	(0.14)	(0.47)	0.03	0.79
K	(0.11)	0.15	(0.02)	(0.15)	(0.44)	(0.29)	(1.08)	(0.38)	(0.38)	(0.17)	(0.06)
L	(1.16)	0.06	(0.09)	0.01	(0.27)	(0.25)	(1.03)	(0.39)	(0.95)	(0.49)	(0.10)
M	(1.37)	(0.27)	0.05	0.14	(0.29)	(0.23)	(1.06)	(0.45)	(1.28)	(0.63)	(0.33)
N	0.38	0.24	0.45	0.12	(0.14)	(0.09)	(0.66)	(0.10)	(0.08)	(0.17)	0.41
P	1.30	1.43	0.64	2.26	1.17	1.70	-	0.53	2.03	0.40	1.32
Q	0.18	0.57	0.03	0.20	(0.13)	(0.20)	(0.93)	(0.29)	(0.31)	(0.16)	0.14
R	0.37	0.18	-	0.09	(0.23)	(0.22)	(0.76)	(0.23)	(0.55)	(0.09)	(0.00)
S	0.26	0.34	0.57	0.41	(0.18)	-	(0.36)	0.00	0.02	(0.08)	0.38
T	-	0.74	0.69	0.84	0.40	0.33	(0.38)	0.10	0.02	0.01	0.86
V	(0.55)	0.99	0.64	0.98	0.54	0.43	(0.57)	0.04	(0.19)	0.04	0.92
W	1.10	0.52	0.45	0.32	(0.08)	0.07	(0.89)	(0.16)	(0.19)	0.29	(0.20)
Y	0.03	0.03	0.10	0.19	(0.17)	(0.22)	(1.07)	(0.24)	(0.18)	0.06	(0.23)

B. Natural sequence frequency for the tenth type III domain of human fibronectin: pfam database summary. Below the full amino acid table, the four most frequent residues at each site are listed.

	V	S	D	V	P	R	D	L	E	V	V	A	A	T	P	T	S	L	L
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	13.1%	12.6%	3.2%	28.6%	1.6%	4.5%	4.1%	2.1%	4.3%	15.0%	4.6%	11.2%	10.1%	2.0%	10.4%	1.7%	6.6%	11.2%	2.9%
C	0.5%	-	-	1.2%	-	-	-	-	-	2.2%	0.7%	0.6%	0.7%	-	0.6%	0.6%	1.0%	3.5%	0.5%
D	1.5%	0.8%	16.0%	-	0.7%	1.0%	13.4%	-	3.7%	0.9%	3.9%	15.8%	1.8%	4.3%	8.8%	13.9%	1.8%	-	3.1%
E	3.5%	1.1%	9.6%	0.8%	0.8%	6.9%	6.2%	-	13.3%	2.2%	5.6%	11.8%	3.1%	2.7%	8.6%	5.5%	5.1%	-	4.7%
F	0.8%	-	-	1.2%	-	1.6%	0.8%	5.0%	1.1%	5.1%	1.9%	0.7%	0.9%	0.9%	-	1.3%	1.2%	4.4%	2.5%
G	7.1%	22.2%	3.8%	3.0%	4.1%	8.8%	9.1%	-	1.6%	2.1%	2.7%	6.7%	4.7%	7.2%	5.0%	2.9%	4.5%	0.6%	1.2%
H	0.6%	-	1.6%	-	-	2.1%	1.9%	-	4.8%	-	1.5%	1.6%	1.1%	0.9%	2.3%	2.8%	1.9%	-	2.4%
I	2.7%	-	0.6%	7.4%	1.2%	2.1%	2.3%	11.8%	2.2%	11.7%	7.2%	1.4%	13.1%	0.8%	1.0%	1.0%	-	21.8%	4.3%
K	1.8%	3.6%	4.0%	1.9%	0.6%	8.5%	5.1%	-	11.5%	1.2%	4.7%	3.3%	3.9%	2.8%	11.7%	2.8%	3.3%	-	5.4%
L	13.2%	0.6%	1.3%	7.6%	0.8%	5.2%	1.4%	42.5%	1.5%	8.4%	6.5%	1.5%	7.8%	1.8%	1.1%	1.1%	-	19.6%	12.1%
M	3.9%	-	-	1.4%	-	1.2%	-	2.0%	0.9%	0.9%	1.4%	0.8%	1.3%	0.6%	0.9%	-	0.6%	5.6%	2.0%
N	0.8%	2.0%	4.3%	0.8%	1.2%	1.8%	25.6%	-	3.6%	1.0%	2.4%	11.1%	2.2%	8.5%	3.5%	11.8%	3.3%	-	3.9%
P	24.0%	14.3%	41.0%	11.1%	81.9%	1.8%	6.4%	8.7%	2.0%	2.0%	1.0%	4.5%	2.9%	0.6%	12.9%	-	-	-	0.6%
Q	2.1%	2.5%	5.1%	1.5%	0.9%	9.6%	3.4%	0.6%	9.2%	0.9%	2.3%	3.2%	2.4%	2.1%	3.3%	2.5%	3.5%	-	4.8%
R	1.9%	2.4%	2.7%	1.1%	-	14.0%	3.1%	-	11.1%	0.9%	3.6%	3.0%	3.7%	3.4%	4.9%	5.2%	2.7%	-	7.9%
S	4.3%	28.4%	3.1%	3.1%	2.0%	10.4%	9.0%	0.8%	9.2%	3.1%	14.0%	13.1%	6.8%	22.1%	16.2%	14.8%	45.8%	-	9.8%
T	4.4%	7.3%	1.5%	5.2%	1.5%	12.4%	5.4%	1.7%	14.0%	5.2%	18.7%	5.2%	7.4%	35.6%	4.9%	27.6%	13.4%	0.9%	21.5%
V	11.6%	0.7%	1.4%	22.8%	1.1%	5.7%	1.5%	21.7%	4.6%	34.8%	14.7%	3.3%	23.4%	1.5%	1.4%	2.6%	1.3%	28.7%	7.8%
W	-	-	-	-	-	-	-	-	-	0.7%	-	0.7%	1.3%	0.9%	0.8%	-	1.1%	1.0%	1.2%
Y	1.9%	-	-	0.6%	-	1.6%	-	-	0.7%	1.2%	2.2%	0.5%	1.1%	0.8%	1.2%	1.1%	1.9%	-	1.4%
Most Frequent																			
1	P (0.24)	S (0.28)	P (0.41)	A (0.29)	P (0.82)	R (0.14)	N (0.26)	L (0.43)	T (0.14)	V (0.35)	T (0.19)	D (0.16)	V (0.23)	T (0.36)	S (0.16)	T (0.28)	S (0.46)	V (0.29)	T (0.21)
2	L (0.13)	G (0.22)	D (0.16)	V (0.23)	G (0.04)	T (0.12)	D (0.13)	V (0.22)	E (0.13)	A (0.15)	V (0.15)	S (0.13)	I (0.13)	S (0.22)	P (0.13)	S (0.15)	T (0.13)	I (0.22)	L (0.12)
3	A (0.13)	P (0.14)	E (0.1)	P (0.11)	S (0.02)	S (0.1)	G (0.09)	I (0.12)	K (0.12)	I (0.12)	S (0.14)	E (0.12)	A (0.1)	N (0.08)	K (0.12)	D (0.14)	A (0.07)	L (0.2)	S (0.1)
4	V (0.12)	A (0.13)	Q (0.05)	L (0.08)	A (0.02)	Q (0.1)	S (0.09)	P (0.09)	R (0.11)	L (0.08)	I (0.07)	A (0.11)	L (0.08)	G (0.07)	A (0.1)	N (0.12)	E (0.05)	A (0.11)	R (0.08)

	I	S	W	D	A	P	A	V	T	V	R	Y	Y	R	I	T	Y	G	E
	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
A	1.9%	6.8%	-	2.6%	18.2%	2.2%	15.0%	9.2%	4.5%	6.9%	2.0%	3.9%	3.1%	2.1%	1.4%	2.9%	1.3%	4.8%	3.0%
C	-	1.0%	-	-	-	-	-	1.8%	-	0.7%	-	0.7%	1.0%	-	-	0.8%	1.8%	1.8%	0.7%
D	-	1.3%	-	16.5%	2.1%	3.7%	4.4%	6.4%	17.5%	6.2%	10.9%	2.2%	2.2%	4.2%	-	3.0%	0.6%	2.5%	7.2%
E	-	5.8%	0.5%	15.2%	6.2%	1.6%	9.4%	4.8%	5.6%	2.1%	6.3%	4.4%	2.9%	9.6%	0.6%	13.6%	1.6%	2.3%	23.9%
F	2.1%	0.6%	1.6%	-	-	-	-	3.0%	0.6%	4.0%	-	2.9%	8.9%	1.2%	1.0%	2.0%	6.0%	1.6%	0.9%
G	-	2.4%	0.5%	3.0%	4.2%	3.1%	13.5%	11.1%	4.8%	11.5%	18.1%	30.7%	4.5%	2.9%	1.3%	5.9%	1.1%	13.5%	-
H	-	2.7%	-	1.8%	1.5%	0.7%	1.2%	2.1%	1.6%	0.7%	1.9%	5.3%	1.4%	1.7%	-	1.5%	2.2%	2.4%	1.6%
I	16.1%	1.0%	-	0.8%	2.3%	-	2.2%	6.5%	1.9%	24.3%	1.7%	3.0%	3.7%	10.9%	37.3%	2.7%	3.4%	5.1%	4.0%
K	-	6.3%	-	8.9%	8.5%	1.6%	4.9%	3.0%	3.3%	0.6%	6.3%	3.4%	1.6%	9.5%	0.7%	4.6%	3.4%	8.8%	7.3%
L	44.5%	1.3%	0.7%	2.1%	2.6%	0.8%	4.1%	8.4%	1.8%	6.8%	5.1%	3.5%	2.3%	6.0%	17.6%	5.1%	5.4%	3.2%	4.4%
M	1.3%	1.3%	-	0.6%	1.1%	-	0.6%	1.6%	-	0.9%	0.5%	1.0%	-	1.0%	1.1%	0.9%	1.8%	1.3%	0.7%
N	-	4.7%	-	8.4%	2.2%	2.4%	5.3%	5.1%	12.1%	3.2%	5.4%	6.6%	1.7%	3.9%	0.5%	4.1%	0.9%	5.0%	2.2%
P	0.7%	1.1%	0.7%	3.7%	22.6%	73.0%	6.8%	5.6%	9.9%	3.0%	1.3%	1.2%	3.5%	3.7%	0.6%	0.8%	-	1.4%	8.5%
Q	-	6.5%	-	7.8%	2.5%	1.3%	4.3%	3.1%	3.7%	0.9%	6.8%	3.1%	1.2%	6.1%	0.6%	7.9%	2.1%	4.9%	5.3%
R	-	6.0%	-	3.8%	5.1%	0.6%	3.0%	2.3%	2.6%	0.6%	10.6%	3.1%	0.8%	15.5%	-	5.7%	4.5%	12.2%	7.1%
S	-	31.6%	-	9.6%	8.4%	4.4%	13.9%	8.0%	14.7%	2.5%	7.4%	11.5%	6.5%	3.7%	0.6%	10.4%	0.9%	9.1%	3.8%
T	0.6%	16.1%	-	12.7%	4.8%	1.4%	4.8%	4.0%	11.7%	2.7%	12.3%	1.9%	1.4%	5.8%	1.9%	17.6%	0.8%	5.5%	3.2%
V	29.7%	2.3%	-	1.4%	6.6%	1.9%	4.9%	8.4%	2.1%	20.4%	2.0%	2.2%	3.3%	9.2%	31.6%	5.6%	4.6%	6.7%	8.2%
W	0.9%	-	92.4%	-	-	-	-	-	-	-	-	0.9%	1.0%	0.6%	-	0.5%	4.8%	2.0%	1.0%
Y	-	0.8%	0.8%	-	-	-	0.8%	5.2%	0.8%	1.4%	0.5%	8.4%	48.4%	2.4%	1.8%	4.3%	52.4%	5.7%	6.4%
Most Frequent																			
1	L (0.45)	S (0.32)	W (0.92)	D (0.17)	P (0.23)	P (0.73)	A (0.15)	G (0.11)	D (0.17)	I (0.24)	G (0.18)	G (0.31)	Y (0.48)	R (0.15)	I (0.37)	T (0.18)	Y (0.52)	G (0.14)	E (0.24)
2	V (0.3)	T (0.16)	F (0.02)	E (0.15)	A (0.18)	S (0.04)	S (0.14)	A (0.09)	S (0.15)	V (0.2)	T (0.12)	S (0.12)	F (0.09)	I (0.11)	V (0.32)	E (0.14)	F (0.06)	R (0.12)	P (0.09)
3	I (0.16)	A (0.07)	Y (0.01)	T (0.13)	K (0.09)	D (0.04)	G (0.14)	V (0.08)	N (0.12)	G (0.11)	D (0.11)	Y (0.08)	S (0.06)	E (0.1)	L (0.18)	S (0.1)	L (0.05)	S (0.09)	V (0.08)
4	F (0.02)	Q (0.07)	P (0.01)	S (0.1)	S (0.08)	G (0.03)	E (0.09)	L (0.08)	T (0.12)	A (0.07)	R (0.11)	N (0.07)	G (0.04)	K (0.1)	T (0.02)	Q (0.08)	W (0.05)	K (0.09)	K (0.07)

	T	G	G	N	S	P	V	Q	E	F	T	V	P	G	S	K	S	T	A
	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
A	6.3%	5.8%	6.1%	2.8%	6.5%	4.1%	6.5%	4.3%	7.4%	3.9%	4.9%	6.3%	5.4%	8.6%	6.5%	5.7%	4.3%	2.6%	16.7%
C	1.6%	1.5%	1.1%	0.6%	1.5%	-	1.0%	0.6%	1.3%	1.7%	1.5%	1.7%	0.7%	2.0%	1.2%	-	2.8%	1.2%	4.4%
D	3.8%	9.9%	7.5%	13.3%	6.2%	6.8%	2.2%	4.1%	8.0%	1.5%	4.9%	1.6%	7.6%	5.3%	7.8%	6.1%	4.8%	2.6%	1.4%
E	7.7%	5.0%	6.7%	12.4%	6.7%	11.4%	4.2%	8.7%	23.7%	2.8%	5.4%	2.8%	7.7%	3.1%	5.3%	12.6%	3.6%	9.0%	3.1%
F	1.6%	1.3%	1.1%	1.2%	1.0%	1.4%	3.3%	0.6%	0.5%	10.6%	0.5%	1.6%	0.6%	0.8%	0.8%	0.8%	1.6%	3.0%	6.4%
G	0.7%	19.6%	27.6%	6.1%	9.7%	5.4%	2.6%	4.9%	2.9%	2.4%	3.1%	1.5%	3.6%	32.9%	3.3%	2.2%	2.9%	1.6%	2.8%
H	0.9%	1.5%	1.2%	2.6%	1.5%	1.4%	0.6%	3.5%	2.1%	2.0%	1.4%	0.8%	1.7%	1.8%	2.6%	2.3%	1.8%	2.1%	4.2%
I	6.8%	3.3%	1.6%	1.8%	1.4%	2.2%	11.9%	1.6%	1.4%	8.4%	2.8%	12.1%	1.7%	2.8%	2.1%	2.2%	1.5%	2.3%	6.2%
K	8.4%	8.4%	5.6%	8.0%	7.4%	6.0%	2.7%	13.3%	7.2%	2.9%	5.2%	2.6%	5.9%	3.9%	4.2%	15.5%	2.6%	4.0%	3.4%
L	7.3%	4.0%	2.9%	2.8%	2.9%	2.8%	10.0%	4.2%	1.5%	11.7%	3.7%	11.2%	2.2%	2.2%	2.7%	3.9%	5.2%	3.6%	5.1%
M	2.1%	1.4%	1.2%	1.3%	0.8%	0.7%	3.1%	1.5%	0.9%	2.9%	1.2%	1.7%	0.7%	0.6%	0.6%	1.0%	1.3%	1.2%	2.0%
N	2.8%	6.6%	2.8%	14.5%	3.6%	2.7%	2.0%	4.7%	6.3%	3.4%	6.4%	1.9%	5.0%	5.7%	9.2%	4.8%	8.9%	5.1%	2.1%
P	2.9%	2.1%	2.7%	2.7%	1.3%	22.7%	0.8%	2.0%	2.2%	0.9%	3.6%	0.7%	28.2%	3.3%	3.8%	2.3%	3.0%	2.7%	2.1%
Q	2.6%	4.5%	3.5%	4.2%	5.2%	4.2%	1.2%	16.5%	3.9%	2.0%	4.6%	1.4%	3.2%	1.6%	3.0%	9.5%	3.3%	4.1%	2.5%
R	5.5%	5.3%	7.4%	5.9%	6.9%	4.1%	4.2%	9.3%	6.4%	3.6%	6.7%	2.4%	3.3%	3.3%	4.0%	10.4%	5.6%	5.5%	2.1%
S	9.2%	8.2%	9.1%	6.2%	18.8%	6.9%	4.3%	7.1%	6.7%	3.3%	12.6%	3.3%	8.4%	8.1%	18.9%	4.8%	17.1%	17.2%	4.1%
T	14.9%	3.3%	5.2%	6.5%	10.8%	7.1%	8.9%	5.2%	8.1%	4.0%	23.6%	6.9%	7.5%	5.8%	16.9%	7.7%	24.0%	23.3%	5.5%
V	11.8%	4.1%	2.7%	2.4%	3.7%	2.1%	21.5%	2.2%	5.2%	14.5%	5.9%	36.2%	5.3%	5.5%	4.7%	5.0%	3.3%	4.5%	11.5%
W	0.6%	1.1%	1.4%	1.3%	1.7%	6.2%	5.5%	3.1%	2.9%	8.9%	1.0%	1.1%	-	1.9%	1.0%	1.9%	0.6%	2.1%	1.1%
Y	2.5%	3.3%	2.5%	3.4%	2.5%	1.5%	3.5%	2.4%	1.3%	8.7%	1.0%	2.2%	0.8%	0.8%	1.3%	0.9%	1.8%	2.3%	13.2%
Most Frequent																			
1	T (0.15)	G (0.2)	G (0.28)	N (0.15)	S (0.19)	P (0.23)	V (0.21)	Q (0.16)	E (0.24)	V (0.14)	T (0.24)	V (0.36)	P (0.28)	G (0.33)	S (0.19)	K (0.15)	T (0.24)	T (0.23)	A (0.17)
2	V (0.12)	D (0.1)	S (0.09)	D (0.13)	T (0.11)	E (0.11)	I (0.12)	K (0.13)	T (0.08)	L (0.12)	S (0.13)	I (0.12)	S (0.08)	A (0.09)	T (0.17)	E (0.13)	S (0.17)	S (0.17)	Y (0.13)
3	S (0.09)	K (0.08)	D (0.07)	E (0.12)	G (0.1)	T (0.07)	L (0.1)	R (0.09)	D (0.08)	F (0.11)	R (0.07)	L (0.11)	E (0.08)	S (0.08)	N (0.09)	R (0.1)	N (0.09)	E (0.09)	V (0.12)
4	K (0.08)	S (0.08)	R (0.07)	K (0.08)	K (0.07)	S (0.07)	T (0.09)	E (0.09)	A (0.07)	W (0.09)	N (0.06)	T (0.07)	D (0.08)	T (0.06)	D (0.08)	Q (0.1)	R (0.06)	R (0.06)	F (0.06)

	T	I	S	G	L	K	P	G	V	D	Y	T	I	T	V	Y	A	V	T
	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
A	1.9%	3.0%	3.3%	3.3%	1.0%	3.5%	6.7%	6.1%	5.3%	3.8%	1.7%	2.7%	3.0%	2.1%	2.1%	3.2%	45.5%	2.2%	2.7%
C	0.8%	0.9%	0.5%	-	-	-	0.7%	-	1.6%	-	-	0.7%	0.9%	1.1%	-	1.1%	1.1%	1.5%	1.1%
D	4.0%	2.9%	6.3%	11.7%	0.9%	3.6%	2.4%	7.2%	-	11.5%	-	3.3%	1.0%	2.1%	1.3%	0.9%	1.8%	0.5%	2.2%
E	5.9%	1.0%	7.3%	4.1%	1.0%	10.9%	9.9%	3.2%	3.1%	21.7%	1.9%	11.0%	1.3%	6.4%	3.3%	3.1%	6.0%	3.8%	5.9%
F	1.1%	6.1%	1.1%	0.8%	3.2%	1.1%	1.4%	3.0%	1.1%	0.5%	4.4%	1.3%	24.9%	0.8%	0.8%	11.4%	-	8.2%	3.2%
G	0.6%	-	3.1%	38.1%	-	0.9%	1.5%	41.5%	0.8%	1.6%	1.4%	2.6%	1.8%	2.5%	2.8%	1.2%	5.5%	0.9%	1.1%
H	2.0%	-	2.5%	2.6%	-	2.3%	0.9%	2.1%	2.2%	2.0%	2.9%	1.8%	-	1.8%	0.5%	2.9%	1.0%	1.2%	1.1%
I	4.3%	22.4%	0.9%	0.7%	4.5%	5.1%	1.9%	0.5%	4.2%	1.1%	1.2%	2.5%	18.2%	0.9%	12.5%	3.5%	1.1%	13.3%	2.0%
K	6.7%	0.9%	7.4%	4.3%	0.5%	16.6%	3.9%	2.4%	4.5%	7.4%	1.8%	8.4%	0.9%	6.1%	1.6%	5.5%	3.9%	1.1%	4.3%
L	5.7%	21.9%	1.7%	1.2%	71.2%	5.8%	1.9%	3.0%	11.6%	4.8%	4.2%	6.0%	13.9%	2.8%	15.3%	8.2%	2.2%	12.6%	2.7%
M	1.3%	2.0%	0.8%	0.5%	1.8%	1.6%	0.6%	0.6%	2.1%	0.8%	0.8%	0.9%	2.8%	0.7%	1.6%	1.8%	0.5%	1.9%	0.9%
N	4.6%	1.1%	5.3%	13.6%	0.7%	4.0%	3.0%	10.9%	2.7%	6.6%	1.2%	6.9%	0.9%	4.9%	2.4%	1.7%	2.6%	1.1%	19.9%
P	2.5%	0.8%	4.2%	1.9%	0.5%	1.4%	45.8%	-	-	2.7%	1.0%	1.9%	0.6%	1.2%	1.4%	1.1%	3.9%	1.0%	1.7%
Q	3.5%	1.5%	4.7%	2.7%	0.7%	8.3%	2.6%	1.6%	3.2%	6.2%	0.9%	6.4%	-	5.4%	0.9%	5.2%	4.1%	0.9%	4.1%
R	6.5%	0.8%	5.2%	3.0%	1.3%	9.5%	2.1%	2.3%	5.3%	4.5%	2.2%	5.7%	-	21.1%	1.1%	8.7%	4.8%	0.9%	5.4%
S	6.1%	0.8%	17.6%	5.4%	0.8%	7.9%	5.0%	7.5%	5.7%	9.2%	1.1%	10.5%	0.7%	11.7%	1.0%	4.7%	7.5%	1.6%	10.1%
T	31.6%	2.3%	23.2%	2.4%	3.3%	10.4%	6.0%	2.6%	31.5%	11.6%	1.4%	20.9%	1.5%	22.4%	1.9%	4.9%	5.0%	3.0%	24.3%
V	9.4%	28.2%	2.4%	1.3%	6.3%	5.3%	2.3%	0.8%	13.1%	2.2%	2.0%	3.9%	21.3%	1.5%	46.6%	4.4%	1.9%	34.7%	2.9%
W	-	-	0.7%	-	-	-	-	0.7%	-	0.6%	0.5%	-	-	1.1%	-	2.0%	-	0.8%	0.5%
Y	1.1%	2.2%	1.9%	1.7%	0.9%	1.4%	1.1%	3.1%	0.7%	-	68.5%	2.0%	4.5%	3.3%	2.0%	24.5%	0.8%	8.7%	3.9%
Most Frequent																			
1	T (0.32)	V (0.28)	T (0.23)	G (0.38)	L (0.71)	K (0.17)	P (0.46)	G (0.41)	T (0.31)	E (0.22)	Y (0.69)	T (0.21)	F (0.25)	T (0.22)	V (0.47)	Y (0.25)	A (0.46)	V (0.35)	T (0.24)
2	V (0.09)	I (0.22)	S (0.18)	N (0.14)	V (0.06)	E (0.11)	E (0.1)	N (0.11)	V (0.13)	T (0.12)	F (0.04)	E (0.11)	V (0.21)	R (0.21)	L (0.15)	F (0.11)	S (0.07)	I (0.13)	N (0.2)
3	K (0.07)	L (0.22)	K (0.07)	D (0.12)	I (0.04)	T (0.1)	A (0.07)	S (0.08)	L (0.12)	D (0.12)	L (0.04)	S (0.1)	I (0.18)	S (0.12)	I (0.13)	R (0.09)	E (0.06)	L (0.13)	S (0.1)
4	R (0.06)	F (0.06)	E (0.07)	S (0.05)	T (0.03)	R (0.09)	T (0.06)	D (0.07)	S (0.06)	S (0.09)	H (0.03)	K (0.08)	L (0.14)	E (0.06)	E (0.03)	L (0.08)	G (0.05)	Y (0.09)	E (0.06)

	G	R	G	D	S	P	A	S	S	K	P	I	S	I	N	Y	R	T
	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
A	18.6%	6.7%	8.8%	1.1%	8.1%	5.6%	42.2%	10.0%	6.6%	2.6%	8.7%	11.4%	7.0%	6.1%	4.8%	4.2%	8.4%	9.4%
C	0.7%	-	1.1%	0.7%	2.0%	0.8%	1.1%	2.6%	1.0%	-	-	1.6%	0.7%	1.0%	-	-	-	-
D	2.6%	0.6%	0.6%	22.8%	0.7%	1.3%	-	1.7%	4.3%	1.3%	2.3%	0.7%	1.2%	3.4%	8.8%	2.2%	2.1%	1.0%
E	5.3%	7.6%	1.5%	17.5%	0.8%	5.3%	2.7%	10.2%	2.4%	19.6%	4.3%	1.2%	1.3%	5.3%	7.8%	10.5%	14.6%	1.8%
F	4.5%	2.0%	4.6%	3.5%	3.7%	2.2%	2.6%	2.5%	1.7%	0.5%	-	7.1%	-	4.0%	-	14.1%	2.5%	0.6%
G	14.6%	-	49.6%	1.1%	9.1%	2.5%	8.6%	13.9%	23.6%	7.0%	8.3%	4.0%	32.0%	6.1%	36.5%	4.7%	2.3%	2.8%
H	1.7%	3.0%	0.8%	1.1%	0.8%	1.2%	-	1.9%	0.6%	2.2%	0.8%	0.5%	-	-	0.6%	3.4%	1.0%	-
I	2.9%	2.3%	2.1%	4.3%	3.4%	6.5%	5.8%	1.8%	1.3%	0.7%	2.4%	13.2%	-	14.8%	-	1.3%	-	1.1%
K	4.9%	13.9%	0.8%	2.8%	3.9%	5.6%	1.0%	3.4%	1.0%	21.3%	1.7%	2.0%	0.8%	1.3%	2.7%	4.1%	9.0%	1.0%
L	2.8%	3.0%	2.3%	4.0%	2.5%	4.0%	6.7%	2.7%	1.2%	1.9%	0.6%	14.2%	-	15.4%	0.7%	4.4%	1.6%	2.1%
M	1.5%	1.1%	1.2%	0.6%	1.0%	3.2%	2.3%	0.8%	-	0.6%	-	3.7%	-	2.5%	-	1.2%	-	0.5%
N	7.6%	1.7%	5.5%	13.4%	6.4%	1.7%	1.4%	5.0%	15.6%	10.2%	3.3%	1.1%	2.5%	0.9%	11.5%	1.1%	1.0%	0.7%
P	3.4%	-	1.0%	-	-	23.6%	-	-	-	0.5%	55.1%	0.6%	-	2.0%	3.5%	11.1%	19.0%	6.4%
Q	3.3%	5.2%	1.4%	4.7%	3.0%	5.4%	1.6%	4.9%	2.0%	9.7%	1.8%	1.2%	1.2%	1.4%	2.6%	2.2%	5.4%	0.8%
R	1.9%	31.2%	0.8%	5.7%	7.4%	8.2%	2.6%	5.7%	1.4%	12.1%	1.3%	0.8%	1.0%	1.3%	1.4%	1.9%	16.3%	0.5%
S	7.6%	1.7%	3.5%	1.7%	19.4%	2.0%	1.2%	15.8%	21.7%	2.8%	2.9%	0.7%	45.9%	2.5%	14.9%	5.0%	7.6%	55.5%
T	3.0%	2.5%	3.7%	3.0%	11.4%	4.3%	1.6%	7.7%	7.9%	3.0%	3.0%	2.6%	4.5%	4.2%	2.8%	1.6%	3.0%	13.0%
V	6.8%	5.4%	3.1%	7.2%	12.0%	12.7%	13.6%	4.9%	4.7%	1.3%	1.8%	24.1%	-	22.6%	-	3.5%	0.8%	1.4%
W	0.6%	-	0.5%	-	-	-	0.7%	-	-	-	-	1.4%	-	2.1%	-	9.5%	2.3%	-
Y	5.7%	10.9%	7.1%	3.7%	4.3%	3.3%	3.2%	4.0%	2.1%	2.1%	0.6%	7.8%	-	2.6%	-	13.6%	1.8%	-
Most Frequent																		
1	A (0.19)	R (0.31)	G (0.5)	D (0.23)	S (0.19)	P (0.24)	A (0.42)	S (0.16)	G (0.24)	K (0.21)	P (0.55)	V (0.24)	S (0.46)	V (0.23)	G (0.37)	F (0.14)	P (0.19)	S (0.55)
2	G (0.15)	K (0.14)	A (0.09)	E (0.18)	V (0.12)	V (0.13)	V (0.14)	G (0.14)	S (0.22)	E (0.2)	A (0.09)	L (0.14)	G (0.32)	L (0.15)	S (0.15)	Y (0.14)	R (0.16)	T (0.13)
3	N (0.08)	Y (0.11)	Y (0.07)	N (0.13)	T (0.11)	R (0.08)	G (0.09)	E (0.1)	N (0.16)	R (0.12)	G (0.08)	I (0.13)	A (0.07)	I (0.15)	N (0.11)	P (0.11)	E (0.15)	A (0.09)
4	S (0.08)	E (0.08)	N (0.06)	V (0.07)	G (0.09)	I (0.07)	L (0.07)	A (0.1)	T (0.08)	N (0.1)	E (0.04)	A (0.11)	T (0.04)	A (0.06)	D (0.09)	E (0.11)	K (0.09)	P (0.06)

C. Complementarity distribution mimicking CDRH3 of antibodies (CDR'), modified from previous work (Hackel et al., 2010).

CDR' amino acid frequency distribution:

A	0.06
C	0.05
D	0.11
E	0.02
F	0.03
G	0.04
H	0.06
I	0.02
K	0.02
L	0.02
M	0.00
N	0.09
P	0.04
Q	0.01
R	0.03
S	0.13
T	0.05
V	0.03
W	0.01
Y	0.17
Z	0.03

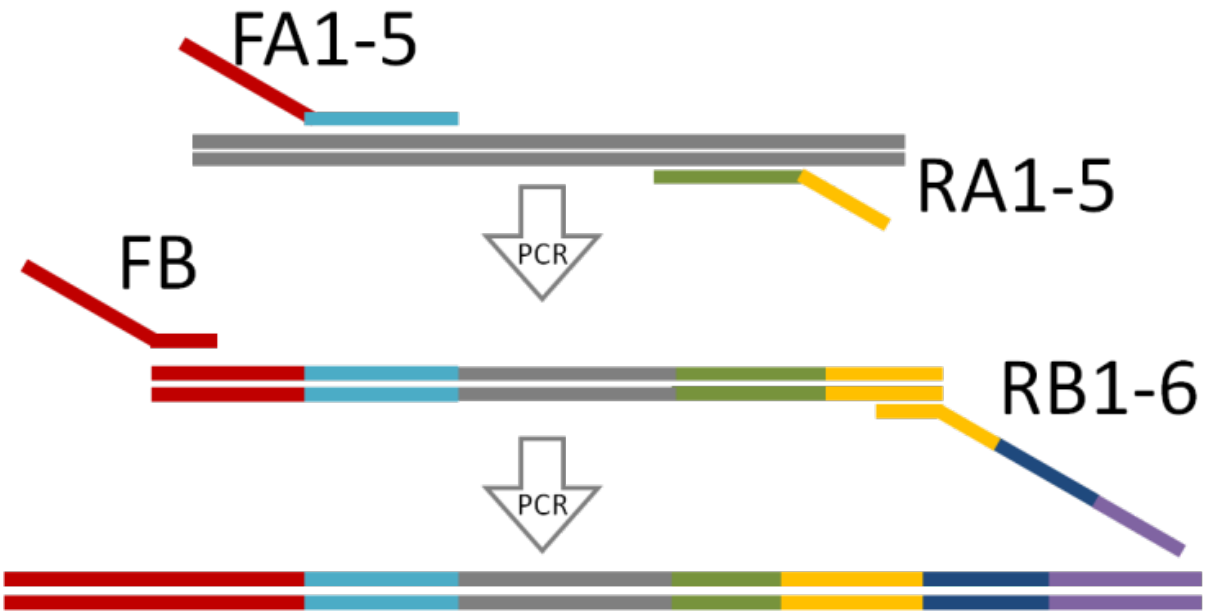
Nucleotide composition of codon design:

A1	C1	G1	T1	A2	C2	G2	T2	A3	C3	G3	T3
0.20	0.15	0.25	0.40	0.50	0.25	0.15	0.10	0.00	0.45	0.10	0.45

D. Target exposure and solvent exposed surface area (SASA) for the BC, DE, and FG loops of fibronectin.

Position	Target Exposure	SASA
D23	0.68	0.35
A24	0.90	0.32
P25	0.70	0.10
A26	0.79	0.75
V27	0.69	0.57
T28	0.84	0.75
V29	0.46	0.03
R30	0.90	0.49
Y31	0.37	0.43
G52	0.54	0.12
S53	0.94	0.83
K54	0.68	0.64
S55	0.85	0.41
T56	0.93	0.48
T76	0.39	0.08
G77	0.60	0.48
R78	0.79	0.81
G79	0.72	0.77
D80	0.70	0.74
S81	0.54	0.69
P82	0.59	0.76
A83	0.76	0.81
S84	0.52	0.54
S85	0.61	0.14
K86	0.58	0.88

Table S2-4. Illumina primer design. Conserved framework positions with regions or sites having low diversity, as is the case with conserved framework positions, require additional considerations during sample preparation to ensure a high level of accuracy during the MiSeq run. The inclusion of variable length degenerate sequence (N_{4-8}) at 5' and 3' ends allow the conserved sites to be offset. Based on TruSeq guidelines, adapter indices are designed to have balanced G/T and C/A content, the following 6 adapter index tags were selected: AD005,6,12,14,18,19. Schematic below demonstrates the two-step PCR used for amplicon library construction. Colored regions of schematic indicate (from left to right) TruSeq universal adapter (red), target primer (cyan), gene of interest (gray), reverse target primer (green), multiplex primer 2.0 (yellow), Illumina index (blue), and Illumina PCR primer (purple). Table at bottom lists individual sequences with item names corresponding to PCR schematic.



Item Name	Length	Sequence
RB1	59	caagcagaagacggcatacgagat**CACTGT**gtgactggagttcagacgtgtgctcttcc
RB2	59	caagcagaagacggcatacgagat**ATTGGC**gtgactggagttcagacgtgtgctcttcc
RB3	59	caagcagaagacggcatacgagat**TACAAG**gtgactggagttcagacgtgtgctcttcc
RB4	59	caagcagaagacggcatacgagat**TTTCAC**gtgactggagttcagacgtgtgctcttcc
RB5	59	caagcagaagacggcatacgagat**GGAACT**gtgactggagttcagacgtgtgctcttcc
RB6	59	caagcagaagacggcatacgagat**GCGGAC**gtgactggagttcagacgtgtgctcttcc
FB	51	aatgatacggcgaccaccgagatctacactctttccctacacgacgctctt
RA1	49	GTTCAGACGTGTGCTCTTCCGATCTNNNNATAAGCTTTTGTTCGGATCC
RA2	50	GTTCAGACGTGTGCTCTTCCGATCTNNNNNATAAGCTTTTGTTCGGATCC
RA3	51	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNATAAGCTTTTGTTCGGATCC
RA4	52	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNATAAGCTTTTGTTCGGATCC
RA5	53	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNATAAGCTTTTGTTCGGATCC
FA1	51	TTTCCCTACACGACGCTCTTCCGATCTNNNNACTACGCTCTGCAGGCTAGT
FA2	52	TTTCCCTACACGACGCTCTTCCGATCTNNNNNACTACGCTCTGCAGGCTAGT
FA3	53	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNACTACGCTCTGCAGGCTAGT
FA4	54	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNACTACGCTCTGCAGGCTAGT
FA5	55	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNACTACGCTCTGCAGGCTAGT

Chapter 3 – ScaffoldSeq: Software for characterization of directed evolution populations

Adapted from: Daniel R. Woldring, Patrick V. Holec and Benjamin J. Hackel. "ScaffoldSeq: Software for characterization of directed evolution populations." Proteins: Structure, Function, and Bioinformatics, 2016. Vol 84(7), 869-874.

3.1. Synopsis

ScaffoldSeq is software designed for the numerous applications – including directed evolution analysis – in which a user generates a population of DNA sequences encoding for partially diverse proteins with related functions and would like to characterize the single site and pairwise amino acid frequencies across the population. A common scenario for enzyme maturation, antibody screening, and alternative scaffold engineering involves naïve and evolved populations that contain diversified regions, varying in both sequence and length, within a conserved framework. Analyzing the diversified regions of such populations is facilitated by high-throughput sequencing platforms; however, length variability within these regions (e.g. antibody CDRs) encumbers the alignment process. To overcome this challenge, the ScaffoldSeq algorithm takes advantage of conserved framework sequences to quickly identify diverse regions. Beyond this, unintended biases in sequence frequency are generated throughout the experimental workflow required to evolve and isolate clones of interest prior to DNA sequencing. ScaffoldSeq software uniquely handles this issue by providing tools to quantify and remove background sequences, cluster similar protein families, and dampen the impact of dominant clones. The software produces graphical and tabular summaries for each region of interest, allowing users to evaluate diversity in a site-specific manner as well as identify epistatic pairwise

interactions. The code and detailed information are freely available at http://research.cems.umn.edu/hackel.

3.2. Introduction

Sequence analysis of diverse protein populations with related functions is valuable in characterizing antibody[58,146–148] repertoires, evaluating homologs (such as for consensus design[35]), guiding combinatorial library design for *de novo* protein discovery[61], and performing deep mutational scanning[149,150] to elucidate evolution, *e.g.* of stability, binding, or catalysis. Sequence analysis can identify – on a sitewise or multi-site motif level – amino acid frequencies that are consistent with the discovery and evolution of stable, functional molecules. These amino acid frequencies can be implemented combinatorially in libraries or precisely in clones. The increased availability of broad data sets of functionally homologous, partially diverse proteins mirrors the growth in deep sequencing and bioinformatics mining. Realizing benefit from these advances requires techniques and software for efficient, accurate, consistent analysis throughout and across fields.

Here we discuss software to analyze diverse protein populations for such purposes. As a particular type of example, we highlight the analysis of populations of small protein scaffolds (fibronectin type III[61] and Gp2 domains[151]) in which three or two regions, respectively, were highly diversified and evolved for various binding functions. The software input is DNA sequences (FASTA/FASTQ); for example, a population encoding for protein domains engineered to bind various epitopes on an antigen. The primary outputs are sitewise amino acid frequency matrices, pairwise epistasis analyses – including epistatic frequency distributions and identification of key positive and negative correlates – and metrics of sequence diversity. The analysis workflow differentiates itself from

existing tools and methods of others[152–159] by being customizable, via a dynamic, easily-navigated user interface, and allows removal of background sequences (*e.g.*, non-functional clones unintentionally isolated during a protein library screen), evaluation and clustering of highly similar sequences, and dominant clone weight dampening. Output files are generated as graphical summaries and in comma-separated value format to facilitate downstream application and project-specific data interrogation (Fig. 1). Along with the annotated script, an accompanying *Software Walkthrough* provides a detailed guide for users as exemplified by Gp2 directed evolution analysis. Beyond the ligand-specific scientific value of sitewise and pairwise amino acid frequency data, analysis of the affibody[160], fibronectin and Gp2 ligand evolution data reveals the benefits of variable dampening of dominant clones.

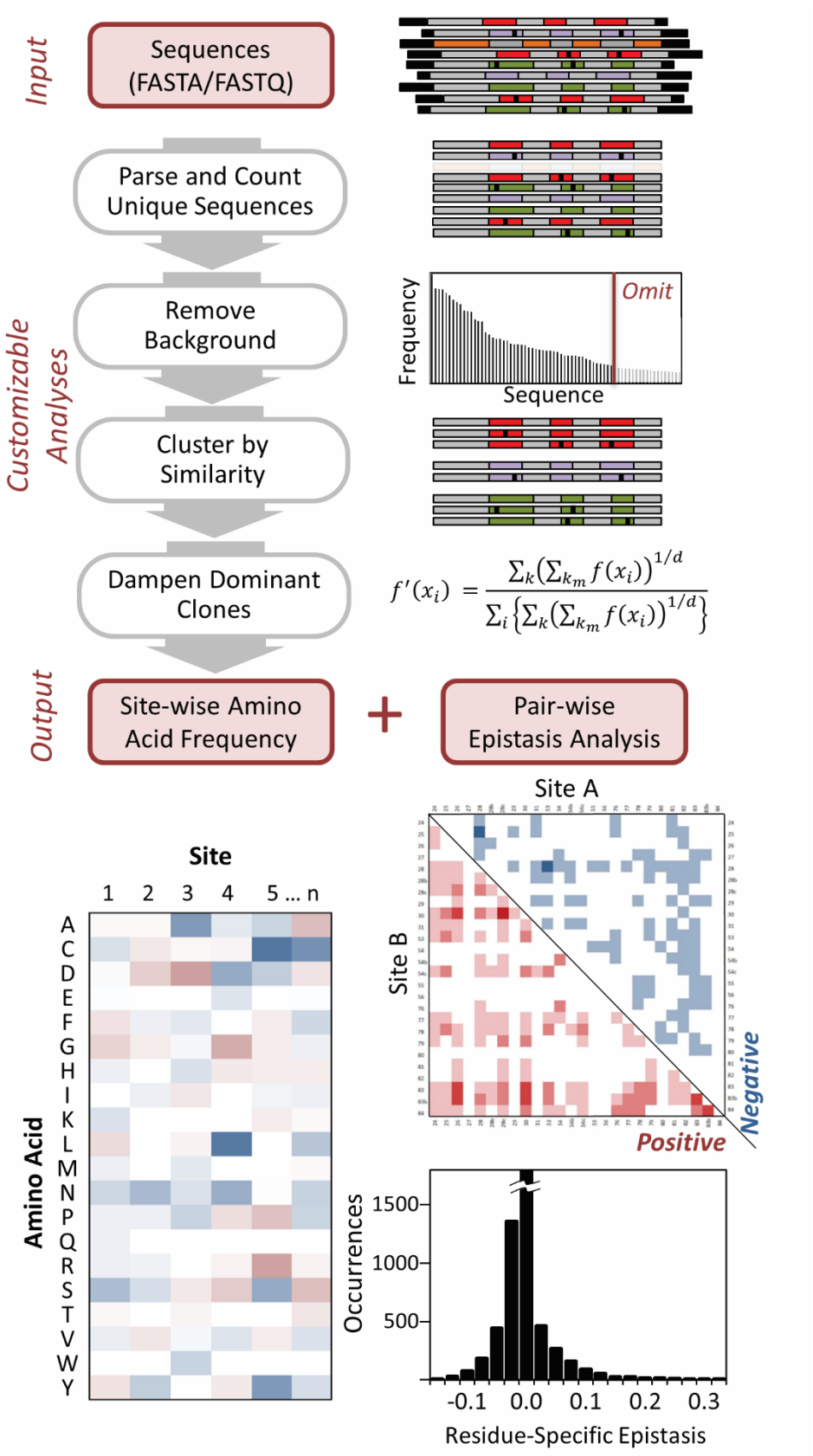

Input
Sequences (FASTA/FASTQ)
Parse and Count Unique Sequences
Customizable Analyses
Remove Background
Cluster by Similarity
Dampen Dominant Clones
Frequency
Omit
Sequence
$f'(x_i) = \frac{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}}{\sum_i \left\{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}\right\}}$
Output
Site-wise Amino Acid Frequency
+
Pair-wise Epistasis Analysis
Site
1 2 3 4 5 ... n
Amino Acid
A C D E F G H I K L M N P Q R S T V W Y
Site A
Site B
Positive
Negative
Occurrences
1500
1000
500
-0.1 0.0 0.1 0.2 0.3
Residue-Specific Epistasis

Figure 3-1. ***ScaffoldSeq evaluates high-throughput sequence data to characterize the diversity within directed evolution and natural populations.*** **The regions of interest within a collection of proteins are identified using tunable similarity thresholds associated with reference sequences. Background sequences are quantified and excluded from the analysis (optional). Highly similar clones are clustered. Population heterogeneities are further elucidated by dampening the impact of highly frequent unique clones. The software generates output files that detail sitewise amino acid distributions and identify pairwise epistasis.**

3.3. Materials and Methods

3.3.1 Interface Design

Scripts, developed using Python (v2.7) with default libraries to ease portability, are compatible with Windows 7/8, Mac OS X and Linux OS. An intuitive interface guides the user through the sequence analysis menus and allows for command line execution. While workflow settings are customizable for essentially any protein, default profiles for fibronectin[61], affibody[160], DARPin[161], knottin (kalata B1)[78] and Gp2[151] are included. Each unique profile containing scaffold-specific parameters and settings declared within the job submission menu are saved via Python's *pickle* module to separately store setting arrays outside the main interface. This allows users to easily retrieve and view settings from prior analyses. A step-by-step tutorial is included as a resource to guide users through selecting the appropriate job settings (see *Software Walkthrough*).

3.3.2 Sequence Alignment

Quality, relevant sequences are parsed from the FASTA/FASTQ input file by looping through individual reads, locating conserved anchors at the 5' and 3' ends of the gene of interest, and removing segments outside of the anchors. Trimmed sequences of acceptable size (dictated by gene length and allowed length variation, such as from loop length diversity) are

aligned to a user-input reference sequence using a cross-correlation test with Python's *difflib* module. This alignment depends on a sufficient fraction of matching nucleotides within the framework regions, directly outside of the diversified regions of interest, having at least 80% matching nucleotides. While the algorithm does not enable FASTQ read quality filtering, the framework alignment threshold can be adjusted by the user with discretion (see *Software Walkthrough*). This has the effect of searching along a trimmed sequence for the transition between a conserved region and a diversified region of interest. The identification method provides rapid location of diversified regions even in cases where the composition and length of the region of interest are not specifically known, which in turn has the advantage of accurately accounting for specific library designs involving loop length diversity within antibody fragments[162,163] and small scaffolds such as fibronectin[93] and Gp2[151]. Length differences are accounted for by inclusion of gap indicators, "-", to maintain alignment in conserved regions.

3.3.3 Background Consideration

Background sequences or noise should be accounted for based on the specific experiments that yielded the sequence set. In directed evolution, functional clones are often isolated by survival selections or screening via genotype-phenotype display technologies[113,164–166] coupled with cell panning[167], bead sorting[115], or flow cytometry[116]. Survival and selection techniques yield a small fraction of false positives or background, which can be quantified via control experiments (non-random 'false' positives resulting from poor assay design must be accounted for separately). These unwanted, random clones, which are the rarest sequences within the data set, are excluded from the analysis by removing either

a user-defined fraction of the rarest sequences (*e.g.* 2% for example directed evolution population isolated by yeast display and magnetic bead sorting) or all sequences with fewer than a user-defined number of occurrences. Accounting for assay-specific background levels has the additional advantage of sufficiently compensating for read error rates of next-generation sequencing platforms, which tend to be <1%[168].

3.3.4 Family Clustering and Frequency Calculation with Dampening

Sampling bias introduced by dominant motifs and selection methods are detrimental to the statistical analysis of raw sequence data[169,170]. To compensate for these biases, ScaffoldSeq allows for separate correction factors to be associated with motif clustering and individual sequence contributions. Similar sequences (default: 80% identity in diversified sites) can be clustered into families, which facilitate identification of key motifs. Families provide an additional metric for population diversity (sequences, unique sequences, and families), which, in turn, enables broader characterization of the sequence set. Notwithstanding, the diversity of a population can be obscured by a given family similar to how a prominent sequence can upstage all other members of a family. To correct for these imbalances and rather emphasize the heterogeneity of functional clones when appropriate, a dampening exponent can be applied to the frequency of unique sequences to lower the impact of dominant sequences[61,151] (Equation 1).

$$f'(x_i) = \frac{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}}{\sum_i \left\{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}\right\}} \quad (1)$$

where $f(x_i)$ is the frequency of amino acid i at site x; k_m is the m^{th} sequence in family k; and f' is the dampened frequency with d^{th} root dampening.

Traditional sequence analysis often treats each sequence as a distinct solution to a problem. However, within a population, two non-identical, but highly similar sequences may share a common structural or functional motif, akin to providing comparable solutions to the same problem. By lowering the *Sequence Similarity Threshold*, the ScaffoldSeq algorithm defines a broader range of related sequences to be a common solution. The contribution of each common solution (i.e. dominant clones and their common-motif variants) can be tuned to suit the needs of the analysis by using family clustering in combination with dampening.

The *Frequency Dampening Power* ($1/d$) will typically be within the range of 0.25 – 1. As this value approaches zero, the data set will be treated as though all duplicate sequences were removed. A value of 1 has the effect of weighting all sequences equally and, consequently, negates all impact of clustering, irrespective of the *Sequence Similarity Threshold*. *Frequency Dampening Power* of 0.5 is suggested for sequence data sets that contain a relatively high number of occurrences for a few dominant clones. Sensitivity analyses (Fig. 2) guide selection of appropriate parameter value.

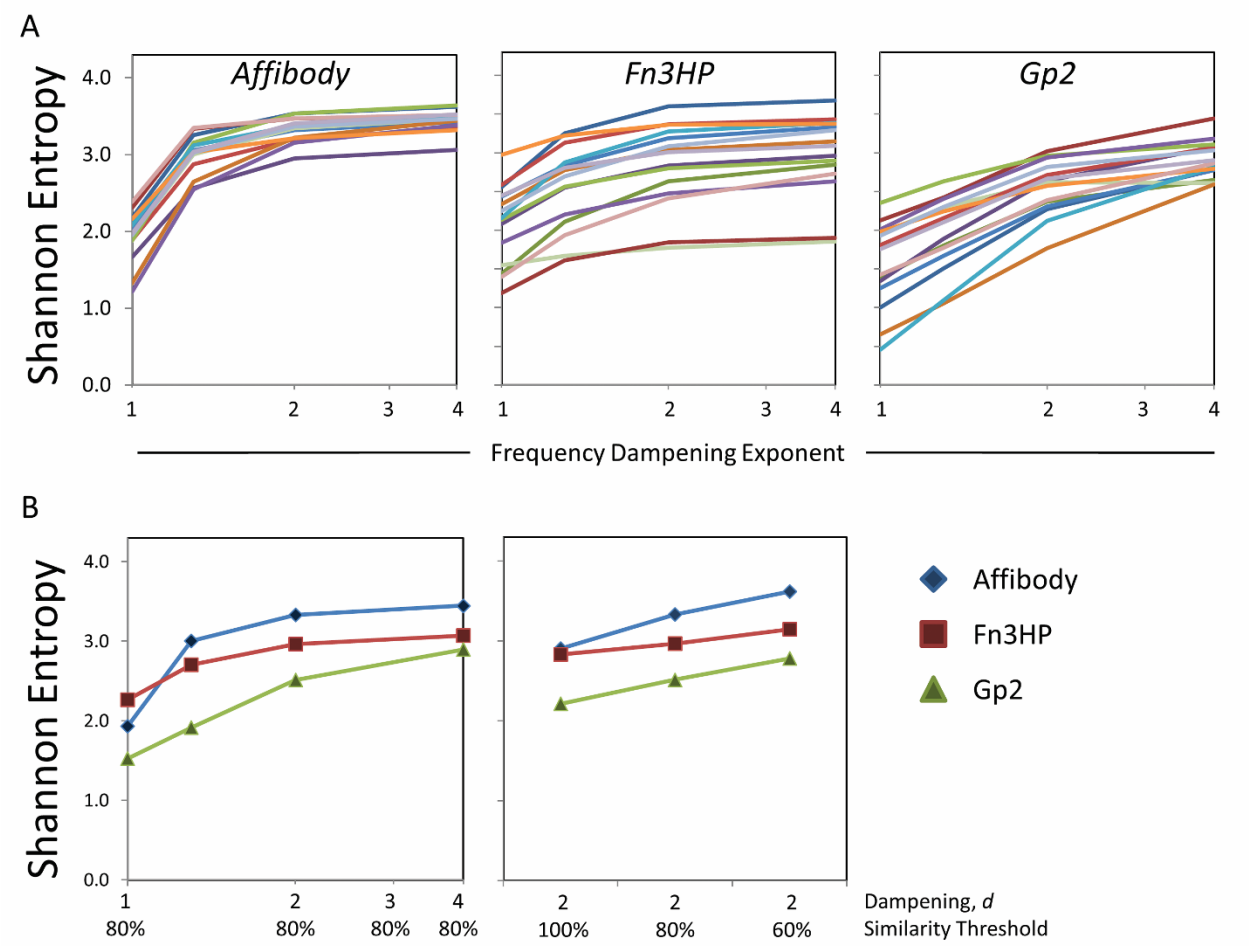


Figure 3-2. *Sensitivity analyses.* (A) Apparent sitewise diversity (i.e. Shannon entropy[130]) within binding populations from recent studies is shown as a function of dampening the effect of dominant sequences by modulating the frequency dampening parameter, *d* (Equation 1). The family cluster similarity threshold is set at 80% for data from all three scaffolds: hydrophilic fibronectin (Fn3HP)[61], Gp2[151], and affibody (unpublished). Data analyses were conducted with frequency dampening coefficient ranging from 1 (no dampening) to 4 (heavy dampening). Each line shows the Shannon entropy of a single site of interest. Conclusions: Accounting for dominant clones, via frequency dampening, has a greater impact on less diverse populations. Dampening uniquely affects each site, as demonstrated by changes in the rank order between levels of dampening. (B) Shannon entropy, averaged across all diversified sites for each scaffold, is shown for a range of frequency dampening coefficients and sequence similarity clustering thresholds. On the left, the similarity threshold is set at 80%. Data analyses were conducted with frequency dampening coefficient ranging from 1 (no dampening) to 4 (heavily dampened). On the right, the frequency dampening coefficient is set to 2 and the similarity threshold is varied. Shannon entropy ($H = -\sum_{i=1}^{20} f_i \log_2 f_i$, where f_i is the fraction of amino acid i at a particular site) describes relative diversity within the range of 0 (fully conserved) to 4.3 (5% of each amino acid).

3.3.5 Pairwise Interactions

Sitewise amino acid frequencies, $f(x_i)$, are most relevant when each site acts independently. In reality, cohorts of residues are likely to interact under evolutionary pressure[171–173]. Therefore, ScaffoldSeq also compares pairwise residue distributions from full-length evolved sequences relative to those predicted by the region specific independent frequency matrix, which empowers identification of positive and negative epistasis[174] (Supplementary Fig. S1-2). Specifically, mutual information, $MI(x,y)$, is calculated for each pair of sites, x and y (Equation 2). Each of the 400 possible amino acid combinations for a site-pair are evaluated based on the predicted sitewise frequency product, $f(x_i)f(y_j)$ and experimentally observed pairwise frequency, $f(x_i,y_j)$ (Supplementary Equation S1). For each of the amino acid-specific contributions, positive values indicate the propensity of two mutations to occur more often than would be predicted by a sitewise frequency analysis alone. Mutations that do not occur within the data set are excluded. The summation of these residue-specific values yields the mutual information for that site-pair. The amino acid-specific components of the mutual information calculation are also output to facilitate epistasis analysis. Mutual information from raw sequences is vulnerable to inaccuracies driven by sequence alignments in multiple scenarios. Broadly diverse or quickly evolving sites with high entropy tend to yield larger mutual information scores, irrespective of paired interaction[175]. Countering these effects through normalization techniques has been discussed[176–179]. Bias is also propagated through redundant sequences and prominent families of highly similar clones[170]. Previous corrective efforts include removing all duplicates[153] and weighting sequence counts inversely proportional to the total cluster

size[180]. Additionally, high background can arise from small samples sizes[181], but can be offset by low count correction[180]. While the ScaffoldSeq algorithm largely overcomes these issues via dampening, clustering, and background removal, the mutual information output data incorporates the average product corrected method[179] (Equation 3)

$$MI(x,y) = \sum_i \sum_j f(x_i, y_j) \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)} \quad (2)$$

$$MI_p(x,y) = MI(x,y) - \frac{MI(x,*)*MI(*,y)}{MI(*,*)} \quad (3)$$

where MI(x,*) and MI(*,y) are the average mutual information values of site-pairs involving site x and y, respectively. MI(*,*) is the average mutual information values across all site-pairs.

3.4. Results and Discussion

ScaffoldSeq has been developed, optimized through extensive testing, and made available for public use along with documentation to facilitate implementation across various applications for the aforementioned functions.[182] Upon completion of the sequence analysis, ScaffoldSeq publishes data in comma-separated value format summarizing each stage of the workflow depicted in Fig. 1. Output data include the total number of quality sequences that were parsed by ScaffoldSeq, number of occurrences for each unique protein sequence, background threshold count by which sequence removal was determined, dampening coefficient applied (*d*), number of family clusters, and protein sequence and frequency for unique clones within each cluster. Sitewise amino acid counts and frequency distributions are presented in matrix form. Following this, lead clones from the population

(i.e. most prevalent clone from each sequence cluster) are enumerated in rank order. Pairwise similarity distances are computed for each lead clone based on the relative Hamming distance as well as the revised BLOSUM64 score matrix[183]. Many analyses are afforded by the ScaffoldSeq output beyond those included in the default package. One potential application is demonstrated whereby the list of lead clones from each family is used to construct a phylogenetic tree, allowing for visual assessment of high level diversity (Fig. 3). Pairwise diversity analysis, evaluated via mutual information and residue-specific epistasis, is conducted for all 400 pairs of residues, *i* and *j*, at all pairs of sites, *x* and *y* (further discussion in Supplementary Figs. S1-2).

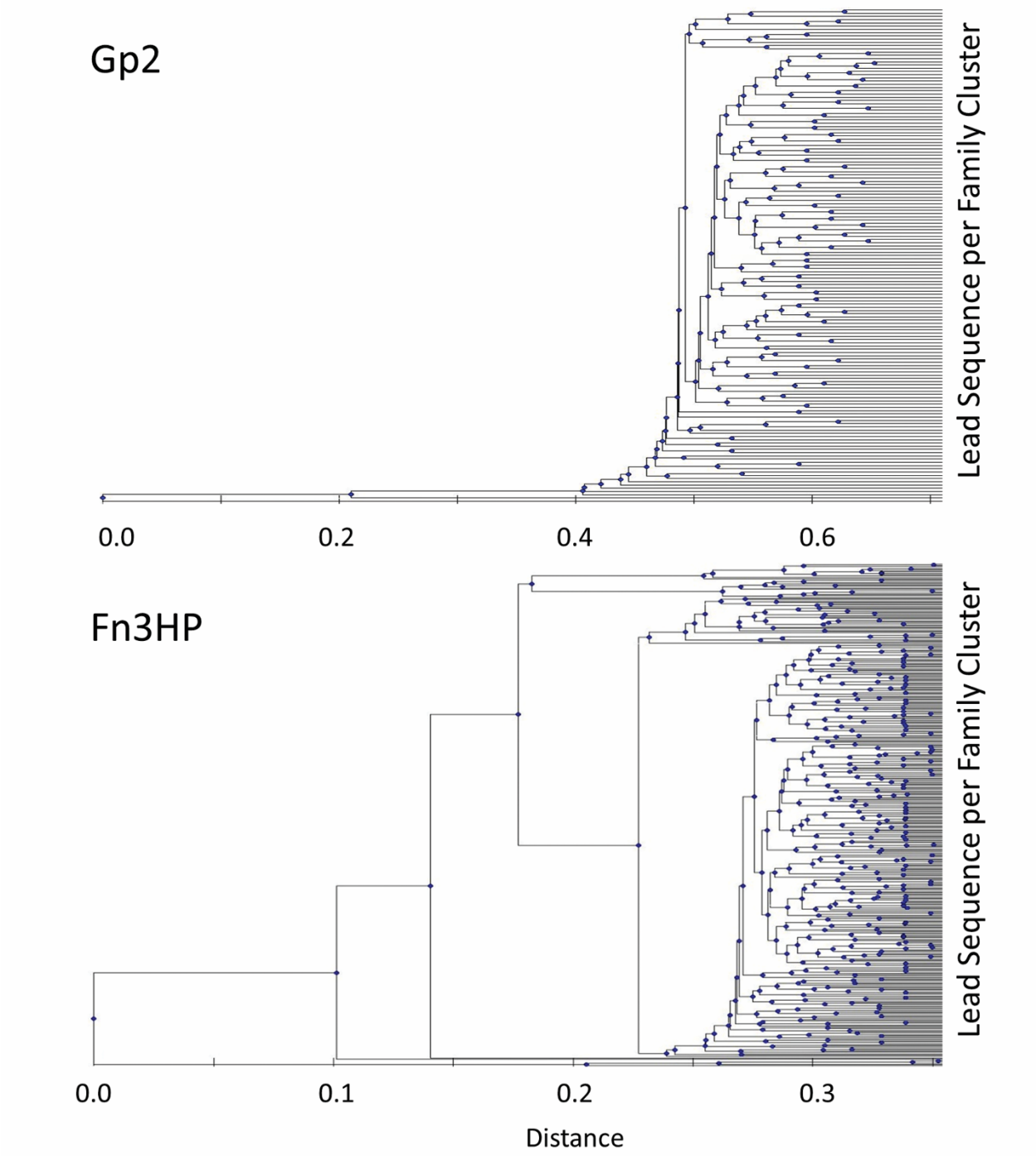


Figure 3-3. ***Family clustering performed within ScaffoldSeq can be further evaluated in a phylogenetic tree visualization.*** **The list of dominant sequences within each clustered family, directly output from ScaffoldSeq in the .csv file, were input into the *seqpdist* and *seqlinkage* functions within MATLAB. Horizontal lines on the far right indicate each unique sequence. The x-axis quantifies the distances between sequences based on the Jukes-Cantor method and blosum50 scoring matrix. The data sets originate from evolved populations of high affinity binders, sequenced by Illumina MiSeq, were analyzed using ScaffoldSeq. Analysis parameters for the Gp2 scaffold (top) and**

hydrophilic fibronectin (Fn3HP; bottom) populations included a clustering threshold of 0.85 and 0.95, respectively and an assay background filter of 10 for both datasets.

High-throughput evolution (directed or natural) and deep sequencing can substantially advance our knowledge of sequence-function relationships to yield improved mutant or combinatorial library designs. In addition to enlightening analysis of single proteins, sitewise (single and paired) consideration of inter- and intra-molecular interactions – quantified via evolutionary prevalence – can aid combinatorial library designs for de novo protein discovery. ScaffoldSeq facilitates such analyses.

3.5. Acknowledgements

This work was supported by the Department of Defense (W81XWH-13-1-0471 to B.J.H.), the National Institutes of Health (R21 EB019518 to B.J.H.) and the University of Minnesota. The authors have no conflict of interest to declare.

3.6. Supplemental Figures

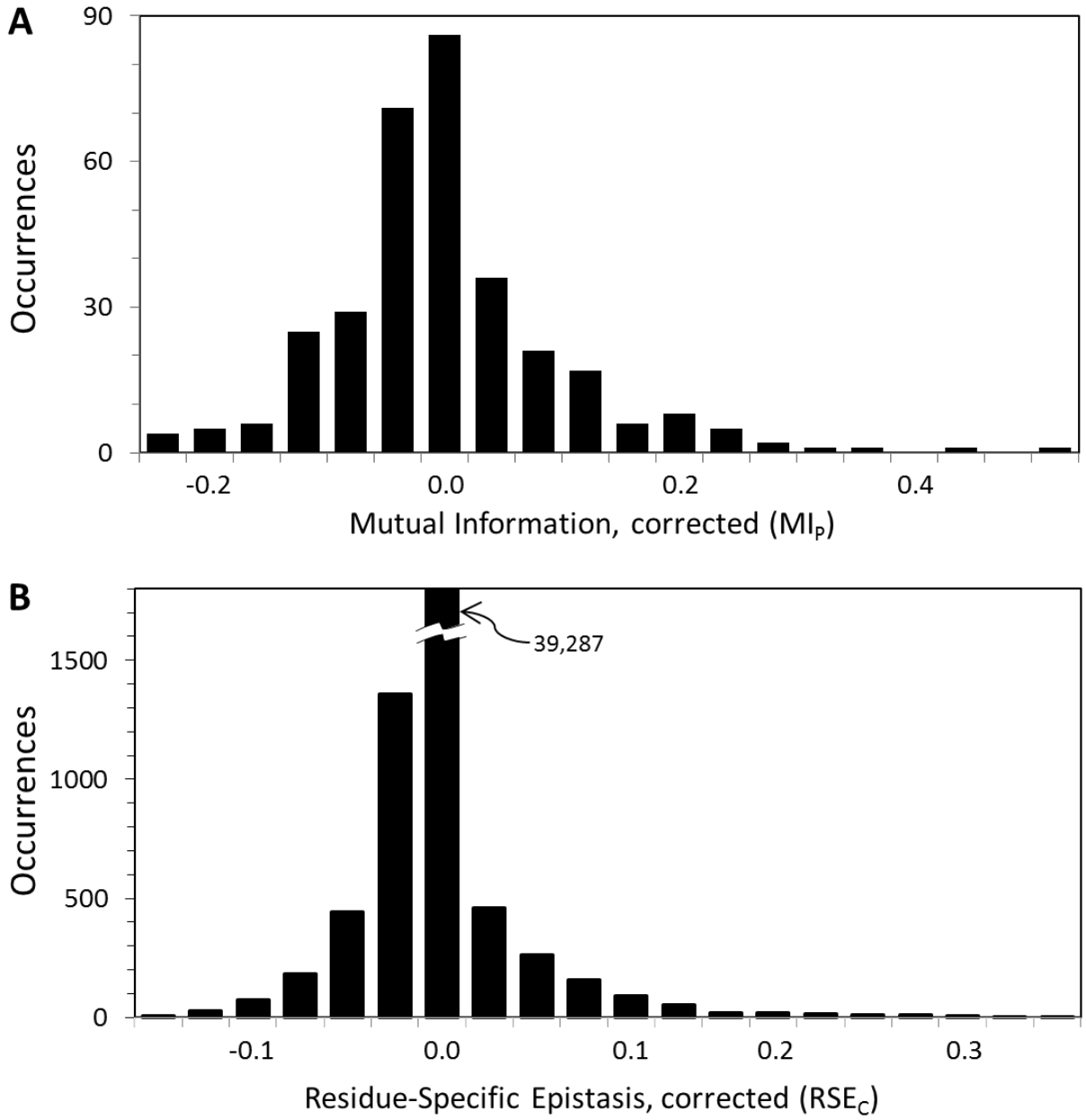


Figure S3-1. ***Histogram of mutual information.*** **(A) and residue-specific epistasis (B) values. In a recent study (Woldring, et al., 2015), high-throughput sequencing was used to elucidate the sequence-function landscape of a hydrophilic fibronectin (Fn3HP) combinatorial library. From a naïve library offering diversity at 26 solvent exposed positions, >10^5 unique strong binding ligands were evolved. ScaffoldSeq used these sequence data to calculate sitewise amino acid distributions as well as conduct pairwise analysis over the 325 possible site-pair combinations. (A) The corrected mutual information, $MI_P(x, y)$, is calculated for each pairwise combination of two positions, x and y (Dunn, et al., 2008). (B) The average product correction method was applied to each combination of amino acids at all 325 site-pairs.**

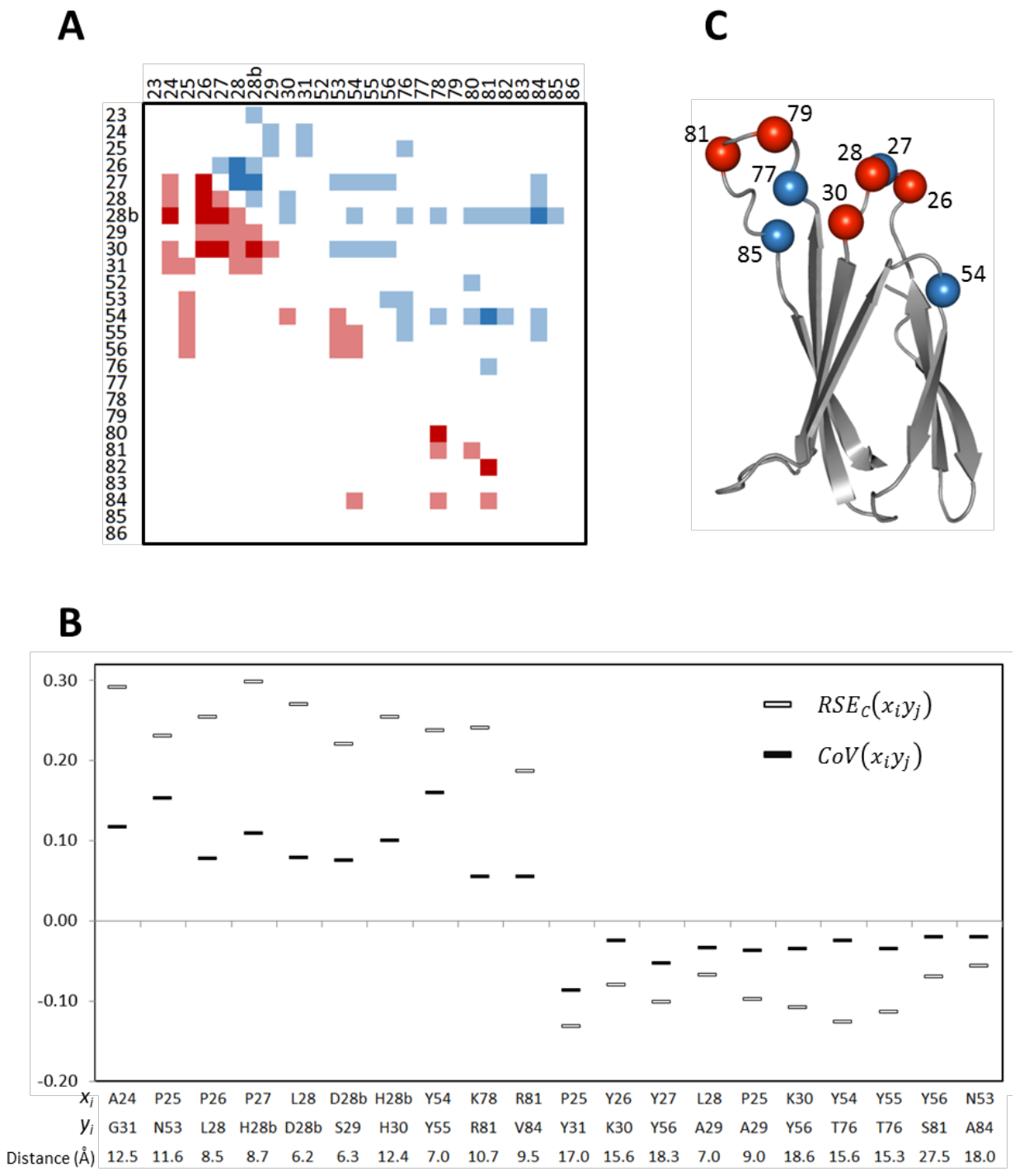


Figure S3-2. ***High affinity binders evolved from a hydrophilic fibronectin (Fn3HP) combinatorial library were Illumina sequenced and analyzed within ScaffoldSeq (Woldring, et al., 2015).*** **Residue-specific epistasis (RSE_C; equation S4) is calculated for each pairwise combination of residues, i and j, at a two positions, x and y. (A) At each pair of sites, the total number of occurrences yielding a value above 0.17 are in**

red on the bottom half of the figure. The summation of simultaneous mutations that yield $RSE_C(x_i y_j)$ values below a threshold of -0.09 are shown in blue on the top portion. Light or dark shades indicate a single occurrence or two occurrences, respectively. (B) The highest and lowest observed RSE_C and covariance (equation S9) values are shown for a representative set of residues and sites. The specific mutations are along the x-axis. Listed below the mutations is the median distance (angstroms) between the two sites based on Fn3 structural information from the protein data bank. Pairwise analysis provides additional information for directed evolution studies and focused library design. The existence of strong correlations, both positive and negative, are found within select pairs of sites, highlighting the potential impact of amino acid choice at diversified sites. In the context of library design, this information can be supplemented with other sources of data such as crystal structures (C; PDB ID – 1TTG) or stability calculations (Woldring, et al., 2015).

Equations S1-9. Pairwise analysis conducted by ScaffoldSeq uses residue-specific epistasis (equation S1) to quantify the disparity between two mutations simultaneously occurring in the observed population, $f(xi,yj)$, and that which you would otherwise predict based on the frequency of each isolated mutation occurring within the same population, $f(xi)f(yj)$. Positive values indicate the propensity of two mutations to occur more often than would be predicted by a sitewise frequency analysis alone. Mutual information is calculated by summing over all residue-specific epistasis combinations at two sites (equation S2). To reduced the inherent bias associated with entropy in this metric, the average product correction method, discussed previously (Dunn, et al. 2008), is shown in equation S3. In the present work, a similar technique was applied to residue-specific epistasis (equation 4). Supporting equations for the average product correction method are shown in equations S5-8. The covariance of two mutations is stated in equation S9.

Residue–Specific Epistasis: $RSE(x_i y_j) = f(x_i y_j) \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)}$ Equation S1

Mutual Information: $MI(x,y) = \sum_i \sum_j f(x_i, y_j)\ \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)}$ Equation S2

MI, corrected: $MI_p(x,y) = MI(x,y) - \frac{MI(x,*)*MI(*,y)}{MI(*,*)}$ Equation S3

RSE, corrected: $RSE_C(x_i y_j) = RSE(x_i y_j) - \frac{RSE(x_i,*)*RSE(*,y_j)}{RSE(*,*)}$ Equation S4

Mean MI at position x: $MI(x_i,*) = \frac{1}{(\sigma-1)} \sum_{y,y\neq x} \sum_{i,j} f(x_i y_j) \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)}$ Equation S5

Mean RSE at position x: $RSE(x_i,*) = \frac{1}{\rho(\sigma-1)} \sum_{y,y\neq x} \sum_{j} f(x_i y_j) \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)}$ Equation S6

Mean MI at all positions: $MI(*,*) = \frac{2}{(\sigma^2-\sigma)} \sum_{x,y,y\neq x} \sum_{i,j} f(x_i y_j) \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)}$ Equation S7

Mean RSE at all positions: $RSE(*,*) = \frac{2}{\rho^2(\sigma^2-\sigma)} \sum_{x,y,y\neq x} \sum_{i,j} f(x_i y_j) \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)}$ Equation S8

Covariance: $CoV(x_i y_j) = f(x_i y_j) - f(x_i) f(y_j)$ Equation S9

Definitions

x,y	individual protein positions
i,j	amino acids
ρ	number of residue options at each position
σ	total number of sites

3.7. ScaffoldSeq – Software Walkthrough

3.8. Software Walkthrough

3.8.1 Overview

ScaffoldSeq is software designed for the numerous applications – including directed evolution analysis – in which a user generates a population of DNA sequences encoding for partially diverse proteins with related functions and would like to characterize the single site and pairwise amino acid frequencies across the population. Importantly, the software provides tools to cluster similar protein families, dampen the impact of dominant clones, remove background, and evaluate diversity.

3.8.2 Workflow

1. ScaffoldSeq reads high-throughput DNA sequences from FASTA/FASTQ files.
2. Regions of interest are parsed; unique sequences are enumerated.
3. Background sequences (i.e. the rarest clones) are quantified and omitted from analysis, if desired.
4. Highly similar clones are clustered.
5. Dampen dominant clones.
6. Output graphical and tabular results for (a) sitewise amino acid frequency and (b) pair-wise epistasis analysis.

3.8.3 Downloads (Two Options)

http://research.cems.umn.edu/hackel
https://github.com/HackelLab-UMN

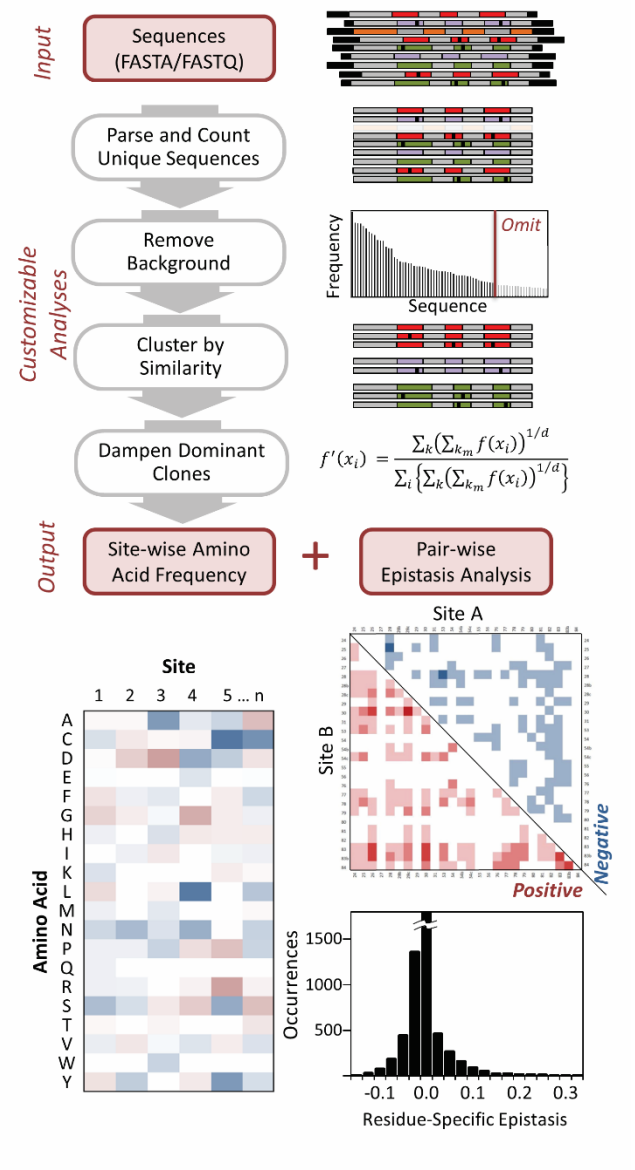


Figure SW1: ScaffoldSeq workflow.

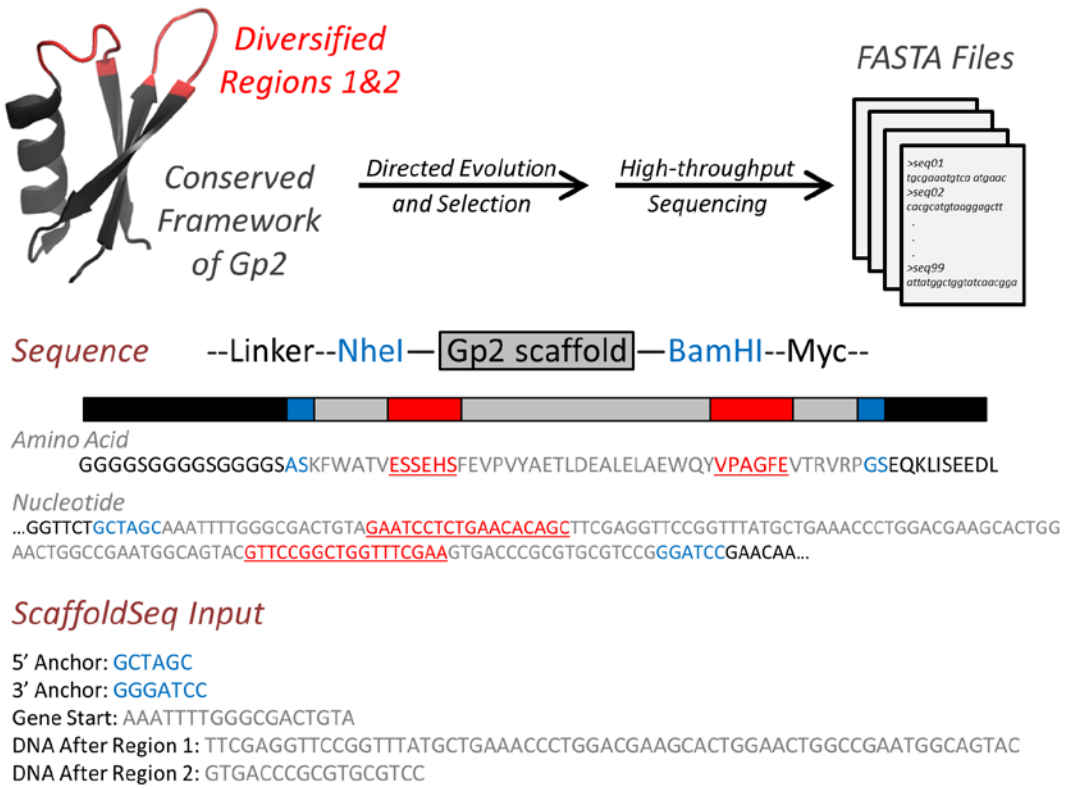


Figure SW2: Representative sequence analysis scenario. The Gp2 scaffold[151] was analyzed using an in-development version of ScaffoldSeq, similar to a previous study[61]. From the 45-amino acid parental domain (PDB: 2WNM), a combinatorial library was employed whereby the two solvent exposed loops (red) were diversified in genetic sequence as well as length, with the inclusion of 6, 7, or 8 residues within each of the two regions. Populations of high-affinity binding clones evolved from this library were sequenced across the entire indicated gene (Illumina MiSeq, paired-end). Raw sequences were groomed using PANDAseq[119], producing FASTA files of full-length reads (see Paired-end Assembly section). Using the FASTA files, ScaffoldSeq evaluated the sitewise and pairwise diversity throughout the two regions of interest (red). To be included in the analysis, an entry within the FASTA file must contain matching segments for both the 5' / 3' anchors (blue) as well as the framework regions (gray) adjacent to the diversified regions (red). Default anchor and framework matching thresholds are 100% and 80%, respectively. This identifies the appropriate genes and localizes the analysis to the intended regions even within a diverse population. Note that the conserved framework positions (gray) are excluded from all future analysis. To analyze the full gene sequences, simply specify the anchor and framework sequences to be directly adjacent to, but not overlapping with the gene region. In the following walkthrough, analysis parameters were set at 0.25 for dampening (1/*d*) with a similarity clustering threshold of 0.8.

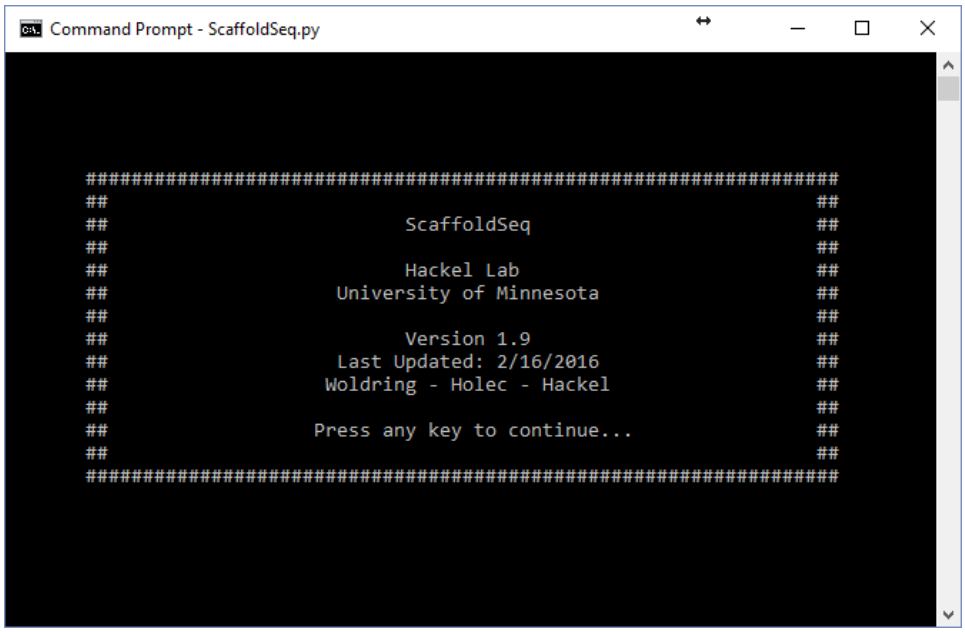
ScaffoldSeq.py is compatible with any common operating system (Windows 7/8/10, Mac OS X or Linux OS) that has Python 2.7 installed.

The software package can be downloaded from either the Hackel Lab research page (http://research.cems.umn.edu/hackel/Hackel/Publications.html) or the GitHub repository (https://github.com/HackelLab-UMN).

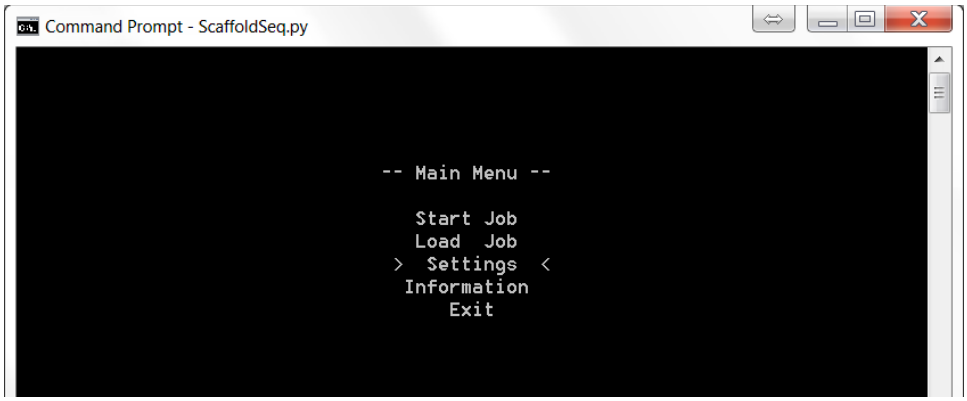
The script is intended to run via the operating system command line or Python terminal, rather than IDLE. Start by ensuring that both the sequence data file and ScaffoldSeq.py are located in the same directory. Then navigate to this directory using the command prompt and load the program, for example:

```
C:\User\Profile\GitHub\ScaffoldSeq>ScaffoldSeq.py
```

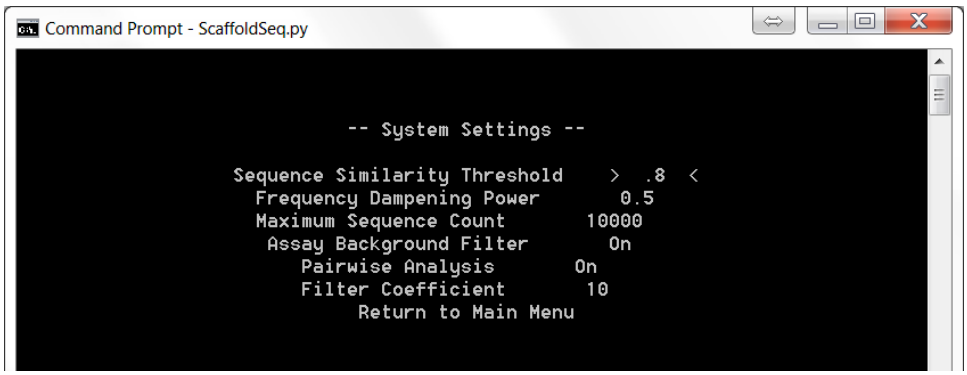
An introduction screen will then be shown.



Press any key to continue. The Main Menu is navigated using keyboard arrows (←→↑↓), then pressing Enter. You can exit at any time by pressing Esc.



Sequence analysis parameters can be specified within *Settings*.



Sequence Similarity Threshold specifies the minimum fraction of sitewise amino acid matches required to place two sequences of the same region into a common cluster. *Frequency Dampening Power* (1/*d*) operates on the individual family clusters by applying a weight to the total count of each residue-position pair, as shown in Equation 1:

$$f'(x_i) = \frac{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}}{\sum_i \left\{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}\right\}} \quad (1)$$

where $f_{i,j}$ is the observed occurrence of amino acid *i* at site *j within* the m^{th} sequence of family *k*; and *f'* is the dampened frequency with d^{th} root dampening. Traditional sequence analysis often treats each sequence as a distinct solution to a problem. However, within a population, two non-identical, but highly similar sequences may share a common structural or functional motif, akin to providing comparable solutions to the same problem. By lowering the *Sequence Similarity Threshold*, the ScaffoldSeq algorithm defines a broader range of related sequences to be a common solution. The contribution of each common solution (i.e. dominant clones and their common-motif variants) can be tuned to suit the needs of the analysis by using family clustering in combination with dampening.

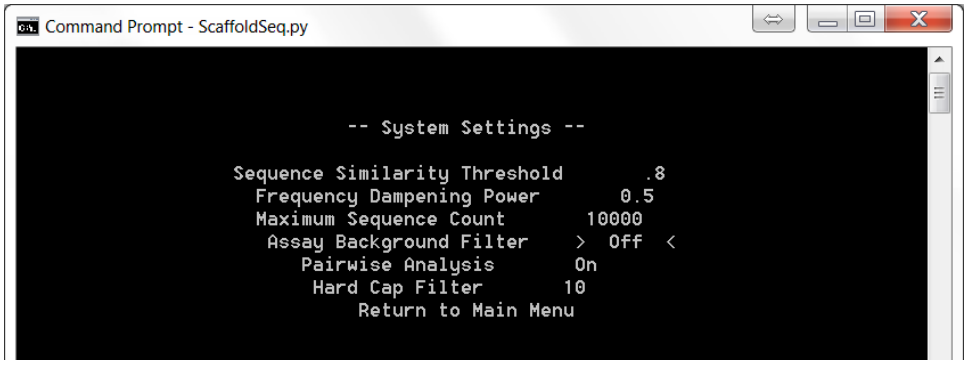
The *Frequency Dampening Power* (1/*d*) will typically be within the range of 0.25 – 1. As this value approaches zero, the data set will be treated as though all duplicate sequences were removed. A

value of 1 has the effect of weighting all sequences equally and, consequently, negates all impact of clustering, irrespective of the *Sequence Similarity Threshold*. *Frequency Dampening Power* of 0.5 is suggested for sequence data sets that contain a relatively high number of occurrences for a few dominant clones.

Maximum Sequence Count sets an upper limit for the number of sequences included in the analysis. This can be set to limiting values to speed analysis time for preliminary explorations.

Background sequences or noise should be accounted for based on the specific experiments that yielded the sequence set. *Assay Background Filter* refers to a quantifiable, assay-specific level of false positives or background. The *Filter Coefficient* is the ratio of total events to false positives. If this ratio is unknown, the *Assay Background Filter* can be turned off using the left and right arrow keys. When turned off (see below), the minimum number of occurrences (*Hard Cap Filter*) can be specified. All unique sequences that are observed less often than this value will be neglected in the analysis. Toggling *Pairwise Analysis* On/Off gives the option of performing this computationally expensive feature.

Selecting *Return to Main Menu* will save these settings.



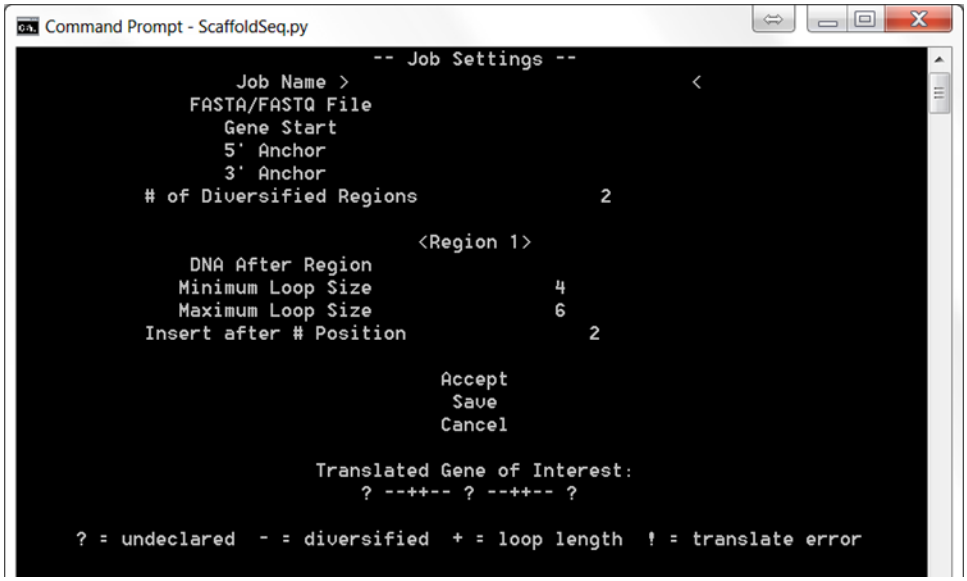
From the Main Menu, the *Start Job* option leads to a screen where the FASTA/FASTQ input file and scaffold specific DNA information are entered. *Job Name* will become the leading name of the output files. The input *FASTA/FASTQ File* is specified on the second row.

DNA sequences should be entered using all caps. Sequences can be typed in manually or pasted into the appropriate field by first selecting the field using the arrow keys, then right clicking and selecting *paste.* The Windows keyboard shortcut *ctrl-c* will likely not be accessible within the Python terminal.

Gene Start refers to all nucleotides within the gene of interest that lead up to the first diversified region. The *5' Anchor* and *3' Anchor* fields should be conserved nucleotide sequences that directly precede and follow the gene of interest, respectively. These are commonly in the form of restriction enzyme cut sites or adapter sequences that were part of the amplicon sample preparation prior to high-throughput sequencing run. The *Gene Start* nucleotides should be a conserved region that directly follows the *5' Anchor* and directly precedes the first diversified

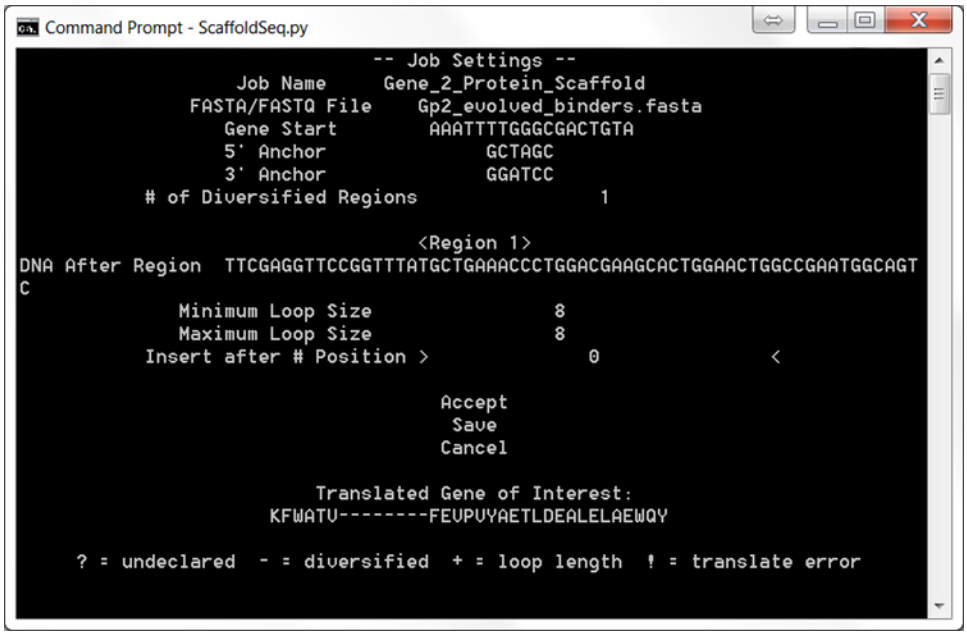
region. If multiple Diversified Regions are being investigated, the *DNA After Region* field should include all nucleotides directly following the selected Region and lead up to the next Diversified Region. The nucleotides that follow the final Diversified Region must begin directly following the diversified region and end at the nucleotide preceding the *3' Anchor* sequence. Note that both the *Anchor* matching threshold and the framework threshold are adjustable global variables within the script (ScaffoldSeq.py): *adaptor_tolerance* (default: 100%) and *framework_match_threshold* (default: 80%), respectively.

The following example coincides with the scenario shown in *Figure 2*.

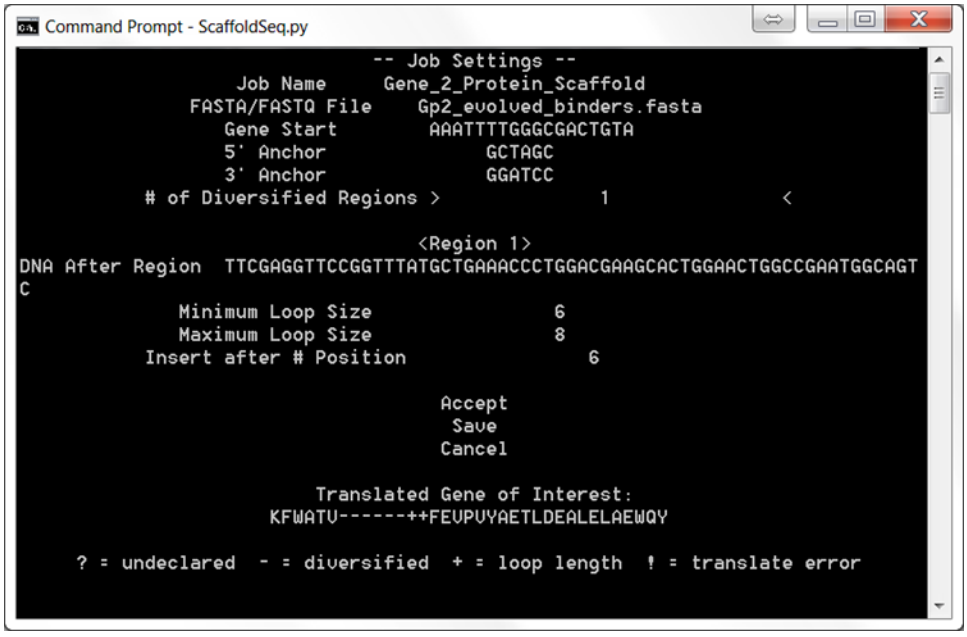


As nucleotides are entered into the *Gene Start* and *DNA After Region* fields, the *Translated Gene of Interest* section will populate.

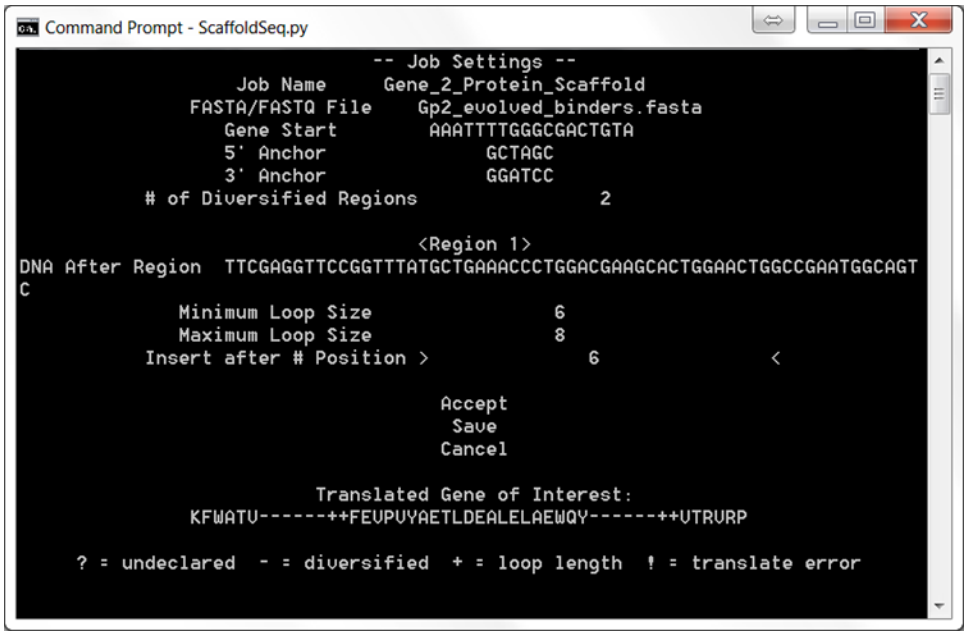
When analyzing Diversified Regions, the nucleotides within that region should not be keyed in; however, the total number of amino acids within the diversified region must be specified with the *Minimum* and *Maximum Loop Size* fields. The Diversified Regions will be displayed as dashes, "-", at the bottom of the screen. The amino acids that are displayed as letters at the bottom will not be included in the sitewise or pair-wise analysis. They are displayed to assist the user in ensuring the diversified regions are accurately positioned for the analysis.



For analyses that harbor loop length diversity within the diversified region, the position of insertion can be selected following the *Minimum* and *Maximum Loop Size* fields. The loop length diversity sites are indicated as '+' within the translated sequence.

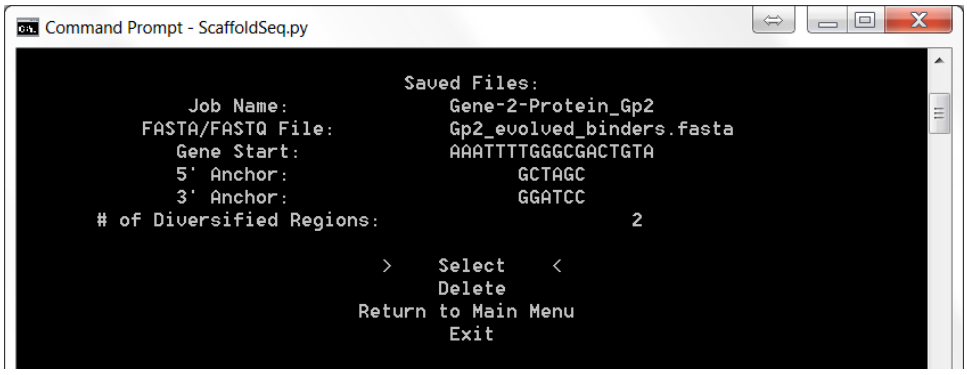


Additional sequence regions can be included in the analysis by using the keyboard arrows to specify the # of diversified regions. Selecting *Save* will store the settings to a file in the working directory.

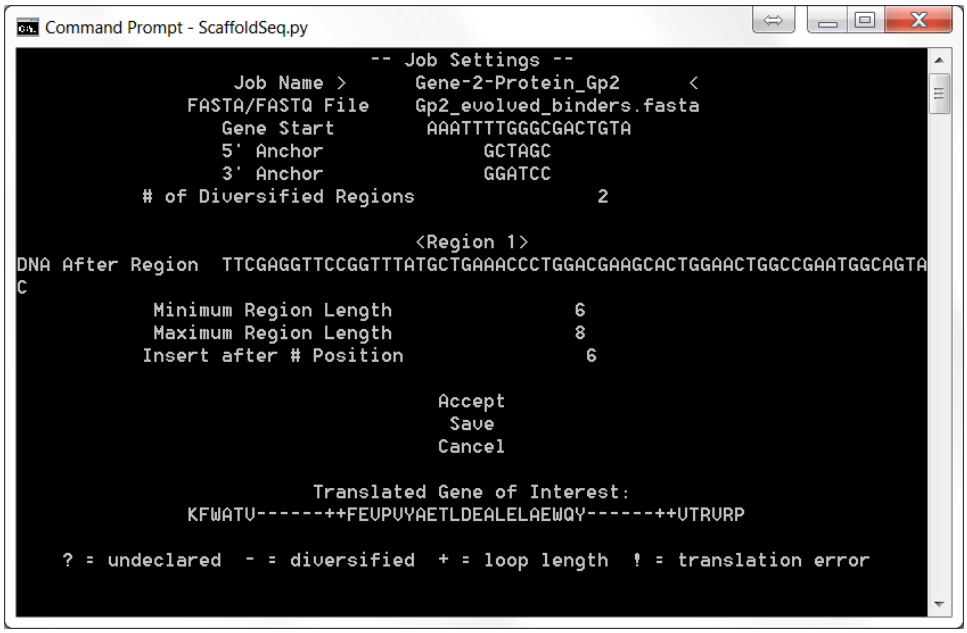


All custom settings saved by the user will be located within the *Load Job* menu for future use.

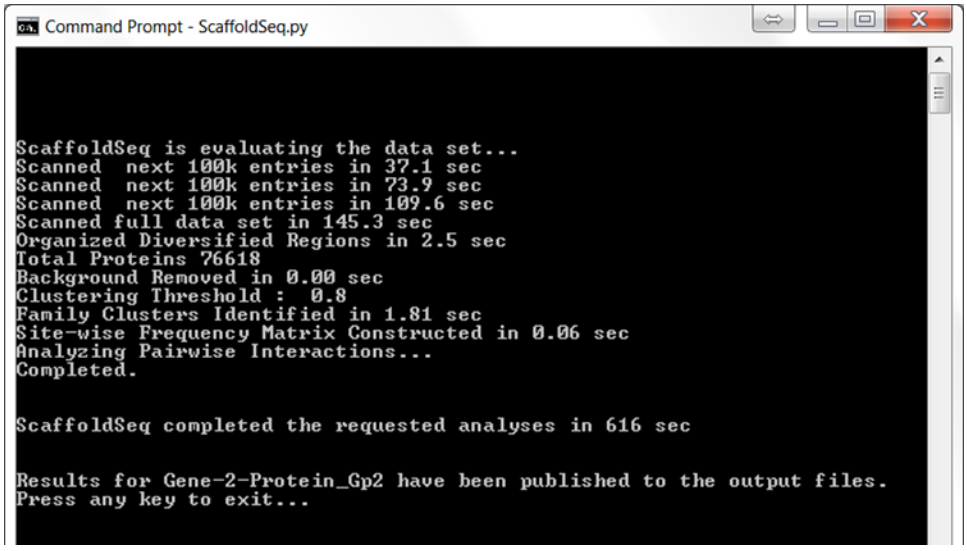
Use the arrow keys to browse summaries for each of the saved jobs.



Upon selecting a job from with the *Saved Files*, you enter the *Job Settings* environment. At this point, press *Accept* to start the job. Confirm by pressing any key or Esc to abort.



As the script is performing each task, a brief status is delivered to the user as shown in the next image. Note that both the sequence file and ScaffoldSeq.py should be contained within the current working directory. Failure to do so will result in an error at which point the user must verify that (a) the sequence data file is located within the current directory and (b) the file name was entered properly. Proper implementation will produce incremental status updates resembling the following:



Output files consists of sitewise amino acid frequency heatmaps (shown below) and tabular summaries (*.csv) of family clusters for each region of interest. If *Pairwise Analysis* was selected, an additional tabular summary (*.csv) is output, which includes mutual information (Equation 2) – with and without average product correction[179] (Equation 3) – and epistasis (amino acid-specific

components of mutual information). Supporting equations for epistasis are shown in equations 5-7.

$$MI(x,y) = \sum_i \sum_j f(x_i, y_j)\ \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)} \quad (2)$$

$$MI_p(x,y) = MI(x,y) - \frac{MI(x,*)*MI(*,y)}{MI(*,*)} \quad (3)$$

where MI(x,*) and MI(*,y) are the mean mutual information values of site-pairs involving site x and y, respectively. MI(*,*) is the mean mutual information values across all site-pairs.

$$\textit{Residue–Specific Epistasis: } RSE(x_i y_j) = f(x_i y_j) log\left\{\frac{f(x_i y_j)}{f(x_i)f(y_j)}\right\} \quad (4)$$

$$\text{RSE, corrected: } RSE_c(x_i y_j) = RSE(x_i y_j) - \frac{RSE(x_i,*)*RSE(*,y_j)}{RSE(*,*)} \quad (5)$$

$$RSE(x_i,*) = \frac{1}{\rho(\sigma-1)} \sum_{y,y\neq x} \sum_j f(x_i y_j) log\left\{\frac{f(x_i y_j)}{f(x_i)f(y_j)}\right\} \quad (6)$$

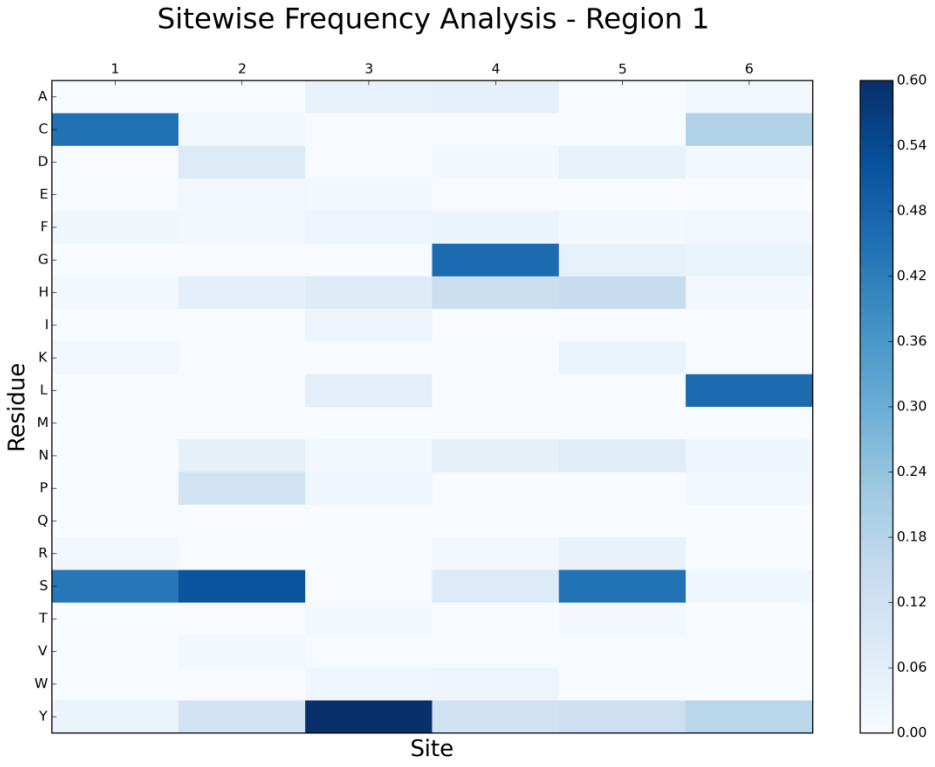
$$RSE(*,*) = \frac{2}{\rho^2(\sigma^2-\sigma)} \sum_{x,y,y\neq x} \sum_{i,j} f(x_i y_j) log\left\{\frac{f(x_i y_j)}{f(x_i)f(y_j)}\right\} \quad (7)$$

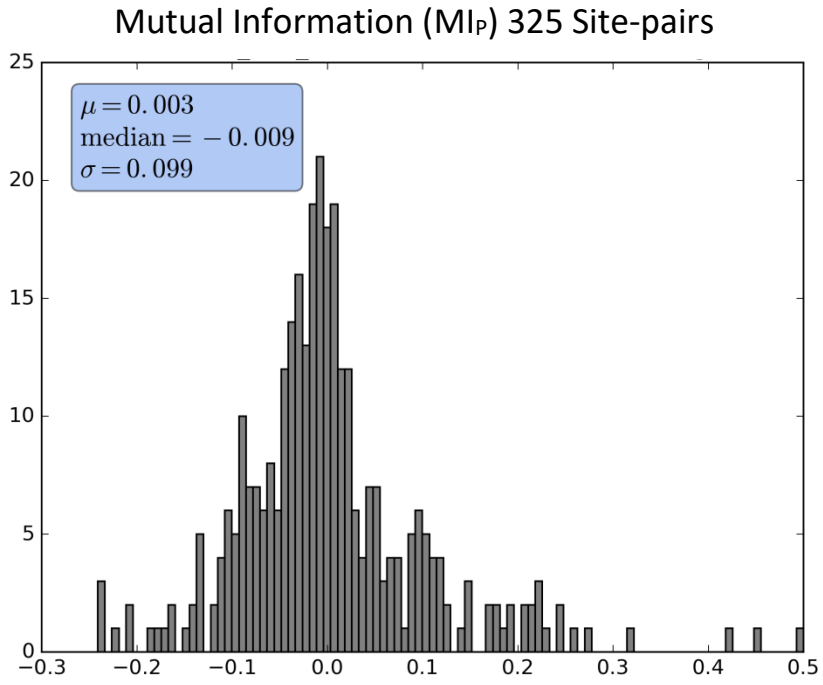
where x and y are individual positions, i and j are amino acids, ρ is the number of residue options at each position and σ is the total number of sites. Additional algorithms that improve upon mutual information have been described by others[175–178,184].

3.8.4 Representative Output Figures

The figures below demonstrate the graphical output provided by ScaffoldSeq. The visualization modules (matplotlib, pandas, numpy) required for figure output are not included within default Python install. These are easily installed using pip via command line (pip.pypa.io/en/stable/reference/pip_install/).

Sitewise amino acid frequency heatmaps are shown for each region of interest with a color scale bar having default range 0 – 60%. Pairwise analysis is summarized by a histogram which includes mutual information using the average product correction method.


Sitewise Frequency Analysis - Region 1
1
2
3
4
5
6
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
Residue
Site
0.60
0.54
0.48
0.42
0.36
0.30
0.24
0.18
0.12
0.06
0.00


Mutual Information (MIP) 325 Site-pairs
$\mu = 0.003$
$\text{median} = -0.009$
$\sigma = 0.099$
25
20
15
10
5
0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5

3.8.5 Silent Mode

To run the software in the absence of the dynamic interface, ScaffoldSeq can be executed from the command prompt in silent mode. This is done by including a text file (e.g. jobname.txt) as the single argument for ScaffoldSeq:

```
C:\User\Profile\GitHub\ScaffoldSeq>ScaffoldSeq.py jobname.txt
```

The job text file must include a complete list of predetermined settings and parameters, shown below:

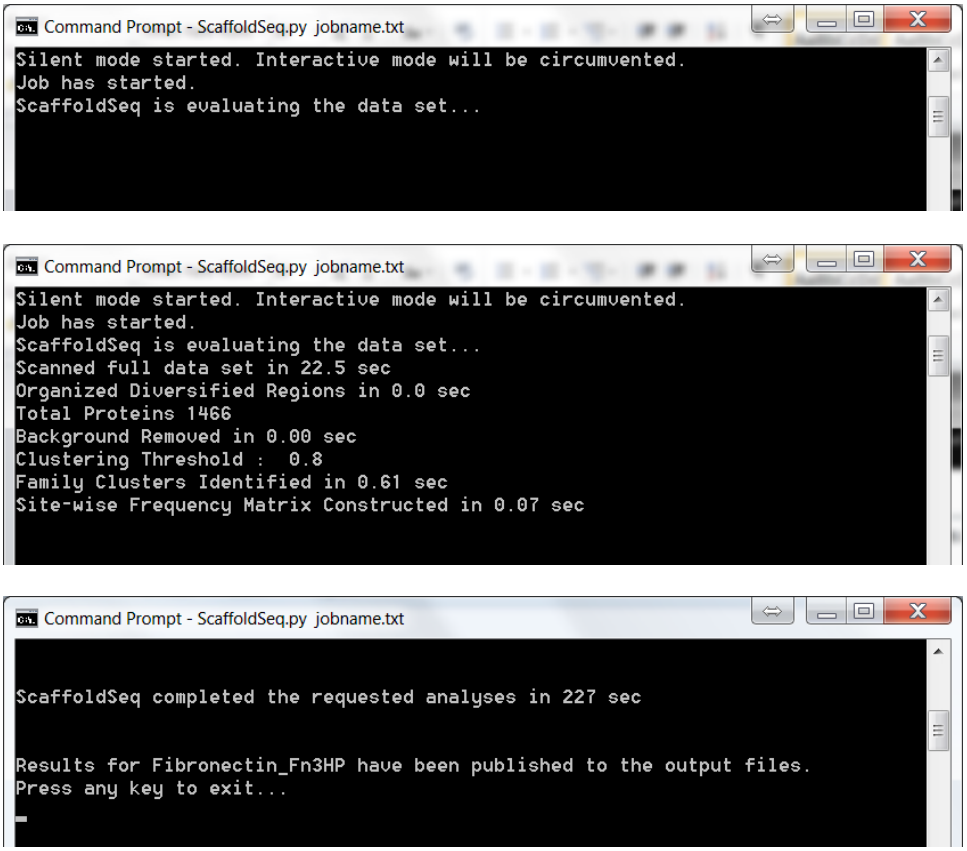
```
Job Name:
FASTA/FASTQ File:
Gene Start:
5' Anchor:
3' Anchor:
# of Diversified Regions:
DNA After Region:
Minimum Region Length:
Maximum Region Length:
Insert after # Position:
Sequence Similarity Threshold:
Frequency Dampening Power:
Maximum Sequence Count:
Assay Background Filter:
Pairwise Analysis:
Filter Coefficient:
```

A sample job file is included in the software package. The file contents are shown below:

```
Job Name: Fibronectin_Fn3HP
FASTA/FASTQ File: High_affinity.fasta
Gene Start:
TCCTCCGACTCTCCGCGTAACCTGGAGGTTACCAACGCAACTCCGAACTCTCTGACTATTTCTTG
G
5' Anchor: GCTAGC
3' Anchor: GGATCC
# of Diversified Regions: 3
DNA After Region:
TACCGTATCACCTACGGCGAAACTGGTGGTAACTCCCCGAGCCAGGAATTCACTGTTCCG,GCGA
CCATCAGCGGTCTGAAACCGGGCCAGGATTATACCATTACCGTGTACGCTGTA,CCAATCAGCAT
CAATTATCGCACCGAAATCGACAAACCGTCTCAG
Minimum Region Length: 6,3,6
Maximum Region Length: 11,7,12
Insert after # Position: 3,1,3
Sequence Similarity Threshold: 0.8
Frequency Dampening Power: 1
Maximum Sequence Count: 10000
Assay Background Filter: On
Pairwise Analysis: Off
```

Filter Coefficient: 10

Upon running a job using silent mode, the window should display brief descriptions of progress:



Output files will be exported to the working directory.

3.8.6 Runtime and Memory Requirements

The analyses discussed in this walkthrough were conducted on a standard desktop PC (Windows 10, Intel i5 4590 @3.3GHz, 16GB RAM). The RAM requirements are dictated by the total number of unique sequences being processed, while the overall runtime governed by the total number of clusters. As a representative high-RAM test case mimicking the analysis of a naïve or unselected population not requiring clustering, $1x10^7$ unique clones were pseudo-randomly generated in the framework of the Gp2 scaffold with NNK codons at each diversified position. This analysis required 2 hours to run with a peak RAM usage of 4.2GB. With 10-fold fewer sequences, this process takes only 12 minutes (400MB). As a representative long-runtime test case mimicking the analysis of broadly diverse population of matured clones, $1x10^7$ sequences (10% unique) were generated in the framework of the Gp2 such that $1x10^5$ family clusters were organized. The more computationally demanding tasks of clustering added 4 hours to the total runtime. When allowing for $1x10^4$ clusters, this results in an 18-fold reduction in time required for clustering.

3.8.7 Paired-end Assembly

Multiple algorithms[185–188] exist for processing paired-end reads. The above examples used PANDAseq[119] for assembling quality sequences into FASTA files. Below, is a basic template for using this method. The forward [-f] and reverse [-r] reads are input as separate FASTQ files. Multi-threading [-T] can be enabled for CPU-bound situations. A sequence quality threshold [-t] can be adjusted to reduce the presence of low-quality reads. This process generates an output file [-w].

```
module load pandaseq
FOR = HACKEL_S1_L001_R1_001.fastq
REV = HACKEL_S1_L001_R2_001.fastq
OUT = panda_assembled.fasta
pandaseq -f $FOR -r $REV -T 4 -t 0.99 –w >$OUT
```

Chapter 4 – A gradient of sitewise diversity promotes evolutionary fitness for binder discovery in a three-helix bundle protein scaffold

Adapted from: Daniel R. Woldring, Patrick V. Holec, Lawrence A. Stern, Yang Du, and Benjamin J. Hackel. "A gradient of sitewise diversity promotes evolutionary fitness for binder discovery in a three-helix bundle protein scaffold." *Biochemistry*. March 2017. DOI: 10.1021/acs.biochem.6b01142

4.1. Synopsis

Engineered proteins provide clinically and industrially impactful molecules and utility within fundamental research. Yet, inefficiencies in discovering lead variants with new desired function, while maintaining stability, hinder progress. Improved function, which can result from a few strategic mutations, is fundamentally separate from discovering novel function, which often requires large leaps in sequence space. While a highly diverse combinatorial library covering immense sequence space would empower protein discovery, the ability to sample only a minor subset of sequence space and the typical destabilization of random mutations preclude this strategy. A balance must be reached. At library scale, compounding several destabilizing mutations renders many variants unable to properly fold and devoid of function. Broadly searching sequence space while reducing destabilization may enhance evolution. We exemplify this balance with affibody, a three-helix bundle protein scaffold. Using natural ligand datasets, stability and structural computations, and deep sequencing thousands of binding variants, a protein library was designed on a sitewise basis with a gradient of mutational levels across 29% of the protein. In direct competition of biased and uniform libraries, both with 1x109 variants,

for discovery of 6x104 ligands (5x103 clusters) toward seven targets, biased amino acid frequency increased ligand discovery 13±3-fold. Evolutionarily favorable amino acids, both globally and site-specifically, are further elucidated. The sitewise amino acid bias aids evolutionary discovery by reducing mutant destabilization as evidenced by 15ºC higher midpoint of denaturation relative to unbiased mutants (62±4ºC vs. 47±11ºC, $p<0.001$). Sitewise diversification, identified by high throughput evolution and rational library design, improves discovery efficiency.

4.2. Introduction

Molecular recognition ligands are valuable tools in fundamental biology, medicine, and industrial biotechnology. Engineered ligands enable control over binding epitope, affinity, selectivity, and the biophysical properties of the ligand. Protein ligands are frequently engineered by modulating amino acids in a select region – known as the paratope – of a protein while conserving a stable underlying framework.[189] A variety of protein topologies have demonstrated efficacy as scaffolds for evolution of novel binding function including natural immune repertoires of antibodies[190] and variable lymphocyte receptors[191] as well as a multitude of synthetically-diversified scaffolds[189,192]. One particular example, the affibody domain, has been effectively used as a ligand scaffold including evolution of binding to numerous targets, with affinities as strong as 20 pM, and application to diagnostics, molecular imaging, and therapy.[193,194] The affibody is a 58-residue, three-helix bundle derived from the Z domain of *Staphylococcal* protein A. It is readily expressed recombinantly in bacteria, highly soluble, and reversibly unfolds with a wild type midpoint of 72ºC[195], although engineered mutants have exhibited destabilization to denaturation midpoints of 37-65ºC (mean: 49ºC)[195–200]. Mutants with novel binding activity have been

discovered and evolved from combinatorial libraries with diversity at 13 residues on one face of the N-terminal and middle helices. Each of the 13 sites was diversified to the 20 natural amino acids using broadly distributed NNK codons. Library screening has been performed using phage display[193,201,202], ribosome display,[203,204] and bacterial display[200,205–207]; yeast display was used for framework evolution[208].

Evolution of novel binding function necessitates mutation of sufficient paratope area to drive the new intermolecular interaction[209] while maintaining sufficient intramolecular stability. Mutation of intramolecularly-critical sites or mutation of semi-tolerant sites to suboptimal amino acids can limit evolutionary potential despite the introduction of an otherwise effective paratope.[21,67,210–213] Thus, identification of the mutational tolerance of each site – within the context of a diverse array of sequences possible within a combinatorial library – can aid evolution. Implementation of variable diversities at each site – both in entropy and specific amino acid preferences – has proven to enhance evolutionary efficacy in synthetic fibronectin domain libraries[51,57] and natural[58] and synthetic[53] antibody repertoires. Sitewise constraint has also been implemented in designed ankyrin repeat proteins[46,59,60] and fibronectin domain sheet libraries[214,215] using rational bias. Sitewise constraint has not yet been published for the affibody domain. Amino acid bias, across sites, has been implemented using amino acids frequently observed in protein-protein interfaces, particularly tyrosine and serine[47–51], as well as glycine in loop paratopes. Sitewise bias has been identified via natural antibody repertoire mimicry[56], wild-type constraint[51], structural analysis[46], and high-throughput evolution and deep sequencing[57]. Identification of detrimental mutations via deep scanning strategies has also been studied.[149,216–219] Computational prediction of mutational impacts on stability[42,44,45,220] and

functional maturation[221,222], though not functional discovery, have been extensively studied and could be used to guide library design.

The current study aimed to identify the sitewise amino acid diversities consistent with efficient evolution of a broad array of binding function (*i.e.* creation of a single library containing specific binders to a multitude of targets) and examine the drivers and implications of these amino acid preferences. We provide a platform for designing small protein libraries in terms of deciding which residues to include or avoid at individual positions. The effectiveness of this approach is demonstrated in the context of discovering high affinity variants from an affibody library.

4.3. Materials and Methods

4.3.1 Preliminary Library Design (First Generation)

As a preliminary attempt to identify the most beneficial diversification strategy for the three-helix affibody scaffold, a combinatorial library was designed which incorporated a wide variety of mutations across select sites. While, traditionally, the affibody scaffold is uniformly mutated at thirteen positions using all 20 natural amino acids, it is largely unknown which amino acids are most effective at any particular site. To better understand these functional diversities, fifteen solvent exposed sites throughout helix 1 and 2 (classic thirteen and sites E15 and I31) were mutated using five separate levels of diversity: i) wild-type (WT) residue, ii) WT or serine (small size and promotes neutral interaction[54]), iii) WT, serine, or tyrosine (frequently drives binding affinity and specificity[49,223]), iv) relaxed, moderate diversity (A, C, D, G, N, S, T, or Y), or v) full diversity mimicking the chemical composition of the third antibody heavy chain complementarity-determining region (CDR-H3). These five combinatorial libraries were separately constructed on the DNA level, then

pooled to form the first generation library. Any single variant within the initial library exhibited sitewise mutations from only one of the five combinatorial sub-libraries.

4.3.2 Gradient Sitewise (GS) Library Design (Second Generation)

The GS library was designed in a sitewise manner by balancing numerous data inputs. Numerous sites were constrained in their amino acid diversity (details in Supporting Information). Amino acid diversity at likely hot spot affibody positions was guided by amino acid prevalence in natural antibody interfaces (Abysis database; CDR-H3 diversity, Kabat sites 95-102) and previously evolved affibodies found in literature[193–195,197,199–203,206,224–247] as well as those generated in-house with the first generation library. The prevalence of each amino acid, except Gly and Arg, were calculated based on equally weighted CDR-H3 diversity via Abysis and previously evolved binder data, yielding the codon design which we call B* for the rest of the manuscript (Figure S1). The B_{LC}* codon closely mimics this design, while also emphasizing low cysteine content (0.5% in B_{LC}* vs 2.5% in B*).

4.3.3 Solvent Accessible Surface Area (SASA)

PDB files for affibody (PBD ID: 1H0T, 1LP1, 2B88, 2KZI, 2OKT, 3MZW) were processed using GetArea[137] to calculate SASA relative to random coil (probe radius, 1.4 Å). At each site, the median value across all six structures was reported. For PDBs that included multiple affibody chains, the mean SASA of the chains was determined prior to calculating median values among all structures.

4.3.4 Computational Stability

t each of the thirteen classically diversified sites (9, 10, 11, 13, 14, 17, 18, 24, 25, 27, 28, 32, 35) and for each naturally occurring amino acid, the change in stability ($\Delta\Delta G_{folding}$) upon mutation was calculated with FoldX[248]. First, an affibody structure was randomly mutated in accordance with the first generation library design to calculate baseline stability. Next, $\Delta\Delta G_{folding}$ was calculated for each single mutation (all 19 natural amino acids substituted, separately, into each site) versus the parental mutant. This process of saturation scanning was repeated for 312 randomly generated parental mutants across five affibody PDBs (PDB ID: 1H0T, 1LP1, 2B88, 2KZI, 3MZW). The median $\Delta\Delta G_{folding}$ was then calculated for each residue-site combination for the thirteen classically diversified sites. 312 iterations were demonstrated as sufficient for convergence of the (de)stabilization values.

4.3.5 Natural Homolog Analysis

Pfam[136] family B (PF02216) was accessed in April 2013 to retrieve 1,484 total sequences, 119 of which were unique. Rather than equally weighting all 1,484 sequences, the unique sequences were given a weight dictated by the square root of the number of occurrences for that sequence. Using this adjusted weighting, a sitewise amino acid distribution was calculated (Table S2).

4.3.6 Relative Helix Propensity

Empirical data from several published studies[249–254] were collectively used to calculate relative propensities of each natural amino acid within helical secondary

structures. The values from the previous studies were linearly averaged, based on destabilizing energetics of folding relative to glycine (propensity = 0) and alanine (propensity = 1). Using this normalized scale, the propensity of proline is calculated to be -2.7.

4.3.7 Library Construction

Each combinatorial library was built using synthetic oligonucleotides (IDT DNA) with degenerative codons, which were assembled by overlap extension PCR. Library genes were transformed into EBY100 *S. cerevisiae* yeast via electroporation[255] wherein the library fragments homologously recombined with linearized pCT vector to yield a construct that enabled yeast surface display of the encoded proteins[113]. Transformation efficiency was quantified using dilution plating with SD-CAA selective media. Sanger sequencing of initial libraries was performed for quality control.

4.3.8 Binder Selection

Yeast libraries were grown in SD-CAA selective media (16.8 g/L sodium citrate dihydrate, 3.9 g/L citric acid, 20.0 g/L dextrose, 6.7 g/L yeast nitrogen base, 5.0 g/L casamino acids) at 30 ºC with shaking at 250 rpm. Surface display was achieved by switching to SG-CAA media (10.2 g/L sodium phosphate dibasic heptahydrate, 8.6 g/L sodium phosphate monobasic monohydrate, 19.0 g/L galactose, 1.0 g/L dextrose, 6.7 g/L yeast nitrogen base, 5.0 g/L casamino acids) with incubation at 30°C with shaking at 250 rpm for 16 hours. Induced libraries were enriched for affibody variants that specifically bound each of the several protein targets (first generation: lysozyme and rabbit

immunoglobulin G (IgG); second generation: death receptor 5, transferrin, cytochrome C, glucose-6-phosphate dehydrogenase, CD276, MET and a G-protein-coupled receptor; separate enrichments performed in parallel) using both magnetic streptavidin coated bead sorting[256] and fluorescence activated cell sorting[257]. Each round of bead sorting consisted of two incubations (two hours each) of induced yeast with either bare beads or beads pre-incubated with an arbitrary biotinylated protein for depleting non-specific binding interactions. Following the depletions, the remaining yeast cells were incubated with beads with biotinylated target for two hours and washed twice with PBSA. Yields for both non-specific binding and target binding were quantified using serial dilution plating. Populations that demonstrated strong specificity (> 10-fold) of target binding relative to non-specific binding were isolated for sequence analysis. For all populations, after two rounds of enrichment using target labeled beads, flow cytometry was conducted. Preparation for cytometry consisted of incubating an induced yeast population with 100 nM biotinylated target and 67 nM of anti-c-Myc epitope tag antibody (9E10, Biolegend) in PBSA for 30 minutes at room temperature. Following the primary labeling step, cells were washed with PBSA, incubated with AlexaFluor647-conjugated goat anti-mouse antibody and AlexaFluor488-conjugated streptavidin for 15 minutes at 4°C, and then washed with cold PBSA. If target binding was observed, all binding events above background, denoted as high stringency binders, were isolated for sequence analysis. Ultra-high stringency binding populations (Figure S5-S7) were obtained through one additional round of cytometry sorting using 5 nM biotinylated target and a stricter sorting gate. In the event that no binding was detected, all full-length, c-Myc positive events were isolated for evolution. Plasmid DNA was zymoprepped from yeast, subjected to dual error-prone PCR

efforts on the full gene and on shuffled helices[258], and electroporated into yeast for additional iterations of bead sorting and flow cytometry.

4.3.9 Affinity and Specificity Analysis

Representative variants were randomly chosen from the high and ultra-high stringency populations (Figure S6A and Figure S6B, respectively). Each variant was induced in yeast, incubated with anti-c-Myc antibody and at various concentrations of their respective target, washed with cold PBSA, incubated with AlexaFluor647-conjugated goat anti-mouse antibody and AlexaFluor488-conjugated streptavidin for 15 minutes at 4°C, and then washed with cold PBSA. Extent of binding was measured via flow cytometry and normalized based on the maximum signal strength associated with each target and the background fluorescence of target-free yeast. Dissociation constants (K_D) were calculated by fitting the data to a two-state binding curve (n=3). Specificity was assessed by separately incubating each high stringency variant with multiple non-target biotinylated proteins at 100 nM or with target protein at 50 nM. The extent of binding, normalized by maximum possible fluorescence associated with each target, is then compared between target and non-target samples (Figure S8).

4.3.10 High-throughput Sequence Analysis

The plasmid DNA from the initial libraries as well as the enriched populations – both high stringency cytometry detectable and magnetic bead selective – were isolated for sequence analysis with a Zymoprep kit (Zymo Research). Illumina MiSeq adapter and indexing sequences were added via PCR. Paired-end (250 base pair) analysis yielded 18

x10^6 quality reads. Raw paired-end read output was groomed and assembled with PANDAseq using a quality score threshold of 0.99, then converted to FASTA[119]. FASTA sequence files were processed using ScaffoldSeq[259]. A sequence homology threshold of 80% was used for clustering similar variants. Clusters were then dampened using a factor of 0.25 to account for enriched sequence frequency while gaining information diversity from less frequent variants.

4.3.11 Library of Origin Analysis

Sequence variants from the initial and evolved populations were assessed for having originated from each of the three library designs. The probability of finding a particular sequence within each of the three sublibraries (NNK, GS, and GS_{LC}) is calculated as the quantity $P(k)_S$. This equation takes into account the amino acid, *i*, present at each position, *j*, of the sequence, *k*. The frequency of *i* at position *j* within the sitewise amino acid design of sublibrary S is given by $f_{\bar{s},i,j}$. The library origin probability, $P(k)_S$, is then obtained by taking the product across each of the 17 potentially diversified sites (Equation 1). Amino acids not offered within a particular sublibrary (e.g. Arg at site 11 within the GS design or Tyr at site 6 of the NNK design) were given a default value of 1 x10^{-4} rather than zero to account for random errors during library construction, evolution of binders, and DNA sequencing.

$$P(k)_S = \prod_{j=1}^{17} f_{\bar{s},i,j} \tag{1}$$

Each sequence was then binned with the sublibrary that yielded the greatest library origin probability. Within each sublibrary bin, the number of occurrences, *n*, of each unique

sequence, *v*, was tallied and square rooted to normalize the contribution of dominant variants. The total sequence count, N_S, within each of the three sublibraries is the summation of normalized counts for each unique variant (Equation 2). Rare variants, having been observed fewer than ten times, were excluded from the analysis.

$$N_S = \sum_S \sqrt{n_v} \tag{2}$$

4.3.12 Stability Measurements

Individual variants were randomly selected from cytometry detectable binding populations. Each gene was shuttled into a pET-24 derivative to include a C-terminal six-histidine tag, expressed in BL21(DE3) bacteria, and purified using cobalt resin columns[57]. Purified proteins were diluted in PBS to 1 mg/mL and assessed via circular dichroism using a Jasco J815 instrument. Temperature scans were conducted using a range of 20-90°C (1°C/min) while monitoring the 220 nm wavelength. Midpoints of thermal denaturation were calculated assuming a standard two-state unfolding curve. For reducing conditions, dithiothreitol (DTT) was prepared fresh and added to the samples to yield a final concentration of 2 mM DTT, then incubated at room temperature for 30 minutes.

4.4. Results and Discussion

4.4.1 First Generation Library

We sought to identify sites, within the typically diversified paratope, that would provide the most substantial benefit from constrained diversity. The main approach to identify sites, and their amino acid preferences, was high-throughput ligand engineering

and deep sequencing analysis. Thus, we designed, constructed, and screened a combinatorial library of affibody domains to discover and evolve functional variants. With an aim to improve the binder frequency to generate a diverse set of functional sequences, modest constraint at select sites was introduced in the initial library. Evaluation of the solvent-accessible surface area (Table 1) and spatial orientation at the 13 classically diversified sites revealed three sites that were hypothesized to provide more evolutionary benefit from at least partial wild-type constraint. Q9 is only 19% accessible to solvent (Figure 1, orange). N11 is 51% accessible but predominantly facing away from the evolved binding surface (Figure 1, yellow). R27 is at the core of the paratope but only 27% accessible (Figure 1, blue). The other diversified sites range from 33-90% accessible (median: 71%). Two sites that were previously conserved, E15 and I31, were modestly accessible (38% and 27%) and oriented near the evolved binding surface (Figure 1, green and violet). Mild diversity of these sites was hypothesized to provide evolutionary benefit.

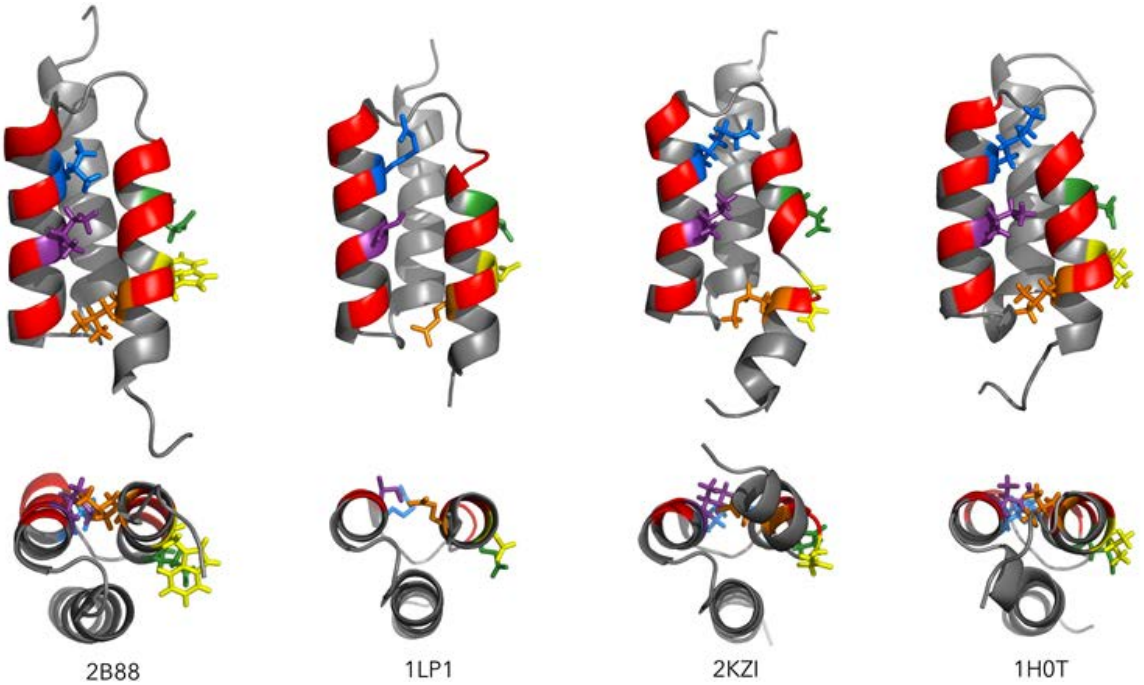


Figure 4-1. ***Affibody structure with constrained sites highlighted.*** **Side chains are shown for Q9 (orange), N11 (yellow), E15 (green), R27 (blue), and I31 (violet) in four solved**

affibody structures[232,236,244,260]. The other ten classically diversified sites are shown in red.

Thus, a combinatorial library was constructed from five sub-libraries in which these constrained sites were varied from wild-type conservation to full amino acid diversity (Table 1). The other ten typical paratope sites were broadly diversified to all 20 amino acids using complementarity-biased amino acid frequencies[47–51]. The library, constructed by overlap extension PCR of degenerate oligonucleotides and homologous recombination in a yeast display system, contained $4x10^8$ variants. The library generally matched the intended design (median absolute deviation from design, $|f_{observed} - f_{design}| = 0.8\%$) although tyrosine and cysteine were modestly higher than desired while alanine and aspartic acid content were lower than designed (Figure S2A).

Table 4-1. ***First generation library design.*** **The site, wild-type (WT) amino acid, relative solvent-accessible surface area (SASA), and diversity in the published and first generation libraries are presented for sites in the N-terminal and middle helix of affibody.**

Site[a]	WT	SASA(%)	Previous Libraries[b]	1st Generation Library[c,d]
N6	A	78	N	A
K7	K	79	K	K
E8	E	39	E	E
Q9	Q	19	20	N / NS / NSYT / 12 / 20*
Q10	Q	66	20	20*
N11	N	51	20	N / NS / NSYT / 12 / 20*
A12	A	4	A	A
F13	F	33	20	20*
Y14	Y	72	20	20*

E15	E	38	E	E / EA / DSYA / 12 / 20*
I16	I	0	I	I
L17	L	10	20	20*
H18	H	67	20	20*
E24	E	88	20	20*
E25	E	90	20	20*
Q26	Q	23	Q	Q
R27	R	27	20	R/RS/RSYCHP/ 12/ 20*
N28	N	71	20	20*
A29	A	74	A	A
F30	F	10	F	F
I31	I	27	I	I / IS / ISYCFN / 12 / 20*
Q32	Q	72	20	20*
S33	S	44	K	A
L34	L	1	L	L
K35	K	76	20	20*
D36	D	66	D	D

a. Sites 1-5 and 37-58 were conserved as wild-type (Table S1).

b. “20” represents a mixture of all twenty amino acids using NNK degenerate codons.

c. “20*” denotes a mixture of all 20 amino acids weighted based on frequency in antibody repertoire but with reduced glycine because of the helical structure (Figure S1).

d. “/” separates sub-library designs

Specific binders to hen egg lysozyme and rabbit IgG were discovered and evolved from the combinatorial library using yeast display with magnetic and flow cytometry selections. Deep sequencing of the evolved variants yielded $6x10^4$ unique sequences in 523

diverse families. Numerous amino acids, at specific sites, exhibited substantial enrichment or depletion from the unselected to the evolved populations, which is indicative of the benefit or detriment of that amino acid at that site in functional affibodies (Figure 2). These data inform effective design of an improved combinatorial library.

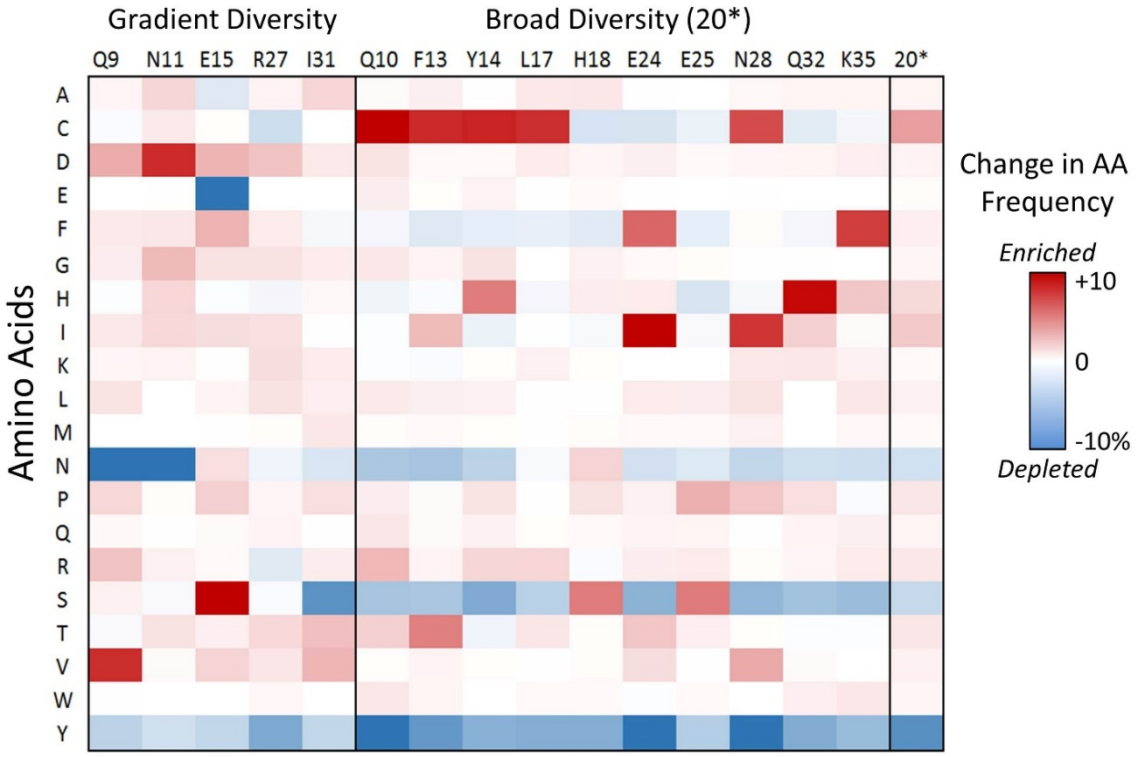


Figure 4-2. ***Sitewise amino acid preferences in the context of the first generation library.*** **The change in amino acid frequency, between the unselected and evolved populations ($f_{evolved} - f_{unselected}$), is shown for each amino acid at each site. 20* represents the aggregated frequencies of the 10 sites with broad diversity (20* in Table 1). $6x10^4$ unique evolved sequences were used.**

4.4.2 Second Generation Library Design: Overall

A second generation library was designed to further evaluate sitewise preferences in the context of a more focused design. On a sitewise basis, amino acids were favored that (i) appeared frequently in binder sequences from the first generation library (Figure 2) and 345 published binder sequences (Figure 3A); (ii) are computationally predicted to be stable

in the context of diverse paratopes (Figure 3B); and (iii) occur naturally in affibody homologs (Figure 4A). Broad diversity was favored in sites that are (i) solvent accessible and oriented towards the proposed binding interface (Table 1, Figure 1); (ii) broadly diverse in natural homologs (Figure 4B); and (iii) computationally predicted to be stable to multiple mutations (Figure 3B). Generally at all sites, amino acids were favored that (i) appear frequently in antibody CDR-H3 (bioinf.org.uk/Abysis; Kabat sites 95-102) (Figure 5A), which has also been implemented in previous synthetic libraries[47–51]; and (ii) favor helix formation[249–254] (Figure 5B). Each undiversified framework position, where mutational diversity is unlikely to provide added quality to the overall library design, was conserved as wild-type in the context of the optimized affibody framework[196]. Library design details are provided in the Supporting Information (Table S3). Commentary on the broadly diversified distribution, cysteine content, and non-traditionally diversified sites follows.

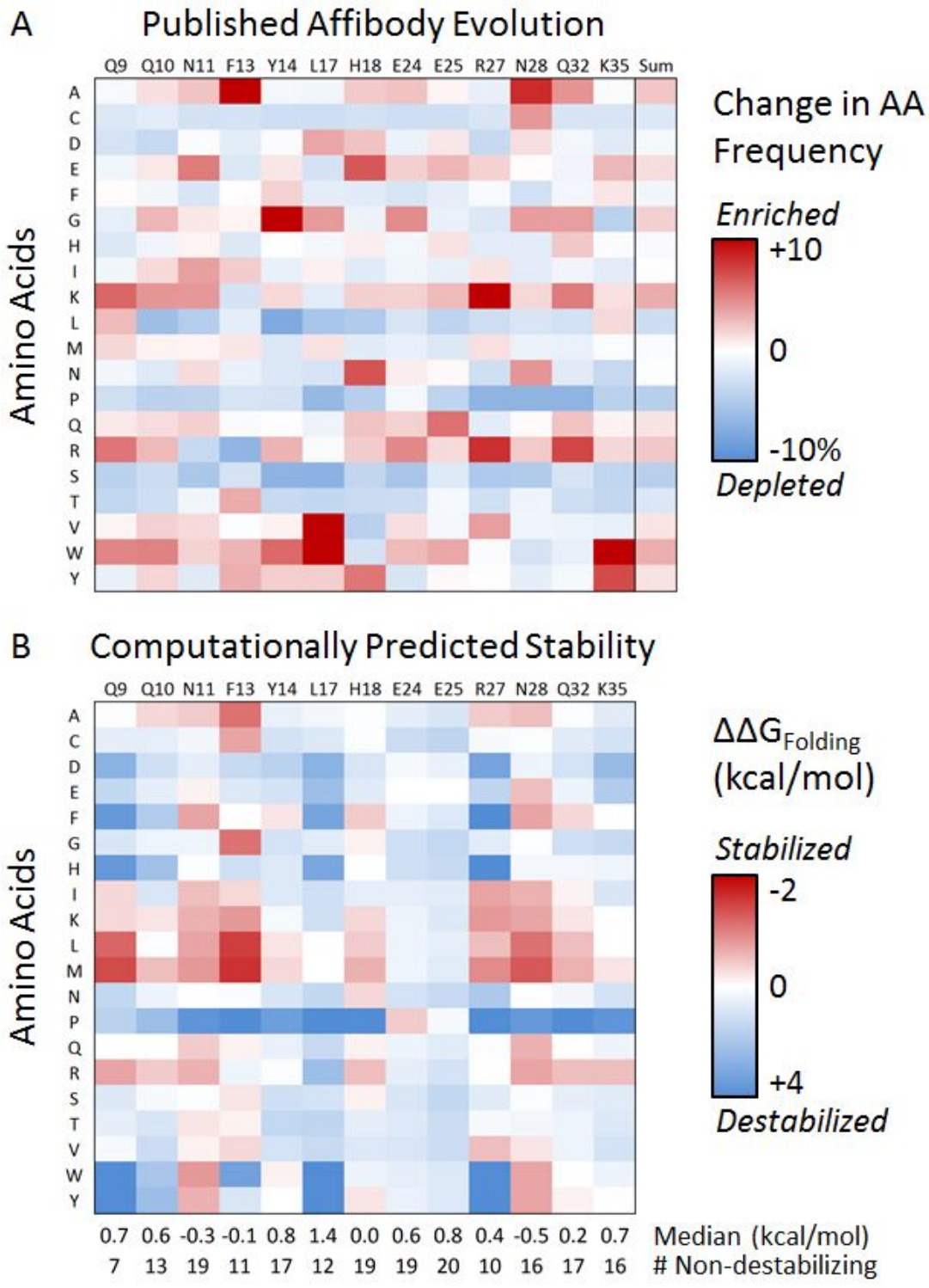


Figure 4-3. (A) ***Sitewise amino acid preferences from published affibody evolution.*** **The change in amino acid frequency, between the unselected and evolved populations ($f_{evolved} - f_{unselected}$), is shown for each amino acid at each of the thirteen traditionally mutated sites. 345 unique evolved sequences were used from numerous references.**[75,193–195,197,199–203,206,224–228,230–247] **(B)** ***Computed destabilization upon mutation in the context of diverse paratopes.*** **The median change in folding free energy ($\Delta\Delta G_f$)**

upon mutation to the indicated amino acid at the indicated site was computed using FoldX. Saturation scanning (calculation of all mutants) was performed at each site for 312 random library variants with five affibody structures. For each site, the median destabilization and number of tolerated amino acids (mutants with $\Delta\Delta G_f$ < 1.5 kcal/mol) are also presented.

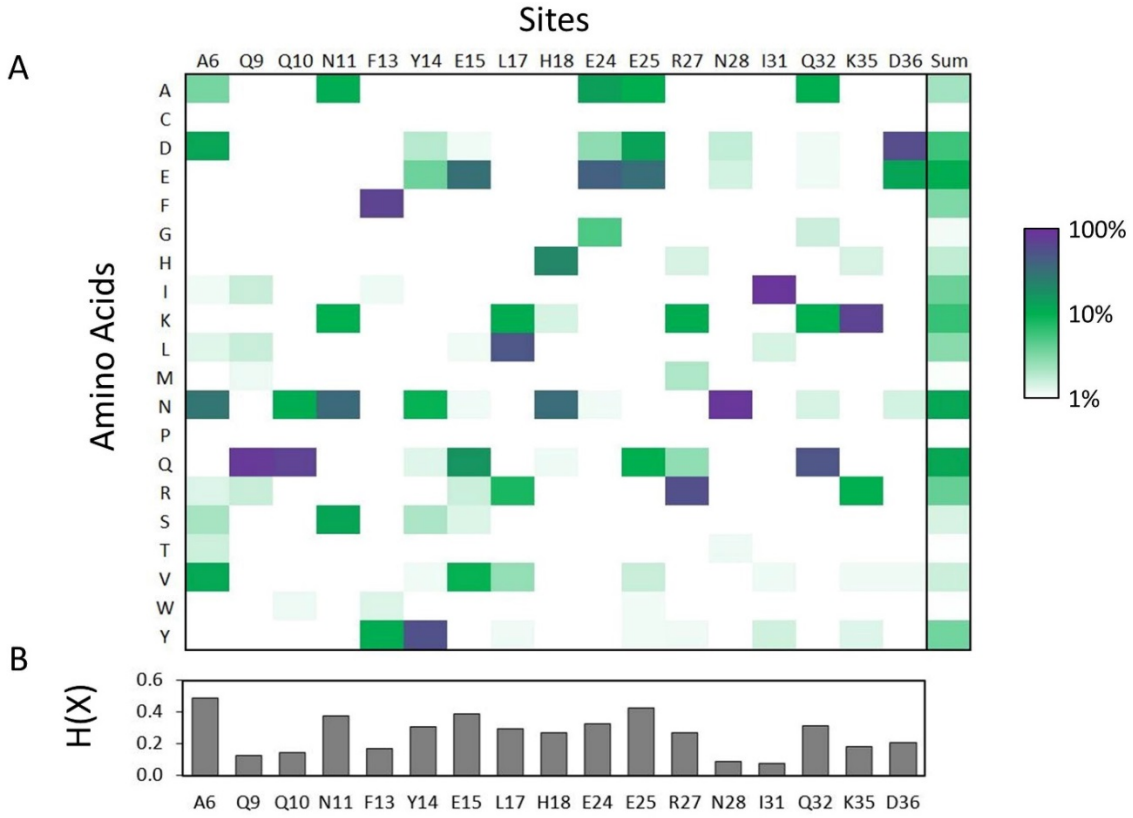


Figure 4-4. ***Sitewise amino acid preferences from affibody homologs.*** **(A) The amino acid frequencies of 1,484 (119 unique) proteins from Pfam[136] family B, PF02216, which are homologous to the affibody sequence, are shown for each amino acid at each site. (B) The Shannon entropy, $H(X) = \sum -f_{aa}\, log_{20}\, f_{aa}$, of each site is indicated.**

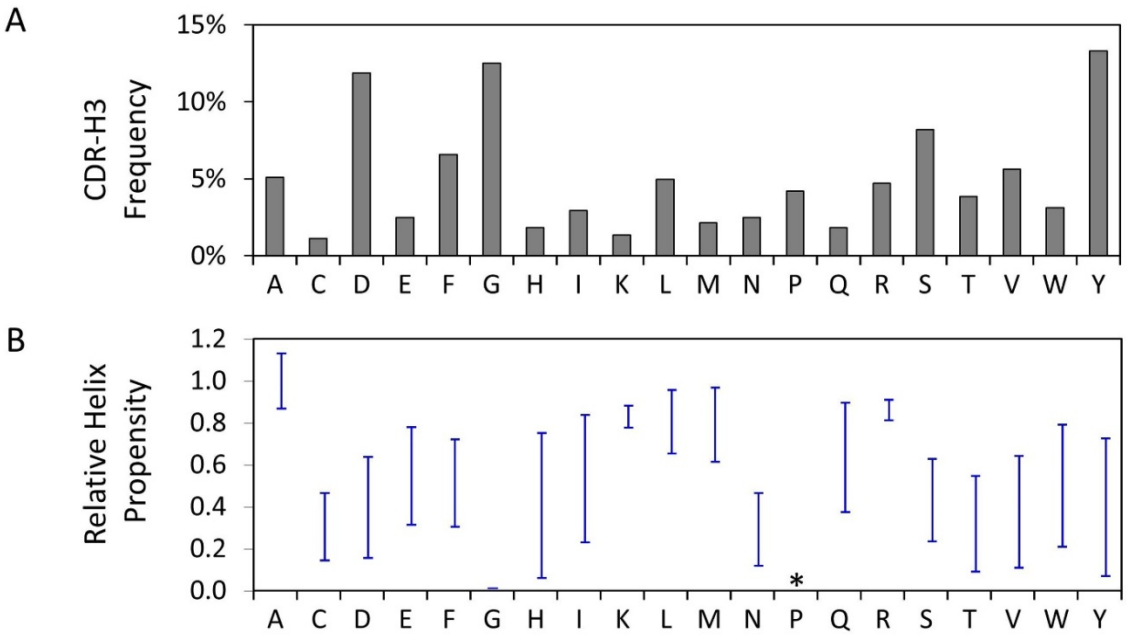


Figure 4-5. ***Aggregate amino acid preferences from natural proteins.*** **(A) The amino acid frequencies observed throughout the third complementarity determining region of antibody heavy chains (CDR-H3; Kabat sites 95-102). (B) The relative helical propensity of each amino acid, based on observed destabilization of helical secondary structure when used for substitution, is calculated as the aggregate of several previous studies[249–254]. The ranges indicate one standard deviation above and below the mean for each residue. Helix propensity of proline (*), calculated to be -2.7, is outside the presented range.**

4.4.3 Second Generation Library Design: Broadly Diversified Sites

At sites that were predicted to benefit from broad diversity, a biased distribution was implemented based 50% on mimicking antibody CDR-H3 diversity and 50% on mimicking evolved affibody sequences in the heavily diversified sites (Figure 2, *20** column). There are two exceptions to this balanced design: glycine and arginine. Glycine is frequently observed (12%) in antibody CDR-H3, predominantly for conformational flexibility[52–54], but does not generally support helical structure (Figure 5B) and was not enriched in functional variants from the first generation library (Figure 2). Thus, glycine content was set based on equally weighting frequency in homologous proteins (2%) and

forecasted frequency in enriched binders (7%, Figure S1), yielding 5% after renormalization. Arginine has been shown to correlate with non-specific binding when moderately present at binding interfaces[54,55]. Thus, although arginine was modestly enriched in published binders (9% naïve to 12% evolved, though not all studies included counter-selections for specificity) and slightly enriched in binders from our first generation library (2.6% to 3.5%), arginine content in the second generation library was restricted to its frequency in affibody homologs: 3%. This broadly diverse distribution, denoted as *B** (broad), was designed to facilitate high-affinity, specific binding interactions as well as accommodate for stable helices.

4.4.4 Second Generation Library Design: Cysteine

Cysteine content in the first generation library was higher than designed (Figure S2A) and was strongly increased in evolved binders (9% to 18%, $p < 0.001$) at five broadly diversified sites – 10, 13, 14, 17, and 28 – and maintained or reduced at all other sites (Figure 6A). Notably, variants with zero or one cysteine were depleted ($p < 0.001$) in evolved binders whereas variants with at least two cysteines were either maintained or enriched ($p < 0.001$, Figure 6B). These sites of cysteine enrichment are spatially clustered (Figure 6C) and several pairs exhibit substantial epistasis (Figure 6D). The variants having two cysteines predominantly came from the lysozyme binding population with cysteines at sites 10 and 28 across four families. Variants with four cysteines were mostly generated from the rabbit IgG binding population with cysteines at sites 10, 13, 14, and 17 across 57 families (Figure S3A). The importance of potential disulfide formation was evaluated by assessing thermal stability in oxidizing and reducing environments. Four evolved variants, each with unique cysteine locations, exhibited substantially greater thermal stability in the

oxidized state (Figure 6E, Figure S3B). Collectively, these observations are consistent with the evolutionary benefit of disulfide bonds.

It is possible that cysteine enrichment resulted from a need for enhanced stabilization in light of the extent of diversification including newly varied sites 15 and 31. This idea is supported by analysis of the subset of evolved binders with wild-type E15 and I31 in which variants with 0 or 1 cysteine are not depleted in the functional population ($p = 0.16$, Figure 6B). Morevoer, in previously published binders from NNK libraries, which conserve E15 and I31, only site 28 exhibited cysteine enrichment (Figure 3A). We aimed to evaluate the ability to evolve affibody ligands with limited cysteines, both to require intramolecular stability in the absence of disulfide bonds and because cysteine-containing variants may hinder production in E. coli, increase likelihood of aggregation or oligomer formation during purification, and complicate the use of thiol chemistries in downstream applications. Thus, in the second generation library experiments an additional broad distribution, B_{LC}*, was tested that has lower cysteine frequency (0.5%).

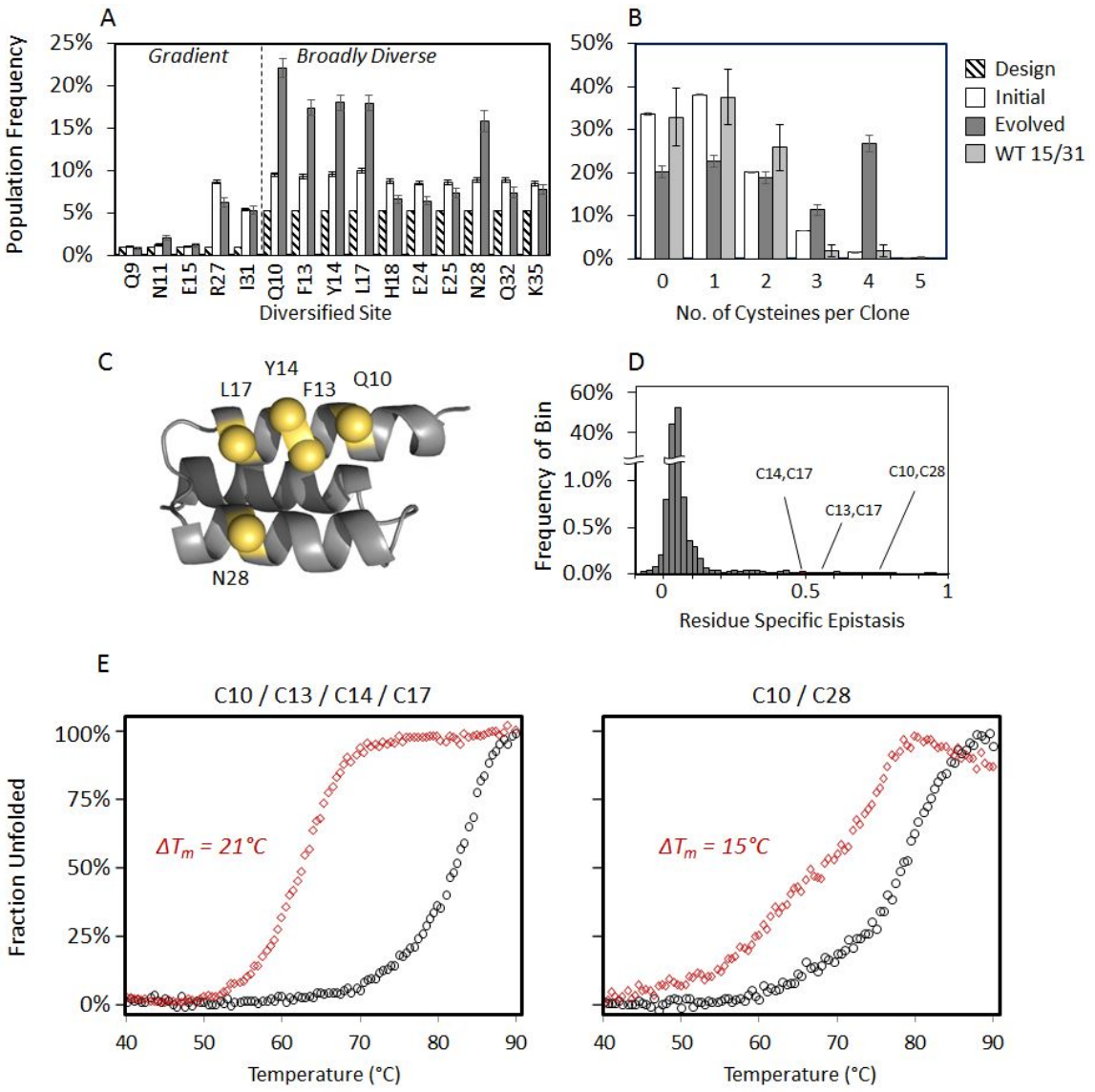


Figure 4-6. (A) Cysteine content in the first generation initial library (solid, white bars) was higher than intended (striped, light bars) and was strongly increased within evolved variants (solid, dark bars) at five broadly diversified sites. (B) The propensity of cysteine(s) within single variants from the first generation initial and evolved populations as well as a subset of evolved variants where wild-type residues are observed at positions E15 and I31 (light gray bars). (C) Cysteine rich positions observed within the first generation evolved variants are labeled and highlighted in gold. (D) Residue specific epistasis[74] was quantified for every observed combination of amino acid and position. The three most prominent cysteine pairs within the first generation evolved variants are indicated. (E) The change in thermal denaturation midpoint (ΔT_m) between reducing (red) and non-reducing (black) conditions is shown for two high affinity variants with the most prevalent cysteine pairings (indicated above each plot).

4.4.5 Second Generation Library Design: Broadened Paratope

Sites 6, 15, 31, and 36 are not traditionally diversified, but natural sequence diversity in affibody homologs (Figure 4A), computational mutational tolerance (Figure 3B), solvent accessibility (Table 1), and amino acid frequency in evolved affibody sequences (Figure 2) inform mild diversification at these sites. Natural diversity at site 6 is high (Shannon entropy: 0.49, Figure 4). Mutations to E and A in the context of a HER2 binder were not destabilizing[196]. The site is exposed to solvent and could be expected to interact with target in many cases; in fact the HER2 ligand mutants impacted affinity. Thus, mild diversity (NSYT) was tested. Site 15 is naturally diverse (Shannon entropy: 0.39) and tolerant of mutation (15 residues), but is pointed away from the paratope and only 38% accessible to solvent. Despite 55% E on natural homologs, biased E was strongly depleted in binding populations (38% to 18%) from the initial library. Wild-type Q (30% in homologs) is conserved in published libraries. S is present naturally (1%) and enriched in binders (5% to 15%). V (10% naturally and enriched 3% to 5% in binders) will also be considered. Thus, QSEV diversity was allowed. I31 is conserved in previous libraries and naturally (96% I; Shannon entropy: 0.08) and is predicted to be poorly tolerant of mutation (only eight tolerant residues). Yet it is in the center of the planned paratope and tolerated diversity in the first generation library analyses. I was maintained at high levels (33% to 32%) but most other amino acids were also tolerated with the most significant depletions being S (18% to 10%) and Y (11% to 8%). A 30:70 mixture of I and the B* codon was used. D36 is conserved in previous libraries and naturally (80% D, 18% E, Shannon entropy: 0.20). Yet it is solvent accessible (66%), could be expected to contact target in some cases, and is computationally predicted to be tolerant to mutation. Diversity could be

beneficial both inter- and intra-molecularly although such diversity should be minimal to limit detriment. DN diversity was used. Collectively, these mild diversifications were evaluated within the second generation library.

4.4.6 Selections and Evolution from Second Generation Library

We aimed to evaluate the discovery and evolutionary efficacy of sitewise designs. Thus, we competed the second generation library design, GS, versus the traditional NNK library with broad, near-uniform diversity. We also included the modified second generation library, GS_{LC}, which has reduced cysteine content as well as modifications in the genetically coupled amino acid preferences (B_{LC}*, details below). The second generation libraries (Table S1 and Figure S1) were synthesized with custom degenerate oligonucleotides assembled by overlap extension PCR and introduced into the yeast display system by homologous recombination. $1x10^9$ variants were achieved for each of the three library designs. High-throughput sequence analysis of each library revealed the expected distributions on an amino acid basis (median absolute deviation from design, $|f_{observed} - f_{design}| = 0.5\%$; Figure S2B-D). To evaluate generalizable discovery/evolutionary efficacy, the pooled libraries were used to identify binders to a broad panel of seven new targets: death receptor 5, transferrin, cytochrome C, glucose-6-phosphate dehydrogenase, CD276, MET, and a GPCR. Specific binders were discovered and evolved from the combinatorial library using yeast display with magnetic and flow cytometry selections (Figure S5B). Affibody variants discovered and evolved from the second generation library exhibited affinities of 2 ± 2 to 82 ± 23 nM (Figure S6A) with high target specificity (Figure S8). Deep sequencing of the evolved variants yielded $6 x10^4$ unique protein sequences in $5 x10^3$

diverse families, where families were dictated by having at least 80% homology throughout the library positions.

The study aimed to identify site-specific amino acid preferences consistent with evolutionary efficacy for a broad array of epitopes. Thus, multiple targets – each with numerous potential epitopes – were used for binder discovery and evolution. While target-specific – and more importantly, epitope-specific – preferences may be present, the diversity of evolved sequences ($6x10^4$ unique sequences in $5x10^3$ families) mitigates the impact of such biases on the broad analysis.

4.4.7 Library Efficacy Comparison

The library origin of the evolved binders was determined by probabilistic sequence analysis after clustering into families based on 80% similarity (Materials and Methods). Variants from the GS library were enriched whereas variants from the NNK library were depleted relative to the initial library ($p < 0.001$, Figure 7), which demonstrates improved evolutionary efficiency of the sitewise amino acid preferences in the GS library design. This evolutionary benefit of the GS design is even more striking (Figure S7) in a population that was further enriched for even stronger affinity binding (0.9 ± 0.1 to 5 ± 2 nM affinities; Figure S6). These results summarize the collective advantage of sitewise constraint in evolutionary discovery. Additional analyses and experiments, which follow, were performed to further elucidate the molecular aspects of this advantage.

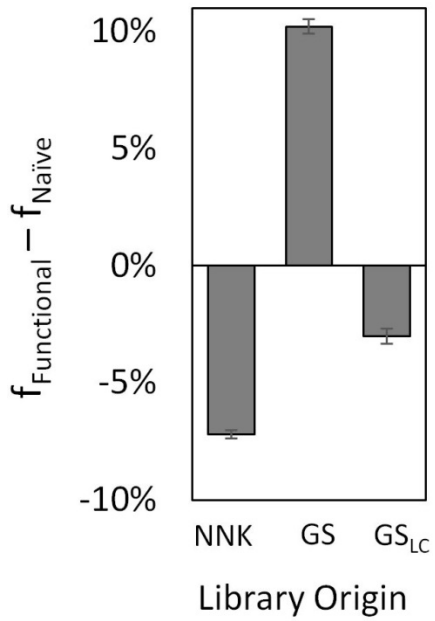


Figure 4-7. ***Enrichment of second generation libraries.*** **Sequences were analyzed for the likelihood of originating from either the NNK, GS, or GS_{LC} library designs. The change in prevalence (frequency in functional population – frequency in naïve library) between the initial library and binding populations is shown. Significant enrichment of the GS library was observed at the expense of depletion of both NNK and GS_{LC} ($p < 0.001$ for GS vs. each).**

GS was also superior to GS_{LC}, which only differed in the codon design at the nine broadly diversified sites. Thus, the evolutionary preference for the B* codon design over the B_{LC}* codon was examined to elucidate relative efficacies of each amino acid. As noted, the motivation to create B_{LC}* was to reduce cysteine content (2.5% to 0.5%) to isolate disulfide-free and thiol-free binders. Unlike the first generation library, in the context of additional amino acid diversity constraint in the second generation library, the cysteine content is depleted during evolution (Figure 8). In particular for the GS library, variants with two or more cysteines are dramatically reduced (59-fold for two-cysteine variants and 51-fold for three-cysteine variants). Cysteine depletion is observed during evolution of binders from the NNK library but to a reduced extent (2- and 28-fold for two- and three-cysteine variants, respectively) relative to the sitewise constrained library.

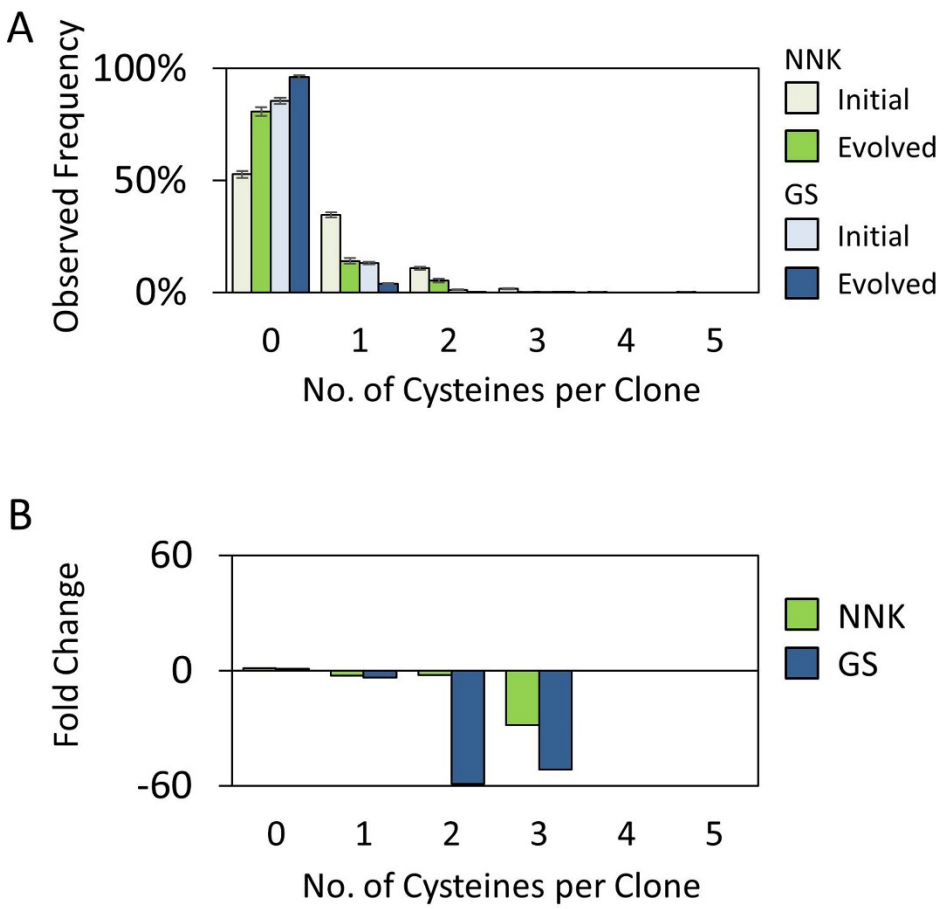


Figure 4-8. ***From the second generation libraries, both the NNK and GS designs yielded fewer cysteines in the evolved populations compared to the initial libraries.*** **The absolute (A) and relative (B) changes in frequency are shown. These modest levels were even slightly depleted among evolved variants.**

Via genetic code coupling, cysteine reduction also results in reduction of F, S, L, Y, W, G, and R; the aim for the antibody- and first generation affibody-inspired amino acid distribution also results in increase of D, E, H, K, N, and Q (Figure 9A). Of the amino acids with higher frequency in the initial B* codon sites vs. B_{LC}*, W and L are enriched and F and Y are maintained at high levels upon binder evolution (Figure 9BC). Conversely, S, G, R, and C are depleted. This is consistent with evolutionary benefit of these aromatics and hydrophobics leading to superiority of B* despite the evolutionary inefficiency of S, G, R, and C. Notably, the results support the aforementioned rationalized reduction of G

and R relative to antibody CDR-H3. Of the amino acids with lower frequency in the initial B* codon sites vs. B_{LC}*, H, K, N, and Q are reduced or maintained at low levels. Conversely, E is enriched and D is maintained at a high level. These results suggest that depleted H, K, N, and Q also contribute to B* superiority vs. B_{LC}* whereas depleted E hinders B*.

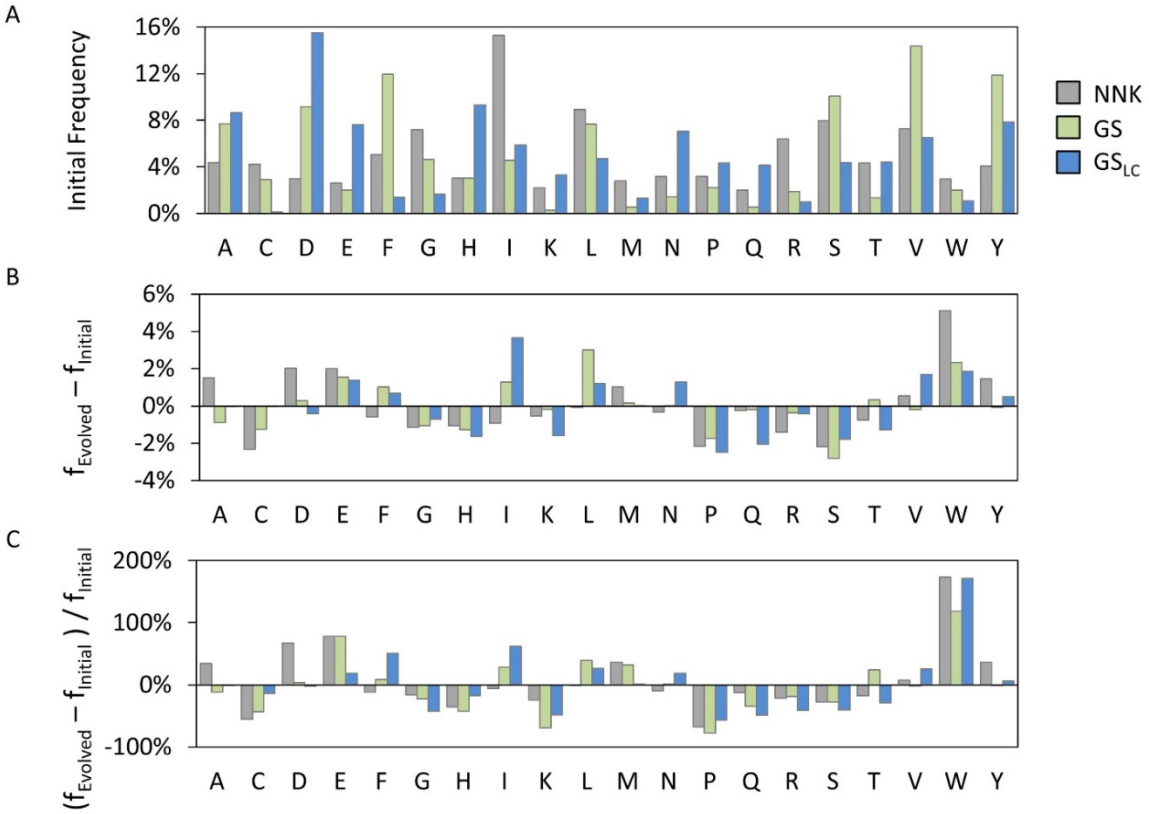


Figure 4-9. ***Nine sites were offered at least 50% GS/GS$_{LC}$ diversity (Q10, Y14, L17, H18, E25, N28, I31, Q32, K35).*** **A comparison between the initial and evolved amino acid frequencies at these sites is shown. (A) Individual frequencies for each of the three sub-libraries. (B) Absolute and (C) relative changes in frequency.**

4.4.8 Sitewise Amino Acid Frequencies

Sitewise amino acid frequencies (Figure 10) provide valuable insight to elucidate the evolutionary benefit of the constrained GS design as well as to guide further refinement. P, Q, K, and C are consistently depleted at broadly diversified sites in evolved binders. S

is depleted at all but site 18. The aforementioned benefit of W is especially observed at sites 14 and 17, adjacent sites on the 'top' of helix 1, and is also mildly enriched at other sites. Two other residues that benefit B* relative to B_{LC}*, F and L (and homolog I), are generally enriched in helix 2.

At site 9, enrichment of W (16% to 31%) and L (12% to 21%), as well as maintenance of large hydrophobic M (5%), in evolved binders demonstrates the value in constraining diversity to increase the initial frequencies of these residues within the GS library. Yet depletion of the remainder of the constrained subset suggest further constraint to WLM could be advantageous to both increase the frequency of these beneficial residues and decrease the frequency of detrimental options, especially cationic R (17% to 5%) and K (7% to 2%). At site 11, which is relatively exposed (51% SASA), the strong hydrophilics D and N are enriched whereas the mid-hydrophilics S, Y, T, and A are depleted. Thus, the GS design benefits from D and N bias but is hindered by bias to S, Y, T, and A. At site 13, aromatics (F and Y) and hydrophobics (F, I, and V from already elevated initial frequencies; and L from rare mutagenic PCR (0.1% to 0.7%)) were enriched in evolved binders whereas hydrophilics (D, N, S, and T) and the small A were depleted. Further hydrophobic constraint of this modestly buried site (27% SASA) to FLIV would likely provide additional benefit. Site 24 (88% SASA) generally tolerates its relatively broad diversity. The reasonably buried (27% SASA) site 27, benefits from bias to hydrophobic I and V, which are strongly enriched in evolved binders, but is hindered by bias to R, K, M, and L, which are depleted upon evolution.

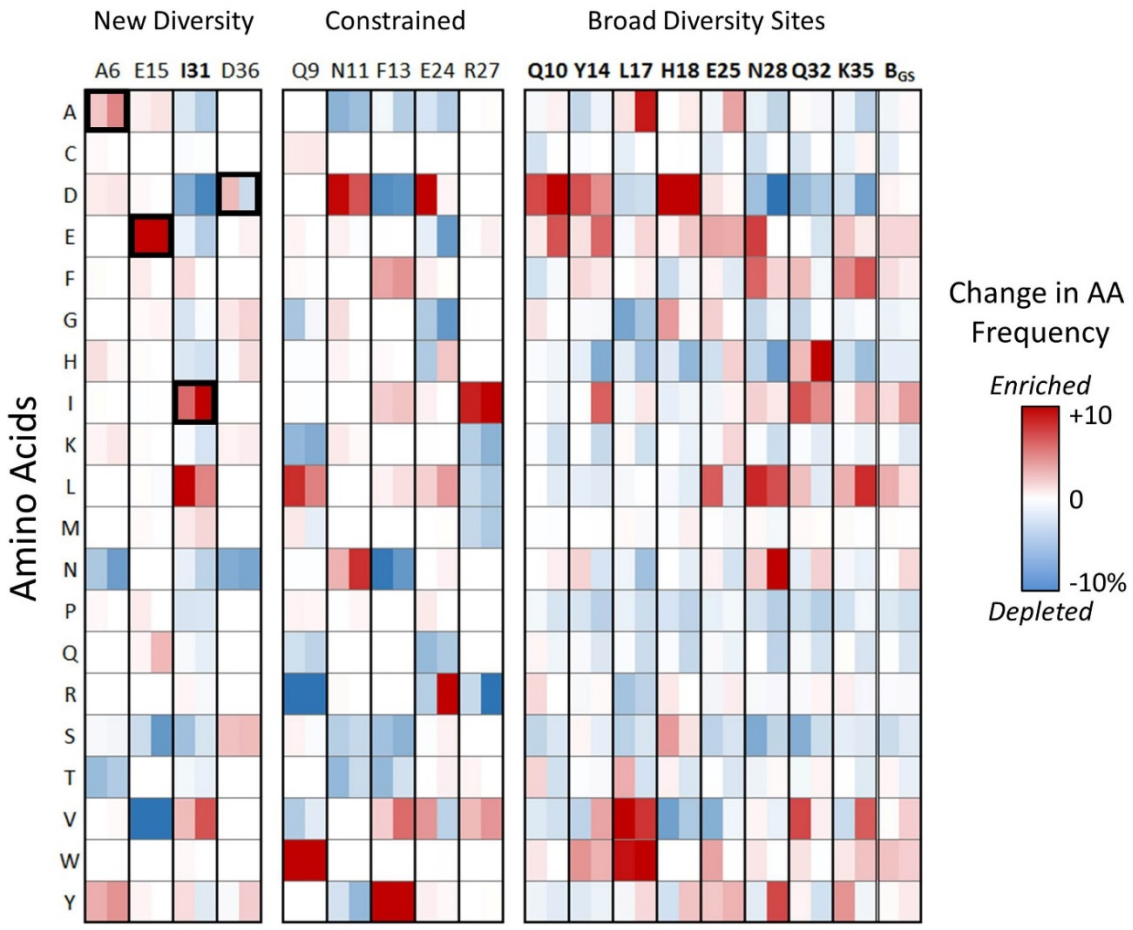


Figure 4-10. ***The changes in frequency between the initial and evolved populations are shown for the GS (left side of each column) and GS$_{LC}$ (right) design campaigns.*** **Bolded sites indicate the nine broadly diversified positions. These nine sites are averaged in the B$_{GS}$ column on the right.**

Strong wild-type enrichment is observed at several of the newly diversified sites (Figure 10 and Figure 11A). At site 15, wild-type E enrichment is countered by depletion of S and V while wild-type homolog Q is slightly enriched. At site 31, in addition to wild-type I, homolog L and similar hydrophobe V are enriched while numerous amino acids are depleted. Site 36 reveals a more modest shift in its wild-type, D, along with significant depletion of the alternative mutant offering, N. At site 6, the Z-domain wild-type N is depleted, as well as T; S is maintained, and Y is enriched. Notably, alanine, conserved in

the first generation library because of its benefit in the optimized affibody backbone[196] and able to appear in second generation via mutagenic PCR and homologous recombination with linearized vector encoding for A, is enriched. As a result of these wild-type preferences, variants with mutation at three or four of these newly diversified sites are depleted (Figure 11B). Interestingly, though wild-type conservation is modestly preferred at site 36, mutation is strongly enriched in evolved binders if it is the only mutation within this set of newly diversified sites (Figure 11D) or it occurs in tandem with mutation at site 6 (Figure 11E); but co-mutation of site 36 with site 15 or 31 is depleted in binders. Dual mutation of 15 and 31, with or without further mutation at sites 6 or 15, is strongly depleted. Overall, in evaluation of the newly created diversity at sites 6, 15, 31, and 36, mild diversity is tolerable, but further constraint from the second generation design would benefit evolutionary efficiency. The ability of the GS library to outperform NNK despite the detrimental over-diversification of these sites indicates significant evolutionary value to the other modifications: B* codon biased diversity rather than NNK at broadly diversified sites and constrained diversity at sites 9, 11, 13, 24, and 27. Analysis of the traditionally diversified 13 sites in functional variants reveals 13-fold enrichment of variants of constrained design (Figure 12).

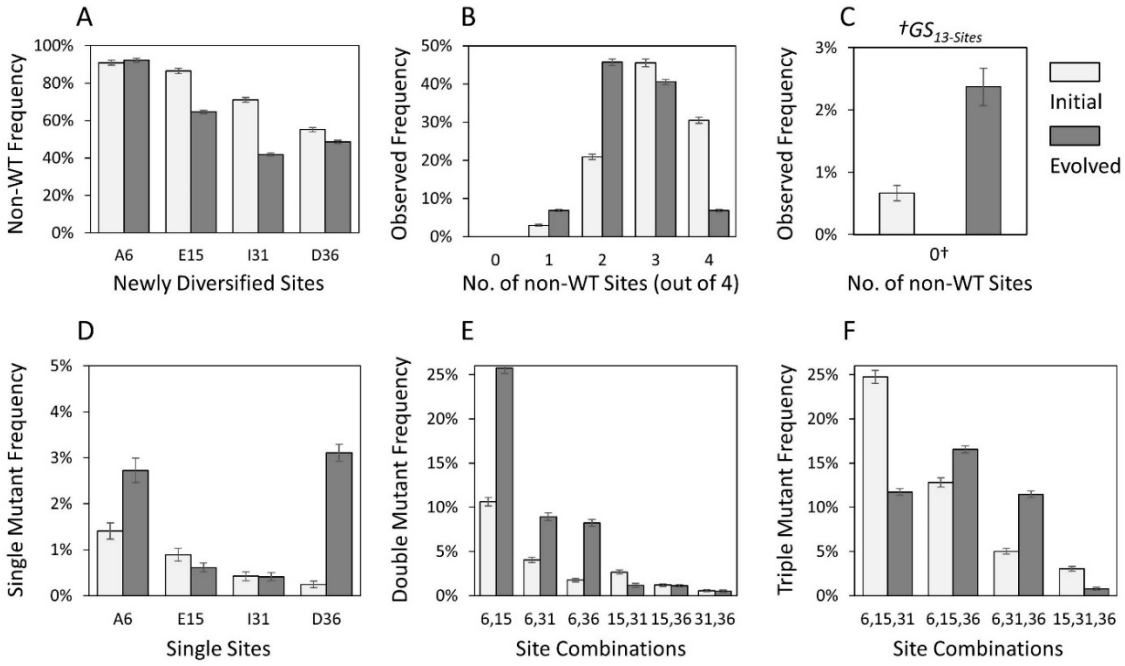


Figure 4-11. ***Among the four newly diversified sites, the tendency for wild type conservation or diversification is shown.*** **(A) The frequency of variants having a residue other than wild type at each site. (B) The distribution of variants from either the initial or evolved populations having exactly 0, 1, 2, 3, or 4 non-wild type amino acids. (C) Frequency of clones having WT at all four newly diversified sites. Frequency of (D) single, (E) double, and (F) triple mutants exclusively at the position(s) listed on the x-axis. The sequences analyzed in panels A, B, D, E, and F all meet the requirement of originating from the GS library based on probabilistic calculations at all 17 diversified sites. However, the sequences contributing to panel C meet the requirement of originating from the GS library based on probabilistic calculations at only the thirteen traditionally mutated sites (†$GS_{13\text{-}Sites}$).**

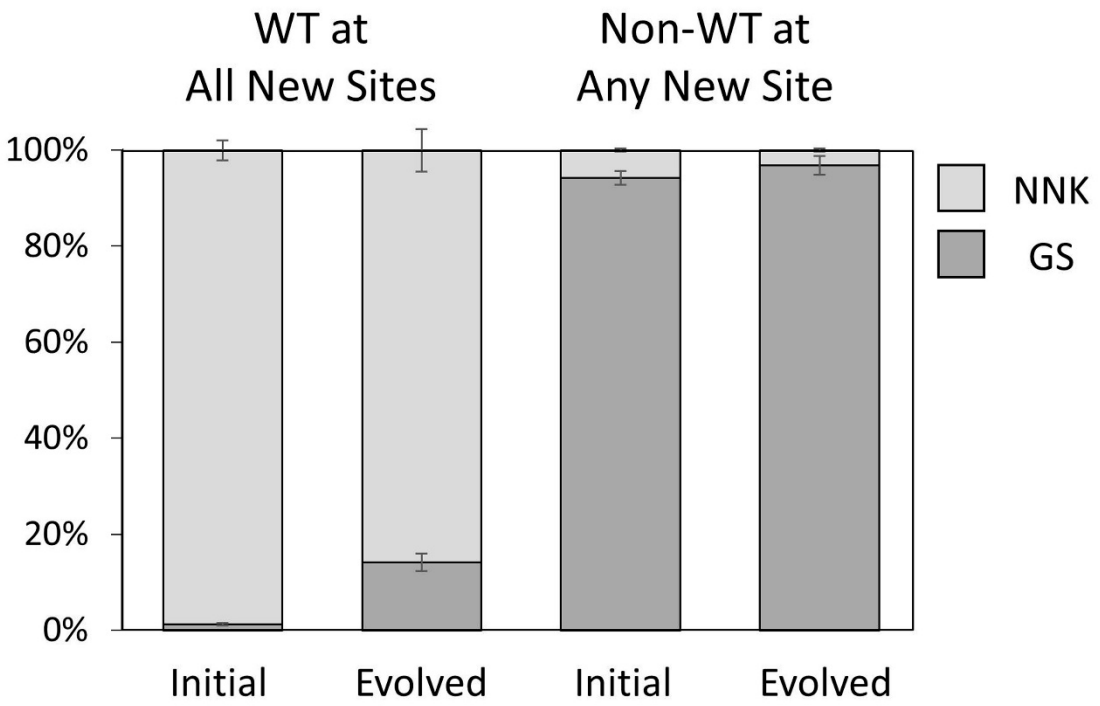


Figure 4-12. ***Wild-type (WT) composition at four newly diversified sites.*** **Sequences were placed into one of two bins: one where variants were entirely WT at all four newly diversified positions (A6, E15, I31, D36); a second bin collected the variants that contained a non-WT amino acid at any of the four new sites. Within each bin, sequences were analyzed at the 13 classically diversified positions for the likelihood of having originated from either the NNK or GS library design.**

The NNK library provides a naïve design benchmark, which demonstrated the overall evolutionary advantage of the GS library (Figure 7), and can also assess the merit of sitewise amino acid preferences. The change in sitewise amino acid frequencies from the initial NNK library to binders evolved from the NNK library (*i.e.* observed experimental evolution) were compared to the changes designed into the constrained sites of the GS library (*i.e* predicted to benefit evolution) (Figure 13A). Experimental values correlate with predicted design (slope = 0.23 ± 0.05; $p < 0.001$). For comparison, evaluation of evolution *away* from the GS library indicates negligible correlation (slope = 0.03 ± 0.04; $p = 0.38$; Figure 13B). This further supports the evolutionary merit of the GS design relative

to naïve NNK. Evaluation of experimental vs. design correlation at each site reveals comparable correlation at most sites with especially strong correlation at site 17 – driven by predicted enrichment of Y and D as well as depletion of R – and lack of correlation at sites 9, 24, and 28 (Figure S4). Overall, the sitewise biases of the GS library design have proven superior to the broad, near-uniform NNK library for binder evolution.

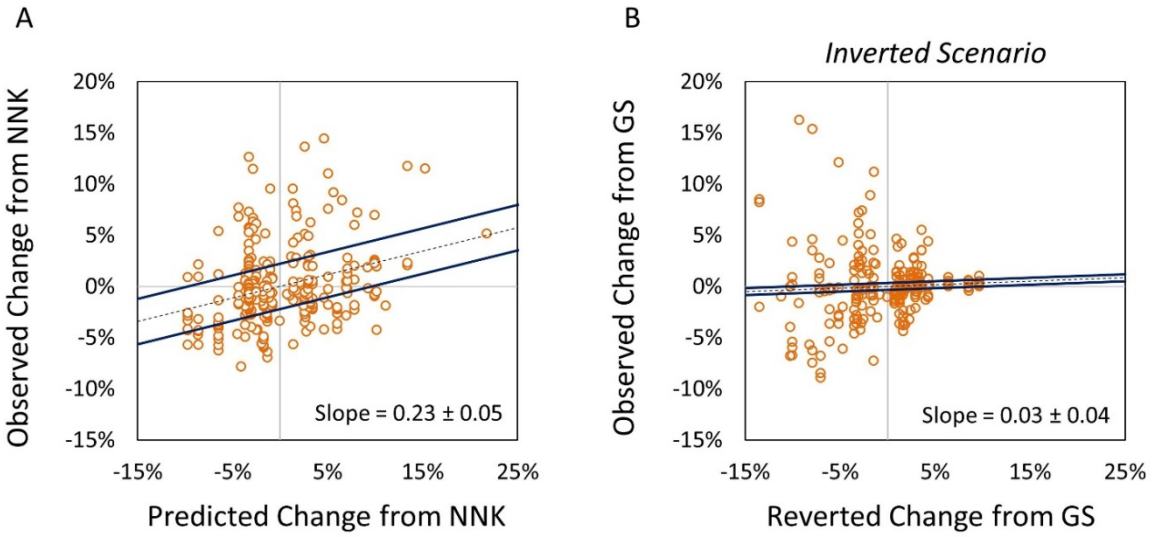


Figure 4-13. ***Evaluating predicted diversity.*** **(A) The observed frequency change reflects the shift in amino acid frequency that was observed within variants having originated from the NNK library design. The predicted frequency change shows the absolute difference in amino acid frequencies between the GS and NNK design. Observed Change from NNK: $f_{AA\ i,\ site\ j\ (NNK\ evolved)} - f_{AA\ i,\ site\ j\ (NNK\ initial)}$. Predicted Change from NNK: $f_{AA\ i,\ site\ j\ (GS\ design)} - f_{AA\ i,\ site\ j\ (NNK\ initial)}$. The 95% prediction interval is shown. (B) Similar analysis to (A), however, it evaluates the predicted and observed mutations away from the GS design. Observed Change from GS: $f_{AA\ i,\ site\ j\ (GS\ evolved)} - f_{AA\ i,\ site\ j\ (GS\ initial)}$. Reverted Change from GS: $f_{AA\ i,\ site\ j\ (NNK\ design)} - f_{AA\ i,\ site\ j\ (GS\ initial)}$.**

4.4.9 Stability

Constrained diversity is designed, in a sitewise manner, to elevate the frequency of evolutionary beneficial amino acids while reducing the frequency of detrimental residues to search more fruitful regions of sequence space. One expected mechanism of this

hypothesis is that select detrimental residues destabilize the scaffold thereby precluding potentially effective binding paratopes because of entropic cost[261,262] or enthalpic destabilization beyond a foldable limit[21,210]. It has been shown that more stable scaffolds enable improved evolution[21,210,263]. A related, but distinct, hypothesis is that reduced destabilization upon mutation improves evolvability. That is, a combinatorial library that contains variants that are less destabilized will contain a higher fraction of folded variants as well as less entropic penalty upon binding (Figure 14A). To partially address this hypothesis, thermal stabilities of several evolved binders were measured by thermal denaturation and circular dichroism spectroscopy (Figure S5A). Random binding variants evolved from the GS library exhibit higher thermal stabilities ($T_m = 62 \pm 4$ºC) than variants from NNK libraries either in the current study ($T_m = 42 \pm 12$ºC; $p = 0.02$) or in the literature (49 ± 8ºC, $p < 0.001$) (Figure 14B). Notably, the NNK-based binders from the current study are not more stable than NNK-based binders in the literature. Thus, the eukaryotic expression machinery of yeast surface display (compared to phage display selections from literature) is *not* responsible for stability enhancement. Rather, the amino acid constraint in the GS library design accounts for this stabilization.

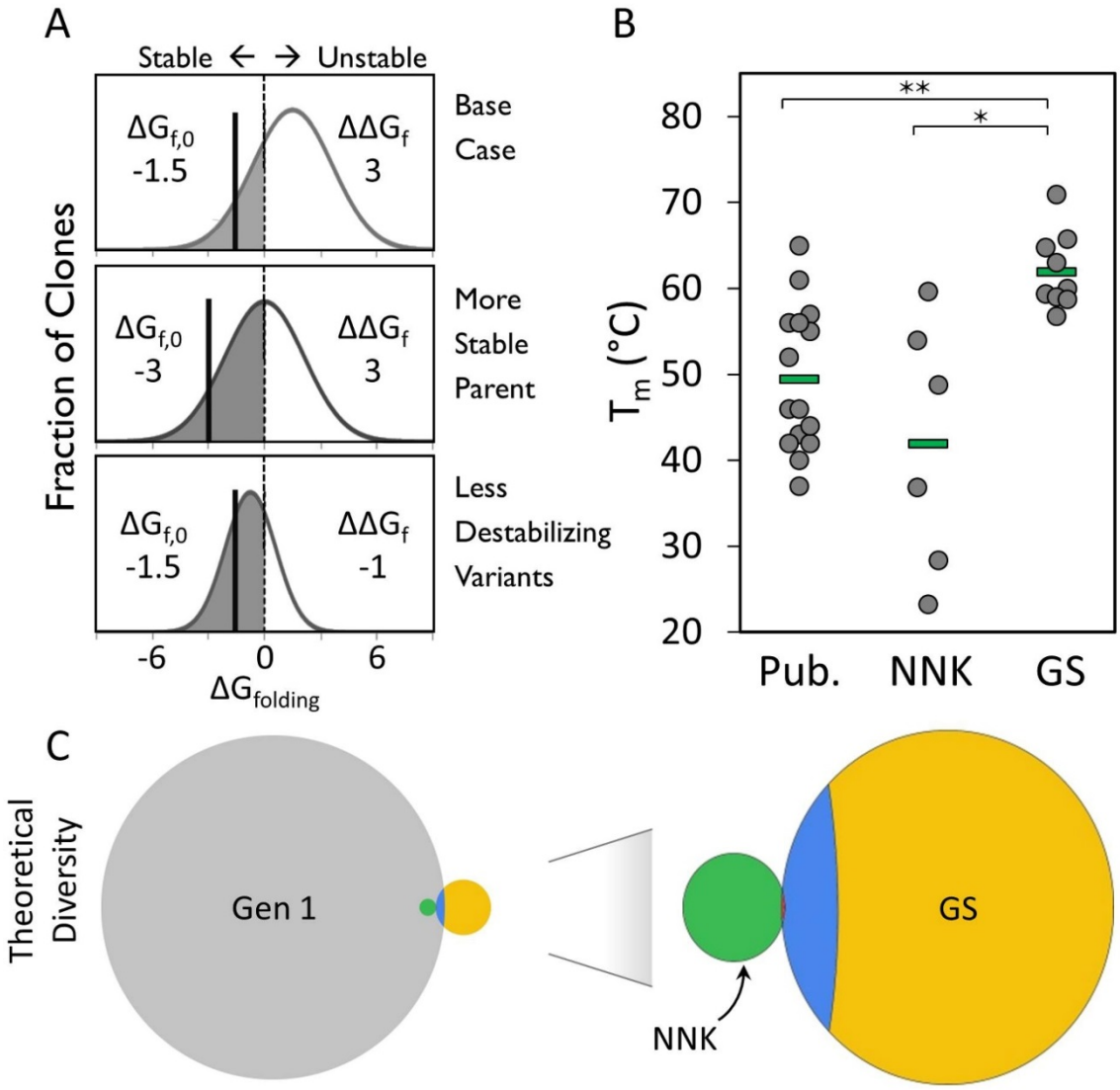


Figure 4-14. ***GS library design yields stable variants.*** **(A) Hypothetical stability distributions of mutant populations derived from a single starting point (i.e. parental clone). A parental clone (vertical solid line labeled $\Delta G_{f,0}$) is randomly mutated, yielding a population with distributed stability. Properly folding proteins are located to the left of the dashed line where the free energy of folding is negative (shaded region). A base case is shown in the top panel. The middle panel presents a more stable parent ($\Delta G_{f,0}$ = -3) with an equivalent broad distribution of destabilization as the base case ($\Delta\Delta G_f$ = 3). A third example is shown in the bottom panel where the parental clone has equivalent stability to the base case ($\Delta G_{f,0}$ = -3); however, upon mutation, destabilization is observed to a much lesser extent throughout the population. This illustrates situations where, relative to the base case, an increased fraction in functional variants is attainable using either a more stable starting point or reduced destabilization upon mutation. (B) Clonal stability of several high-affinity binders were analyzed using circular dichroism. Randomly chosen clones having originated from either the gradient sitewise (GS) or NNK design are shown alongside stability**

measurements previously reported in the literature. Average stabilities (green bars) were found to be 49 ± 8, 42 ± 12, and 62 ± 4 °C for published, NNK, and GS variants, respectively. (C) The relative sequence space afforded by each of the libraries discussed in this work: first generation (Gen 1; gray), traditional design (NNK; green), and second generation (GS; gold). Sequence diversity coverage shared between Gen 1 and GS is highlighted in blue. Sequence space existing within both NNK and GS is shown in red between green and blue regions. Relative to NNK (13 library sites), the theoretical sequence diversity of GS (17 sites) and Gen 1 (15 sites) are larger by 10- and 400-fold, respectively.

The improved performance of the GS library was achieved by biasing amino acid diversity, including eliminating select amino acid options, within the 13 sites traditionally diversified while also expanding diversity by varying four previously conserved sites. The constraint reduced possible sequence space 80-fold and also biased the search of the possible space by preferential occurrence of select amino acids. This biased diversity was 13 ± 3 fold more effective than uniformly applied NNK diversity (Figure 12). The introduced diversity at sites 6, 15, 31, and 36 increased sequence space 800-fold; variants with 0-2 mutations, particularly at sites 6 and 36, were found to be evolutionarily effective whereas triple and quadruple mutants were less functional (Figure 11). Thus, while diversity at these sites is functional, they would benefit from further bias. Overall, the GS library has a 10-fold greater potential sequence space (Figure 14C), but searches this space in a more biased manner than the NNK library. The average uniformity of diversity, as measured by Shannon entropy, of the NNK library is 0.97 versus 0.74 for the GS library (Figure S9). Overall, the result of the sitewise bias in the GS library is a higher frequency of binder discovery (Figure 7, Figure S7) and more stable binders (Figure 14B).

4.5. Conclusions

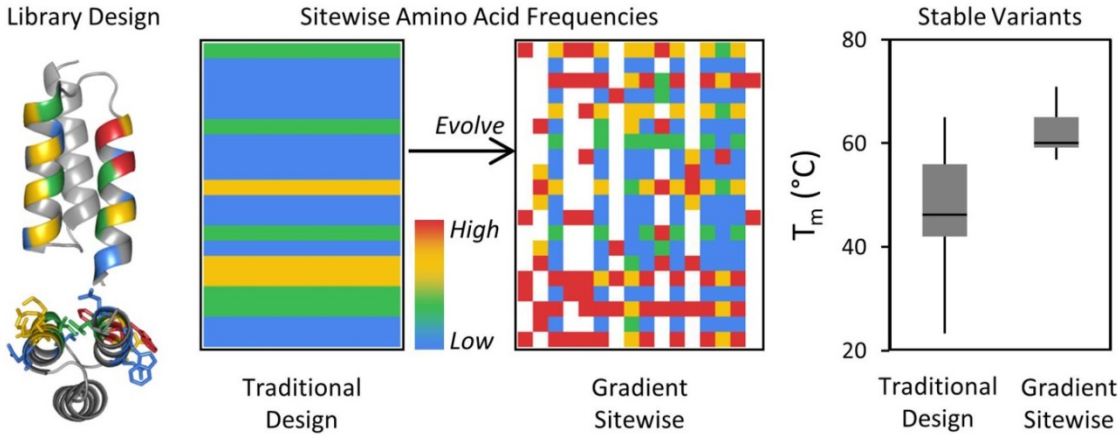
In this study, amino acid frequency distributions for each site in a combinatorial affibody library were designed from a multitude of inputs: high-throughput binder evolution, computed mutant stabilities, target accessibility, helical propensity, chemical complementarity, and frequency in natural homologs. We evaluated the efficiency of the gradient sitewise (GS) design by direct competition of multiple libraries in the context of binder discovery against a collection of protein targets. Site-specific amino acid bias, as well as introduced diversity at four additional positions, in the GS library enabled more efficient evolution than the traditional approach of homogeneous diversity across all diversified sites in the NNK library. Analysis indicates that GS's evolutionary benefit resulted from the sitewise constraint and in spite of the broadened diversity at sites 15 and 31. Amino acid preferences, both overall and at select sites, were revealed. The first generation library, designed to explore sitewise amino acid preference from a broad range of paratopes, exhibited much broader diversity than optimal, as assessed by sitewise frequencies in evolved repertoires. In fact, these 'overdiversified' paratopes frequently required cysteine pairs to achieve functionality. Conversely, reduction of undesirable amino acids at select sites in the second generation library greatly reduced the presence of cysteine pairs in evolved clones. Moreover, the sitewise bias yielded variants of higher stability (15ºC higher midpoint of thermal destabilization) than NNK-based variants. Evolutionary efficiency benefits from increased frequency of not only amino acids that drive intermolecular interactions but also intramolecularly tolerated mutants that reduce destabilization. Overall, the library design approach favoring constrained amino acid diversities that take into account complementarity, amino acid frequencies in previously

discovered binders, diversity in natural homologs, and solvent exposed surface area produces binding ligands more efficiently than the unconstrained NNK library design in the affibody scaffold.

4.6. Acknowledgments

We appreciate assistance from Baradan Panta and Nicholas Heise in assembling the database of published evolved affibodies, Justin Klesmith for valuable suggestions within the manuscript, the University of Minnesota Genomics Center for assistance with Illumina sequencing, the Masonic Cancer Center Flow Cytometry Shared Resource, and the University of Minnesota Supercomputing Institute for computational resources during sequence analysis.

4.7. Table of Contents Figure



4.8. Supplemental Figures

Table S1 Wild-type Affibody Sequence and Library Design Summary

Table S4-1. Wild-type affibody sequence and library design summary. The diagram compares traditional design (NNK), first generation (Gen 1), and gradient sitewise (GS) library designs, as well as framework mutations established by Feldwisch, et al.[196] The traditional design uniformly diversifies thirteen sites with an NNK degenerate codon. The first generation design broadly diversifies ten sites and provides a gradient of diversity at five sites.

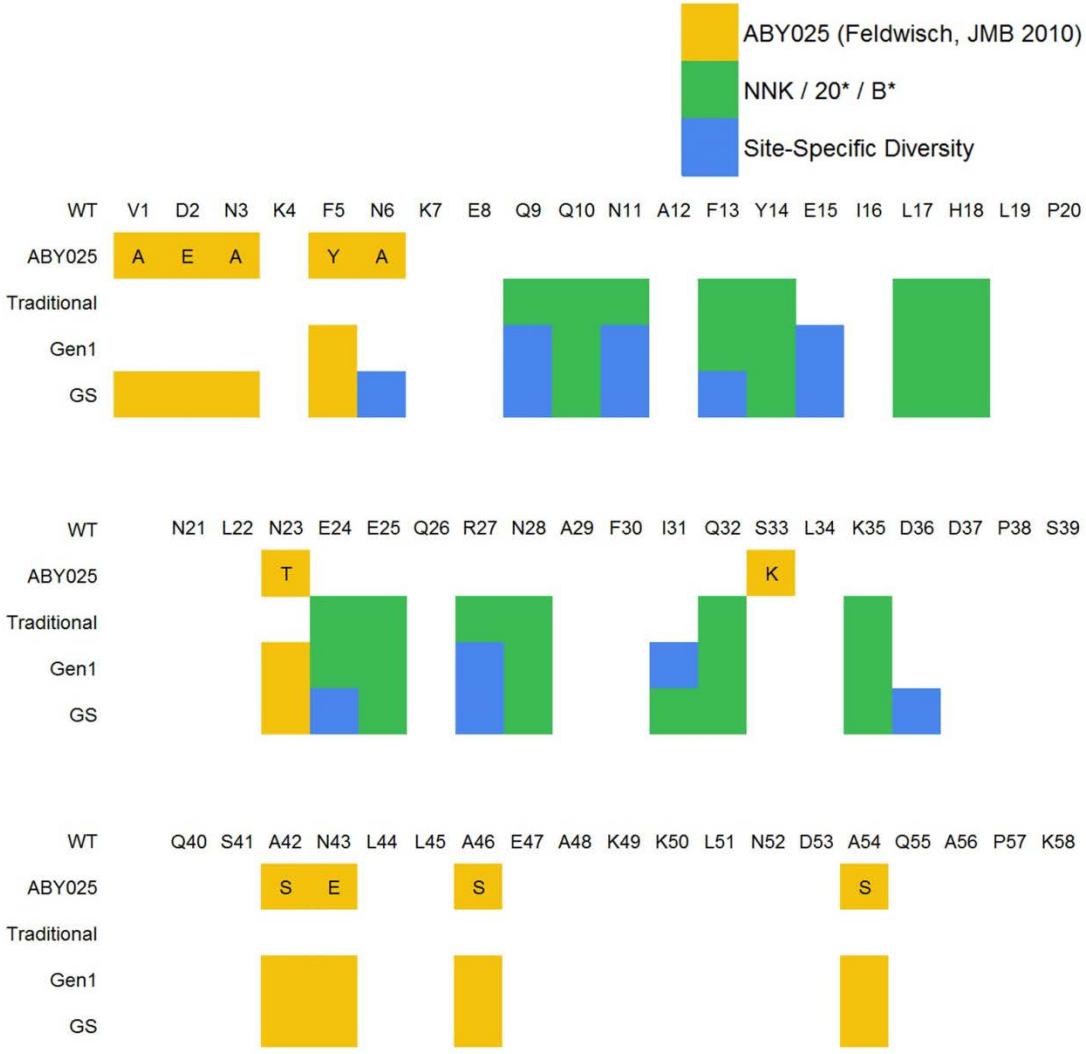


Figure S4-1. Comparison of broad diversity codon design. (A) Amino acid frequencies for broadly diversified sites within the first (20*) and second (B* and B_{LC}*) generation gradient sitewise (GS) library designs. The primary second generation GS design (B*) mimics natural antibody repertoires (Abysis database) and frequencies forecasted from evolved affibodies (B) while also emphasizing reduced levels of glycine, motivated by low the low helix propensity observed in previous studies[249–254], and potentially non-specific arginine[49,54,55]. The B_{LC}* codon design resembles B* while also reducing the cysteine content. Note that because degenerate codons were used to implement 20*, B* and B_{LC}* designs, additional restrictions were imposed by the genetic code. (B) Affibody sequences from existing literature and high-throughput sequence analysis of binders selected from the first generation affibody library were aggregated. The amino acid frequencies associated with both the initial libraries and evolved populations are shown. The 'forecast' is based on an extrapolation from the trend observed between the initial and evolved populations as described by: $f_{forecast} = f_{binder} + 0.25(f_{binder} - f_{initial})$.

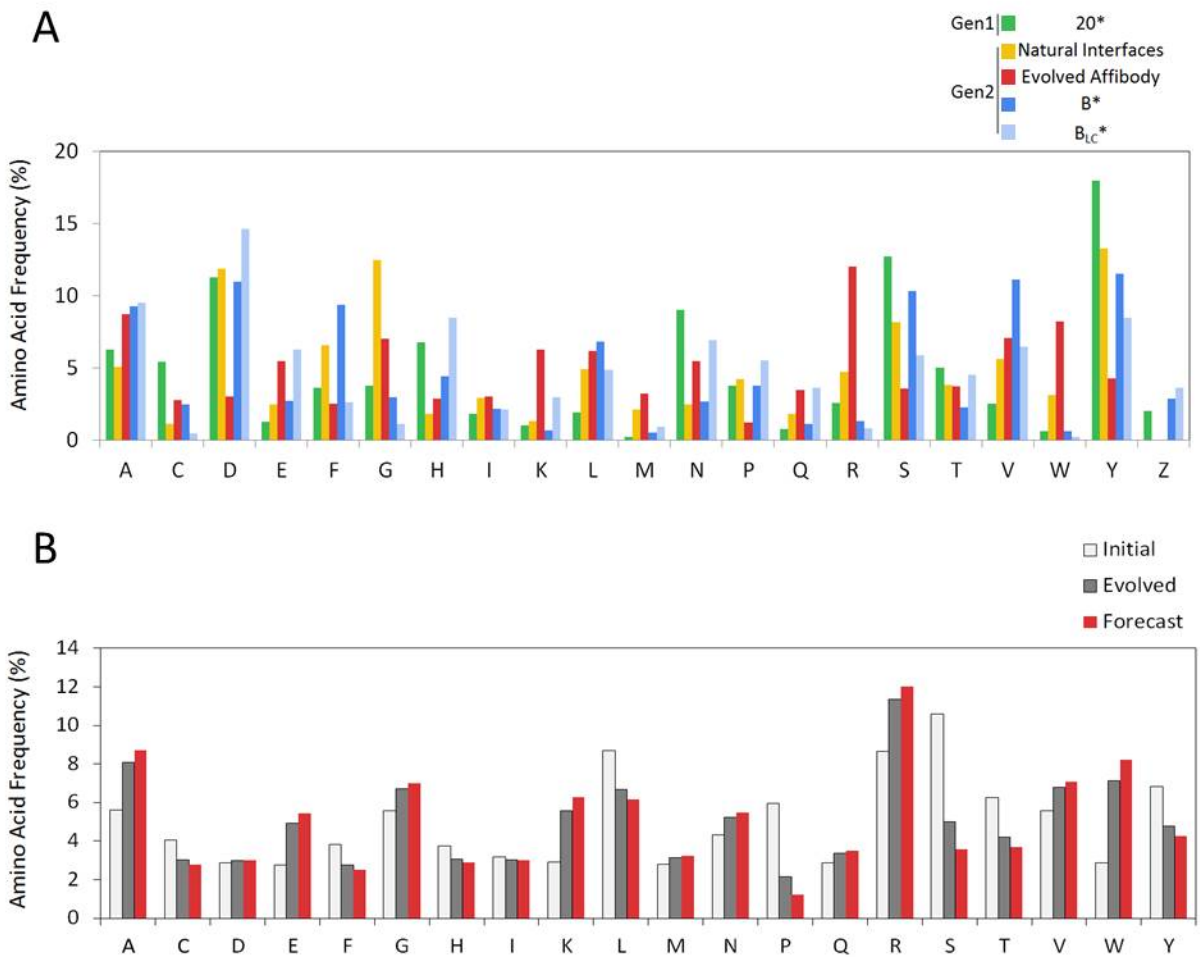
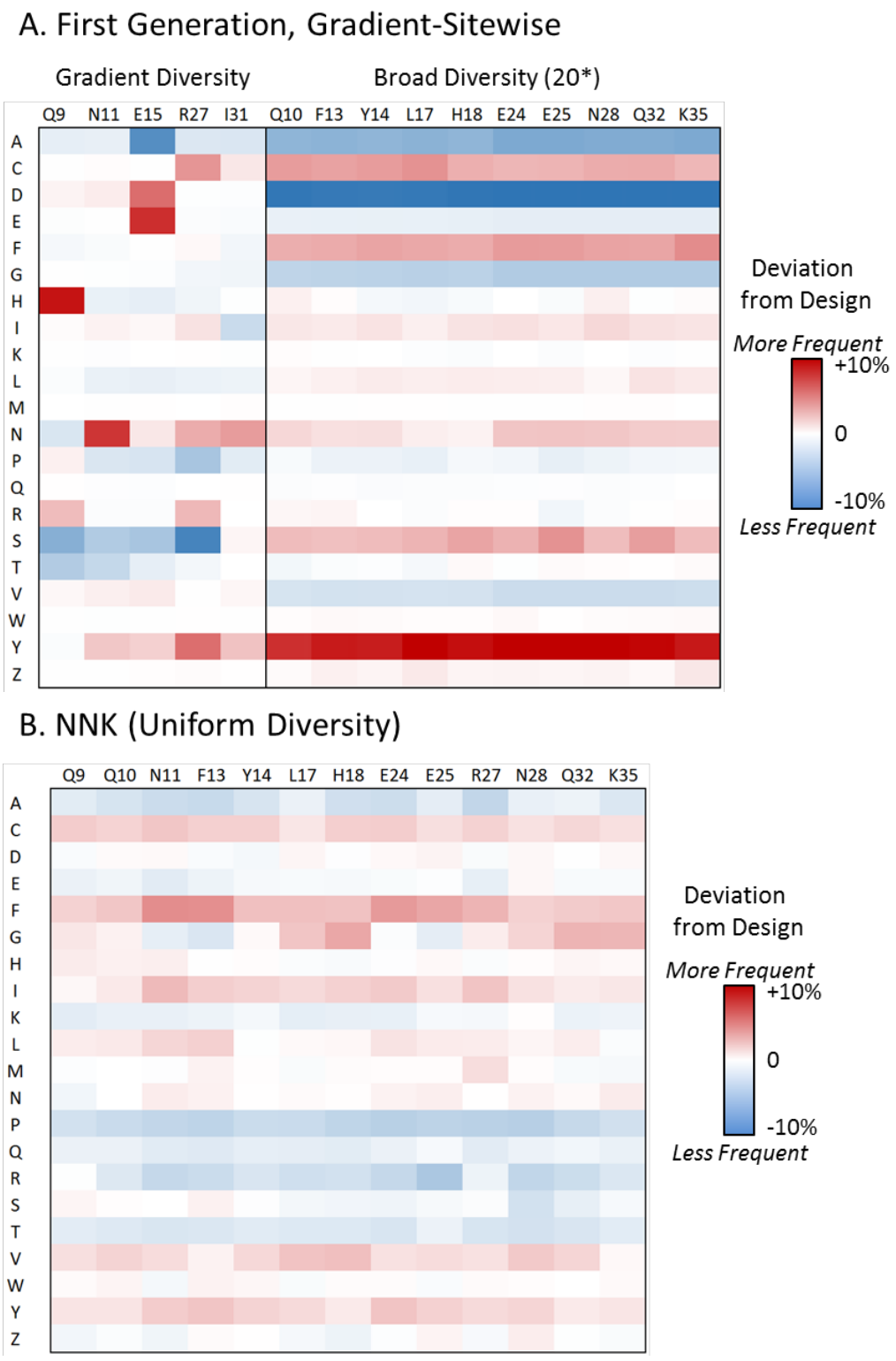


Figure S4-2. Design vs observed initial libraries. Sitewise comparison between design and observed amino acid frequencies ($f_{observed}$–f_{design}) within the initial populations of the first generation library (A) and for each of the three second generation library designs: (B) NNK, (C) GS, (D) GS_{LC} .



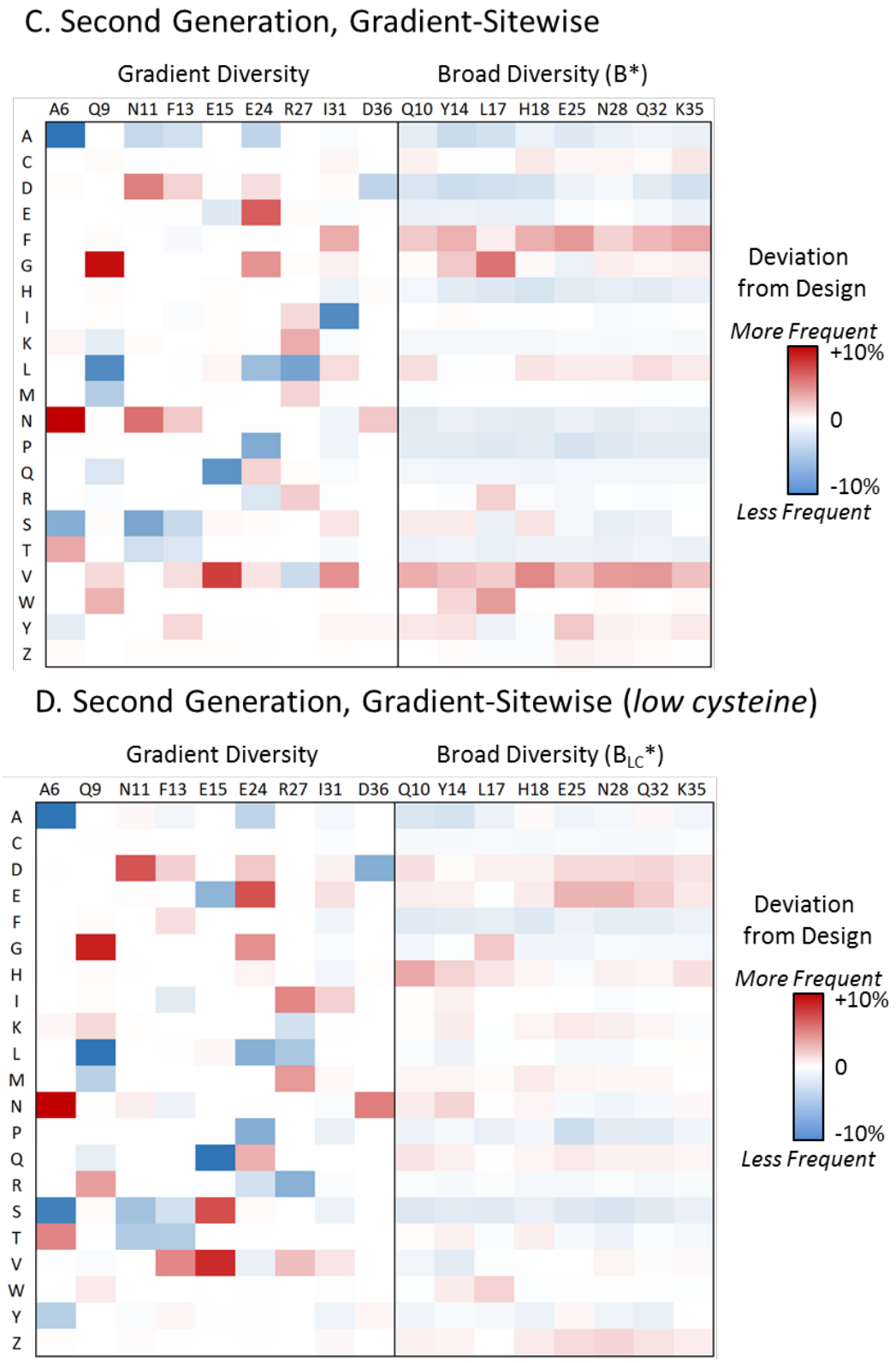
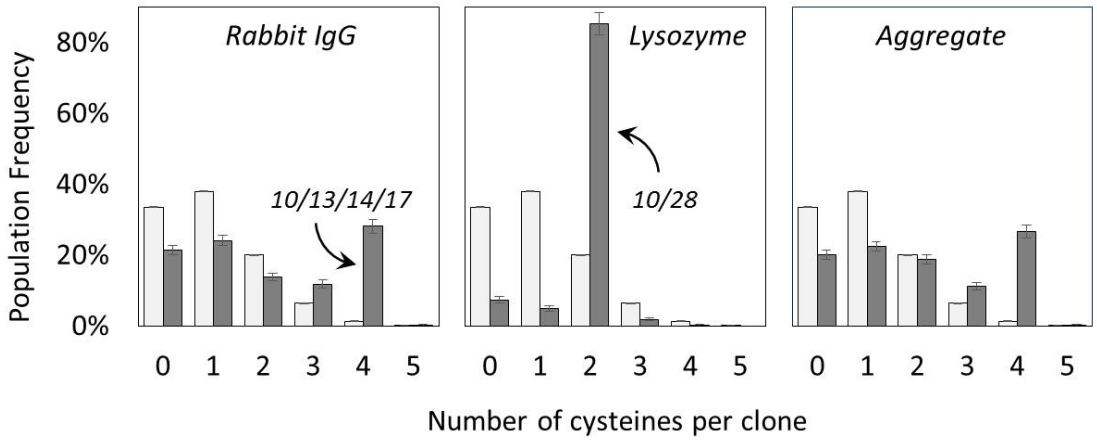


Figure S4-3. First generation evolved binder campaign cysteine content and stability. (A) Cysteine content within individual target campaigns using the first generation affibody library is shown. The sites contributing to the most enriched cysteine motifs are indicated for each

campaign. Aggregate data was calculated by weighting the relative number of family clusters from each campaign where rabbit IgG outnumbered lysozyme 479 vs 44. (B) Thermal stability of cysteine-rich clones was evaluated using circular dichroism under reducing (2mM DTT, red) or non-reducing (0mM DTT, black) conditions. The 220 nm wavelength was monitored over a range of temperatures. Cysteine positions are labeled above each denaturation plot and variant names are indicated on the lower right (e.g. "AαrIgG-1").

A



B

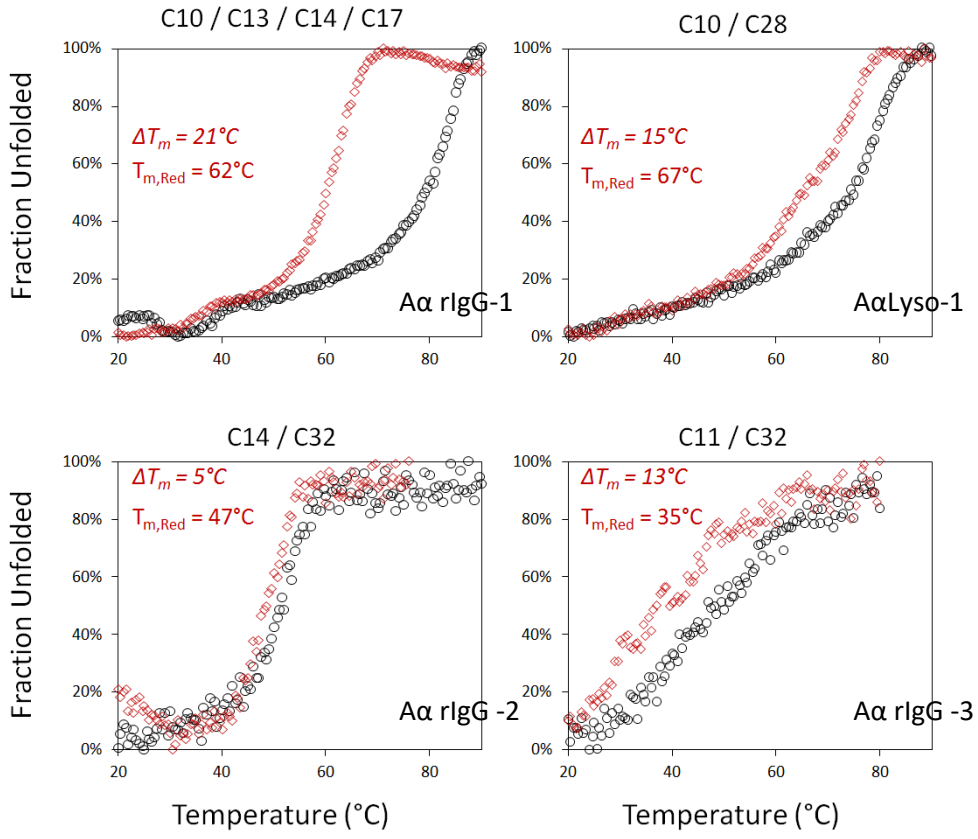
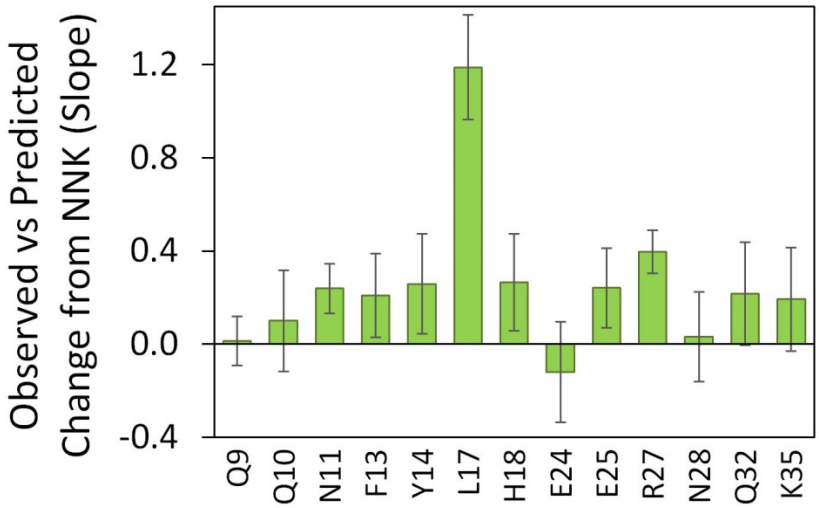


Figure S4-4: Predictive analysis of GS vs NNK library designs. Observed and predicted changes in sitewise amino acid frequencies were calculated within variants determined, via maximum posterior probability, to have originated from the NNK library. The observed frequency change reflects the change in amino acid frequency that was observed upon discovery and evolution: $\Delta f_{obs} = f_{AA\ i,\ site\ j\ (NNK\ evolved)} - f_{AA\ i,\ site\ j\ (NNK\ initial)}$. The predicted frequency change shows the absolute difference in amino acid frequencies between the GS and NNK design: $\Delta f_{pred} = f_{AA\ i,\ site\ j\ (GS\ design)} - f_{AA\ i,\ site\ j\ (NNK\ initial)}$. Δf_{obs} was plotted versus Δf_{pred} for each site (with each amino acid providing a data point for each site). (A) The slope (± standard deviation) of the linear regression of Δf_{obs} vs. Δf_{pred}. (B) The Δf_{obs} vs. Δf_{pred} plots for the two sites having the highest (i.e. 'Best Predictions') and lowest (i.e. 'Worst Predictions') slopes and their respective 95% prediction intervals.

A



B

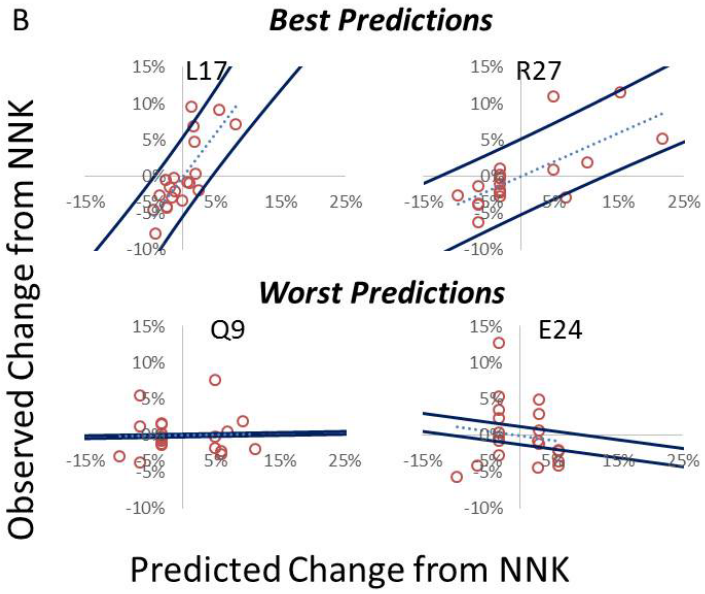


Table S4-2: Sitewise amino acid frequencies from natural homolog sequences of affibody, Pfam family B(PF02216). The most frequently observed amino acids for each site are listed at the bottom. Values are shown as percentages.

	A1	E2	A3	K4	Y5	A6	K7	E8	M9	W10	A11	A12	W13	E14	E15	I16	R17	N18	L19	P20
A	38	17	18	0	0	5	10	21	0	0	12	82	0	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
D	1	36	1	6	0	18	2	23	0	0	0	0	0	3	1	0	0	0	0	11
E	1	5	0	1	4	0	19	38	0	0	0	0	0	6	55	0	0	0	0	2
F	0	1	1	0	49	0	0	0	0	0	0	12	86	0	0	0	0	0	0	0
G	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	1	4	0	16	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0
I	2	0	0	0	1	1	0	0	2	0	0	0	1	0	0	59	0	0	0	0
K	9	3	0	24	0	0	45	2	0	0	10	0	0	0	0	0	12	2	0	10
L	1	0	0	0	4	1	0	0	2	0	0	0	0	0	1	12	74	0	47	0
M	2	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	50	0
N	14	12	53	38	1	52	4	5	0	12	60	1	0	10	1	0	0	58	2	1
P	6	0	18	0	1	0	2	0	0	0	0	2	0	0	0	0	0	0	0	72
Q	16	24	0	16	0	0	5	11	93	87	0	0	0	1	30	1	0	1	0	0
R	1	0	0	10	10	1	1	0	2	0	0	0	0	0	2	0	9	0	0	0
S	0	0	1	2	0	4	0	0	0	0	18	0	0	3	1	0	0	0	0	3
T	9	0	2	1	0	2	9	0	0	0	0	2	0	0	0	0	0	0	0	0
V	0	1	1	0	1	16	1	0	0	0	0	1	0	1	10	28	4	0	0	0
W	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
Y	0	0	0	0	13	0	0	0	0	0	0	0	12	77	0	0	1	0	0	0
Most Frequent																				
1	A (38)	D (36)	N (53)	N (38)	F (49)	N (52)	K (45)	E (38)	Q (93)	Q (87)	N (60)	A (82)	F (86)	Y (77)	E (55)	I (59)	L (74)	N (58)	M (50)	P (72)
2	Q (16)	Q (24)	A (18)	K (24)	H (16)	D (18)	E (19)	D (23)	R (2)	N (12)	S (18)	F (12)	Y (12)	N (10)	Q (30)	V (28)	K (12)	H (39)	L (47)	D (11)
3	N (14)	A (17)	P (18)	Q (16)	Y (13)	V (16)	A (10)	A (21)	L (2)	W (1)	A (12)	P (2)	W (1)	E (6)	V (10)	L (12)	R (9)	K (2)	N (2)	K (10)
4	T (9)	N (12)	H (4)	R (10)	R (10)	A (5)	T (9)	Q (11)	I (2)	Y (0)	K (10)	T (2)	I (1)	S (3)	R (2)	Q (1)	V (4)	Q (1)	Y (0)	S (3)

	N21	L22	T23	G24	W25	Q26	M27	T28	A29	F30	I31	A32	A33	L34	V35	D36	D37	P38	S39	Q40
A	0	0	0	22	11	0	0	0	5	0	0	11	1	0	0	0	0	0	0	1
C	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	1	0	0	4	19	0	0	2	1	0	0	1	0	0	0	80	78	0	12	0
E	0	0	0	66	56	11	0	2	0	0	0	1	1	0	0	18	0	0	10	1
F	0	0	0	0	0	0	0	0	0	77	0	0	0	0	0	0	0	0	0	0
G	11	0	0	7	0	0	0	0	72	0	0	2	0	0	0	0	0	0	0	0
H	1	0	0	0	0	1	2	0	0	0	0	0	0	0	2	0	11	0	0	0
I	0	12	1	0	0	0	0	0	0	0	96	0	0	13	0	0	0	0	0	0
K	0	0	0	0	0	0	12	0	0	0	0	11	1	0	86	0	0	0	0	1
L	0	86	0	0	0	0	0	0	0	1	2	0	0	87	0	0	0	0	2	0
M	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
N	83	0	61	1	0	0	0	95	11	0	0	2	0	0	0	2	11	2	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0
Q	1	0	0	0	10	87	4	0	11	0	0	74	11	0	0	0	0	0	1	57
R	0	0	0	0	0	1	78	0	0	0	0	0	1	0	11	0	0	0	0	15
S	0	0	5	0	0	0	0	0	0	0	0	0	74	0	0	0	0	0	76	1
T	0	0	32	0	0	0	0	1	0	0	0	0	10	0	0	0	0	0	0	2
V	0	2	0	0	2	0	0	0	1	0	1	0	2	0	1	1	0	0	0	21
W	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	4	0	0	0	1	0	1	0	0	22	2	0	0	0	1	0	0	0	0	1
Most Frequent																				
1	N (83)	L (86)	N (61)	E (66)	E (56)	Q (87)	R (78)	N (95)	G (72)	F (77)	I (96)	Q (74)	S (74)	L (87)	K (86)	D (80)	D (78)	P (98)	S (76)	Q (57)
2	G (11)	I (12)	T (32)	A (22)	D (19)	E (11)	K (12)	D (2)	Q (11)	Y (22)	Y (2)	A (11)	Q (11)	I (13)	R (11)	E (18)	N (11)	N (2)	D (12)	V (21)
3	Y (4)	V (2)	S (5)	G (7)	A (11)	H (1)	Q (4)	E (2)	N (11)	L (1)	L (2)	K (11)	T (10)	Y (0)	H (2)	N (2)	H (11)	Y (0)	E (10)	R (15)
4	D (1)	Y (0)	I (1)	D (4)	Q (10)	R (1)	M (3)	T (1)	A (5)	W (0)	V (1)	G (2)	V (2)	W (0)	Y (1)	V (1)	Y (0)	W (0)	L (2)	T (2)

	S41	A42	N43	L44	L45	A46	E47	A48	K49	K50	L51	N52	D53	A54	Q55	A56	P57	K58
A	12	48	0	0	0	35	0	76	0	5	3	0	2	3	0	44	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	9	0	0	0	1	0	0	0	0	1	80	0	0	0	0	0
E	2	0	19	0	0	0	92	1	1	0	1	0	2	3	55	0	0	5
F	0	0	0	0	12	0	0	0	0	0	1	0	0	0	1	0	0	0
G	0	0	0	0	0	26	0	0	0	0	0	0	0	3	0	0	0	0
H	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
I	0	0	0	24	2	1	0	1	2	0	1	0	0	3	0	1	0	3
K	0	13	1	2	0	0	0	0	72	87	0	12	0	0	33	0	0	18
L	0	0	0	37	75	0	0	2	0	0	78	0	0	0	0	1	0	0
M	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
N	0	5	62	0	0	3	0	0	0	2	0	76	1	0	1	1	0	62
P	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	79	0
Q	0	18	10	0	0	4	5	0	21	1	0	10	1	0	6	1	0	0
R	0	0	0	0	0	0	0	0	1	3	0	0	0	0	2	0	0	5
S	87	1	0	0	1	18	1	18	0	1	0	0	14	78	0	48	11	8
T	0	14	0	0	0	0	2	1	0	0	5	0	0	0	0	4	11	0
V	0	0	0	38	1	13	0	1	0	0	12	0	0	10	0	2	0	0
W	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Most Frequent																		
1	S (87)	A (48)	N (62)	V (38)	L (75)	A (35)	E (92)	A (76)	K (72)	K (87)	L (78)	N (76)	D (80)	S (78)	E (55)	S (48)	P (79)	N (62)
2	A (12)	Q (18)	E (19)	L (37)	F (12)	G (26)	Q (5)	S (18)	Q (21)	A (5)	V (12)	K (12)	S (14)	V (10)	K (33)	A (44)	T (11)	K (18)
3	E (2)	T (14)	Q (10)	I (24)	W (10)	S (18)	T (2)	L (2)	I (2)	R (3)	T (5)	Q (10)	E (2)	I (3)	Q (6)	T (4)	S (11)	S (8)
4	Y (0)	K (13)	D (9)	K (2)	I (2)	V (13)	S (1)	V (1)	H (2)	N (2)	A (3)	D (1)	A (2)	G (3)	R (2)	V (2)	Y (0)	R (5)

Sitewise Design of Second Generation Library

At site 9, low SASA (19%), constrained homolog sequences (Shannon entropy: 0.12), and poor theoretical mutational tolerance (only seven residues tolerated at 1.5 kcal/mol) motivate sequence constraint while position near the center of the paratope supports mild diversity. Wild-type Q appears 93% in homologs, is maintained (3 to 4%) in binders, and is predicted to be stable. L, M, and R occur naturally, are predicted to be stabilizing, and increase in binding populations. Aggregate data from gradient library evolved binders and previously published binders show an increase in W from 2% to 7%. V and K are predicted to be stable and enriched in binders. Note that constrained N decreases 37% to 20% in the gradient library. Thus, QLMRRK/VWLG (G part of degenerate codon) was chosen for diversity. Site 10 is naturally constrained (87% Q, 12% N, and 1% W) but is well-exposed (66% SASA) and tolerant of mutations theoretically (13 tolerant amino acids) and in evolved binders. *B** diversity was used. N11 shows reasonable solvent exposure (51%), albeit with only modest proximity to the paratope core. It is naturally diverse (Shannon entropy: 0.37) and predicted tolerant of mutation (19 tolerant amino acids). N and A occur naturally (60% and 12%), are prevalent in binders (8% and 8% from 10% and 6%), and predicted to be stable. S occurs 18% naturally and is predicted to be stable, although it is depleted from binders (10% to 6%). D, Y, and T also appear in the NAS degenerate codon. D is maintained in binders (3% to 4%). Y is predicted to be stable and is prevalent in antibody CDRH3, although slightly depleted at this site in evolved binders. T is predicted to be stable and maintained at 6% in binders. NASTDY was used. At F13, binder sequences and FoldX stability calculations (11 tolerated amino acids) indicate a modest tolerance for mutation, while homology shows significant conservation (Shannon entropy: 0.16), and the sidechain is reasonably buried (33%). The following amino acids are enriched in binders: A (5% to 15%), T (6% to 10%), Y (7% to 8%), W (2% to 5%). Yet W is predicted to be highly destabilizing. F is naturally frequent (86%) and

maintained at 4% in binders. Moderate diversity (FYAT, which includes VINDS on the degenerate codon) was afforded in the subsequent library. S is tolerated and modifies from 10% in libraries to 8% in binders. V is stable and maintained at 5% in binders. I is present naturally, predicted to be stable and mildly enriched from 4% to 5% in binders. D and N are mildly destabilizing and slightly depleted in binders. Site 14 is exposed (72%), naturally diverse (Shannon entropy: 0.3), and tolerant (17 amino acids tolerant). Broad diversity was incorporated. Additionally, the strong increase in G (5% to 17%) and W (2% to 8%), both underrepresented in B*, may benefit the evolvability. Thus, a mixture of 95% B* and 5% GW was used (to yield 5% G). L17 is near the middle of the binding paratope and reasonably naturally diverse (Shannon entropy: 0.30) although only 40% solvent accessible and poorly tolerant of mutation. Binders have elevated levels of V (5% to 16%), W (2% to 12%), G (5% to 9%), D (3% to 6%), and Y (7% to 8%). The two dominant naturally occurring side chains, L (74%) and K (12%), are depleted (8% to 4% for L; 3% to 2% for K) in binders while two other naturally occurring residues, R (9%) and Y (1%), are enriched (8% to 9% for R; 7% to 8% for Y). To better match the amino acid observed within binder sequences, this site received both 86.5% broad diversity (B*) and 13.5% WGR to yield 5% W, 6% G, and 5% R. Site H18 used B* diversity because of broad tolerance (19 stable residues), reasonable diversity in nature (Shannon entropy: 0.28), and high exposure (67% SASA).

Within the second helix, site 24 is well-exposed (88%), naturally diverse (Shannon entropy: 0.32), and tolerant (19 tolerant residues). The breadth of diversity suggests B* although Y and S – frequent in B* – are depleted in binders (7% to 3% and 10% to 5%) while R and G – rare in B* – are enriched (8% to 13% and 5% to 9%). Two amino acids coupled in the genetic code to Y and S are not desired: C and F are depleted in binders (4% to 1% and 4% to 3%), although W is enriched (2% to 5%) and modestly stable. Thus, a diverse codon encoding 16 amino acids aside from C, F, and Y (and also W, unfortunately) was used. Site 25 is well-exposed (90%), naturally diverse

(Shannon entropy: 0.42), stable to mutation (all 20 residues tolerant within 1.5 kcal/mol), and diverse in binders. B* diversity was used. Site 27, while near the center of the paratope and mildly diverse naturally (Shannon entropy: 0.28), is fairly buried (27%) and relatively unstable to mutation (only 10 tolerant residues). Wild-type R and fellow cation K are frequent in homologs (78% and 12%), stable, and enriched in binders (13% to 20% and 3% to 16%). Small-to-medium hydrophobics (A, I, L, M, and V) are also predicted to be stable and reasonably maintained in binder development (6% to 5%, 4% to 4%, 8% to 6%, 2% to 4%, and 6% to 9%). RKVMIL was used. Site 28 is exposed (71%), near the paratope center, and predicted to be tolerant of mutation (16 tolerant amino acids) although constrained naturally (95% N; Shannon entropy: 0.09). Only P is strongly depleted in binders (6% to 1%) while S (10% to 6%) and Y (7% to 4%) are mildly depleted. B* diversity was used. Site 32 is exposed (72%) although pointed slightly away from the paratope core. It is naturally diverse (Shannon entropy: 0.3), predicted to be tolerant to mutation (17 amino acids), and exhibits broad diversity in binders (with the exception of proline depletion at 6% to 1%). Enriched amino acids include R (8% to 15%), K (3% to 7%), G (5% to 9%), and A (5% to 9%). B* diversity was used. Site 35 is well-exposed (76%) and predicted tolerant (16 residues) although naturally conserved (86% K, 11% R, Shannon entropy: 0.19). Binders exhibit reasonable diversity although numerous residues are mildly depleted in exchange for enrichment of W (2% to 11%), Y (7% to 12%), E (2% to 5%), F (4% to 6%), and L (8% to 10%). B* diversity was used.

Sites 6, 15, 31, and 36 are not traditionally diversified. Natural diversity at site 6 is high (Shannon entropy: 0.49), and mutations to E and A in the context of a HER2 binder were not destabilizing[196]. The site is exposed to solvent and could be expected to interact with target in many cases; in fact the HER2 ligand mutants impacted affinity. Mild diversity (NSYT) was tested. Site 15 is naturally diverse (Shannon entropy: 0.39) and tolerant of mutation (15 residues), but is pointed away from the paratope and only 38% accessible to solvent. Despite 55% E on natural homologs, biased E

was strongly depleted in binding populations (38% to 18%) from the gradient library. Wild-type Q (30% in homologs) is conserved in published libraries. S is present naturally (1%) and enriched in binders (5% to 15%). V (10% naturally and enriched 3% to 5% in binders) will also be considered. Thus, QSEV diversity (on three separate oligonucleotides) was allowed. I31 is conserved in previous libraries and naturally (96% I; Shannon entropy: 0.07) and is predicted to be poorly tolerant of mutation (only eight tolerant residues). Yet it is in the center of the planned paratope and tolerated diversity in the gradient library analyses. I was maintained at high levels (33% to 32%) but most other amino acids were also tolerated with the most significant depletions being S (18% to 10%) and Y (11% to 8%). A 30:70 mixture of I and the B* codon was used. D36 is conserved in previous libraries and naturally (80% D, 18% E, Shannon entropy: 0.21). Yet it is solvent accessible (66%), could be expected to contact target in some cases, and is predicted to be tolerant to mutation. Diversity could be beneficial both inter- and intra-molecularly although such diversity should be minimal to limit detriment. DN diversity was used.

Table S4-3: Second generation library design summary table. Top – Description of amino acids allowed at each site. When a residue is listed multiple times, it indicates the relative abundance of the residue at that site. Bottom – Degenerate codons are listed that encode for the amino acids at each site.

Second Generation, Gradient-Sitewise (GS) Amino Acid Design

	A6	Q9	Q10	N11	F13	Y14	E15	L17	H18	E24	E25	R27	N28	I31	Q32	K35	D36
Amino Acid Design 1	NSYT	QRRKLM (50)	B*	NSTYAD	FYATVINDS	B* (95)	QE (33)	B* (86.5)	B*	LLPPHQRRRSNKT TIMVVAADEGG	B*	RK (50)	B*	I (30)	B*	B*	DN
Amino Acid Design 2		LVWG (50)				GW (5)	S (33)	WGR (13.5)				VMIL (50)		B* (70)			
Amino Acid Design 3							V (33)										

Second Generation, Gradient-Sitewise *low cysteine* (GS_{LC}) Amino Acid Design

	A6	Q9	Q10	N11	F13	Y14	E15	L17	H18	E24	E25	R27	N28	I31	Q32	K35	D36
Amino Acid Design 4	NSYT	QRRKLM (50)	B_{LC}*	NSTYAD	FYATVINDS	B_{LC}* (95)	QE (33)	B_{LC}* (86.5)	B_{LC}*	LLPPHQRRRSNKT TIMVVAADEGG	B_{LC}*	RK (50)	B_{LC}*	I (30)	B_{LC}*	B_{LC}*	DN
Amino Acid Design 5		LVWG (50)				GW (5)	S (33)	WGR (13.5)				VMIL (50)		B_{LC}* (70)			
Amino Acid Design 6							V (33)										

Second Generation, Gradient-Sitewise (GS) Degenerate Codon Design

	A6	Q9	Q10	N11	F13	Y14	E15	L17	H18	E24	E25	R27	N28	I31	Q32	K35	D36
Codon Design 1	WMC	MDG (50)	B*	DMC	DHC	B* (95)	SAG (33)	B* (86.5)	B*	SNS	B*	ARA (50)	B*	ATC (30)	B*	B*	RAC
Codon Design 2		KKG (50)				KGG (5)	TCC (33)	BGG (13.5)				VTR (50)		B* (70)			
Codon Design 3							GTG (33)										

Second Generation, Gradient-Sitewise *low cysteine* (GS_{LC}) Degenerate Codon Design

	A6	Q9	Q10	N11	F13	Y14	E15	L17	H18	E24	E25	R27	N28	I31	Q32	K35	D36
Codon Design 4	WMC	MDG (50)	B_{LC}*	DMC	DHC	B_{LC}* (95)	SAG (33)	B_{LC}* (86.5)	B_{LC}*	SNS	B_{LC}*	ARA (50)	B_{LC}*	ATC (30)	B_{LC}*	B_{LC}*	RAC
Codon Design 5		KKG (50)				KGG (5)	TCC (33)	BGG (13.5)				VTR (50)		B_{LC}* (70)			
Codon Design 6							GTG (33)										

(B) Oligonucleotide sequences used for the construction and assembly of the second generation libraries.
Name Sequence

B_a-a-a-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)SAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACC
TGACC
B_a-a-b-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)TCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
B_a-a-c-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)GTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACC
TGACC
B_a-a-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)SAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-a-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)TCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-a-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)GTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-b-a-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGSAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-b-b-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGTCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-b-c-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGGTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-b-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGSAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-b-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGTCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-b-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGGTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-a-a-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)SAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
B_b-a-b-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)TCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
B_b-a-c-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)GTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
B_b-a-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)SAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-a-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)TCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-a-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHC(N1)(N2)(N3)GTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-b-a-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGSAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-b-b-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGTCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-b-c-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGGTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-b-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGSAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-b-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGTCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_b-b-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:09153739)(N2:37250830)(N3:00402040)DMCGCGDHCKGGGTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-a-a-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)SAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACC
TGACC
Bc_a-a-b-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)TCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
Bc_a-a-c-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)GTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACC
TGACC
Bc_a-a-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)SAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-a-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)TCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC

Bc_a-a-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)GTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-b-a-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGSAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-b-b-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGTCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-b-c-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGGTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-b-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGSAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-b-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGTCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_a-b-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAMDG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGGTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-a-a-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)SAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
Bc_b-a-b-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)TCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
Bc_b-a-c-a
GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)GTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCT
GACC
Bc_b-a-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)SAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-a-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)TCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-a-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHC(N1)(N2)(N3)GTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-b-a-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGSAGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-b-b-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGTCCATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-b-c-a GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGGTGATC(N1)(N2)(N3)(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-b-a-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGSAGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-b-b-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGTCCATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
Bc_b-b-c-b GGTTCTGCTAGCGCCGAAGCAAAATACWMCAAAGAAKKG(N1:18223822)(N2:55250317)(N3:00353035)DMCGCGDHCKGGGTGATCBGG(N1)(N2)(N3)CTGCCGAACCTGACC
B_a-a CTGCCGAACCTGACCSNS(N1:09153739)(N2:37250830)(N3:00402040)CAGARA(N1)(N2)(N3)GCATTCATC(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG
B_a-b CTGCCGAACCTGACCSNS(N1:09153739)(N2:37250830)(N3:00402040)CAGARA(N1)(N2)(N3)GCATTC(N1)(N2)(N3)(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG
B_b-a CTGCCGAACCTGACCSNS(N1:09153739)(N2:37250830)(N3:00402040)CAGVTR(N1)(N2)(N3)GCATTCATC(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG
B_b-b CTGCCGAACCTGACCSNS(N1:09153739)(N2:37250830)(N3:00402040)CAGVTR(N1)(N2)(N3)GCATTC(N1)(N2)(N3)(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG
Bc_a-3 CTGCCGAACCTGACCSNS(N1:18223822)(N2:55250317)(N3:00353035)CAGARA(N1)(N2)(N3)GCATTCATC(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG
Bc_a-3 CTGCCGAACCTGACCSNS(N1:18223822)(N2:55250317)(N3:00353035)CAGARA(N1)(N2)(N3)GCATTC(N1)(N2)(N3)(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG
Bc_b-3 CTGCCGAACCTGACCSNS(N1:18223822)(N2:55250317)(N3:00353035)CAGVTR(N1)(N2)(N3)GCATTCATC(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG
Bc_b-3 CTGCCGAACCTGACCSNS(N1:18223822)(N2:55250317)(N3:00353035)CAGVTR(N1)(N2)(N3)GCATTC(N1)(N2)(N3)(N1)(N2)(N3)GCACTG(N1)(N2)(N3)RACGACCCGTCCCAGAGCTCTG

Figure S4-5: Biophysical characterization. (A) Circular dichroism was used to assess secondary structure and thermal denaturation for several clones. The molar ellipticity of a representative affibody clone is shown over a range of temperatures and wavelengths. The thermal stability (i.e. midpoint of thermal denaturation) was obtained while monitoring a wavelength of 220nm. All clones demonstrated the ability to refold upon cooling, as shown on the right. (B) Flow cytometry was used to isolate clones having high-affinity. High stringency populations were obtained using 100-150 nM target protein labelling conditions and collecting all double positive events (red box). From the high stringency population, an additional sort was done under more stringent conditions (5 nM target) to yield the ultra-high stringency populations (red triangle). Representative binding populations are shown.

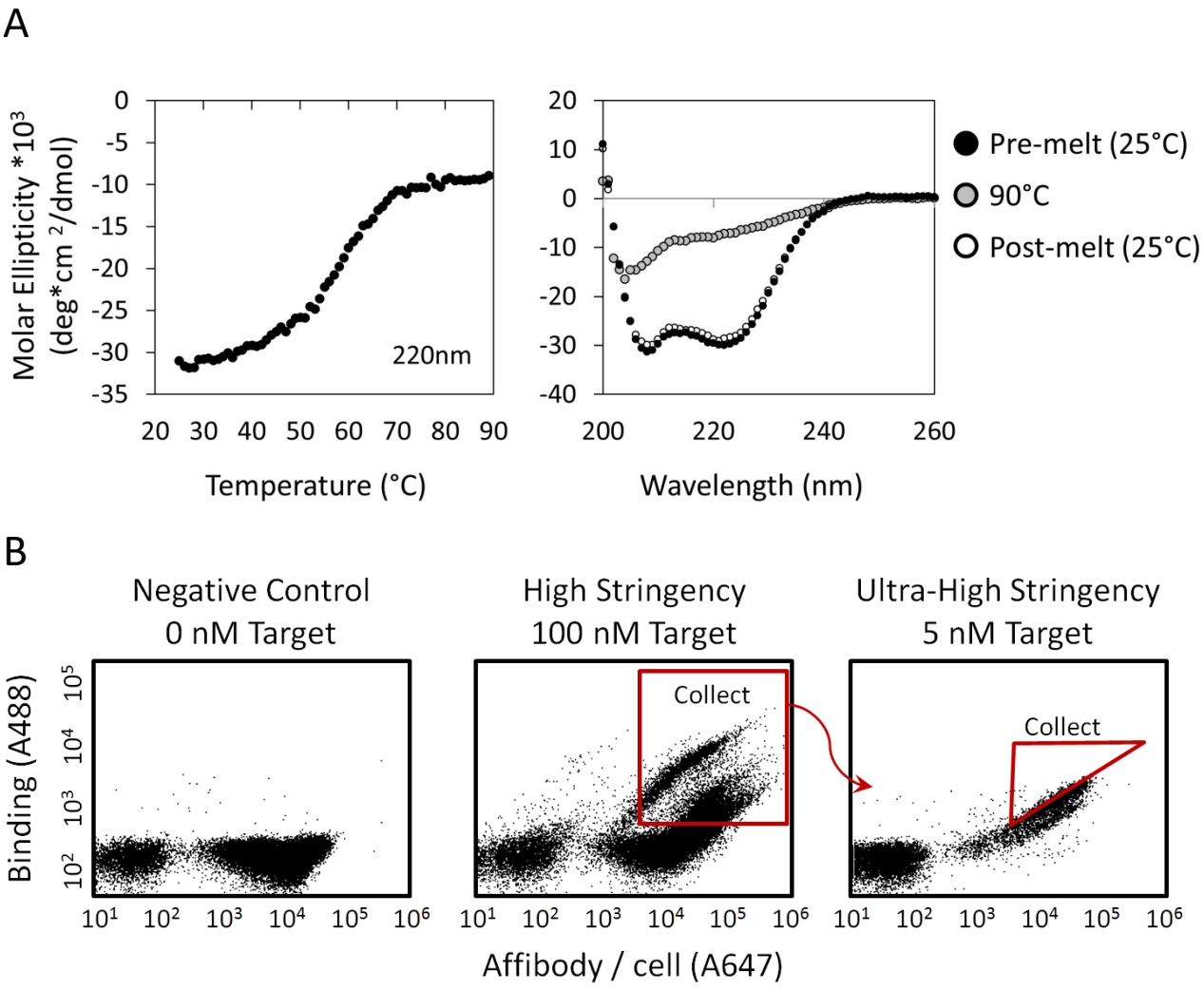


Figure S4-6: Affinity titration of high (A) and ultra-high (B) stringency variants. Randomly selected variants were isolated using either high (100-150 nM) or ultra-high (5 nM) stringency sorting conditions. Each variant was incubated at a range of concentrations, then analyzed via flow cytometry to measure the fraction of displayed protein bound to the target of interest. Panel A indicates the binding affinities (K_D; mean ± standard error; n=3) of representative clones from each target campaign. Panel B shows binding affinities for two representative unique variants from three target binding campaigns.

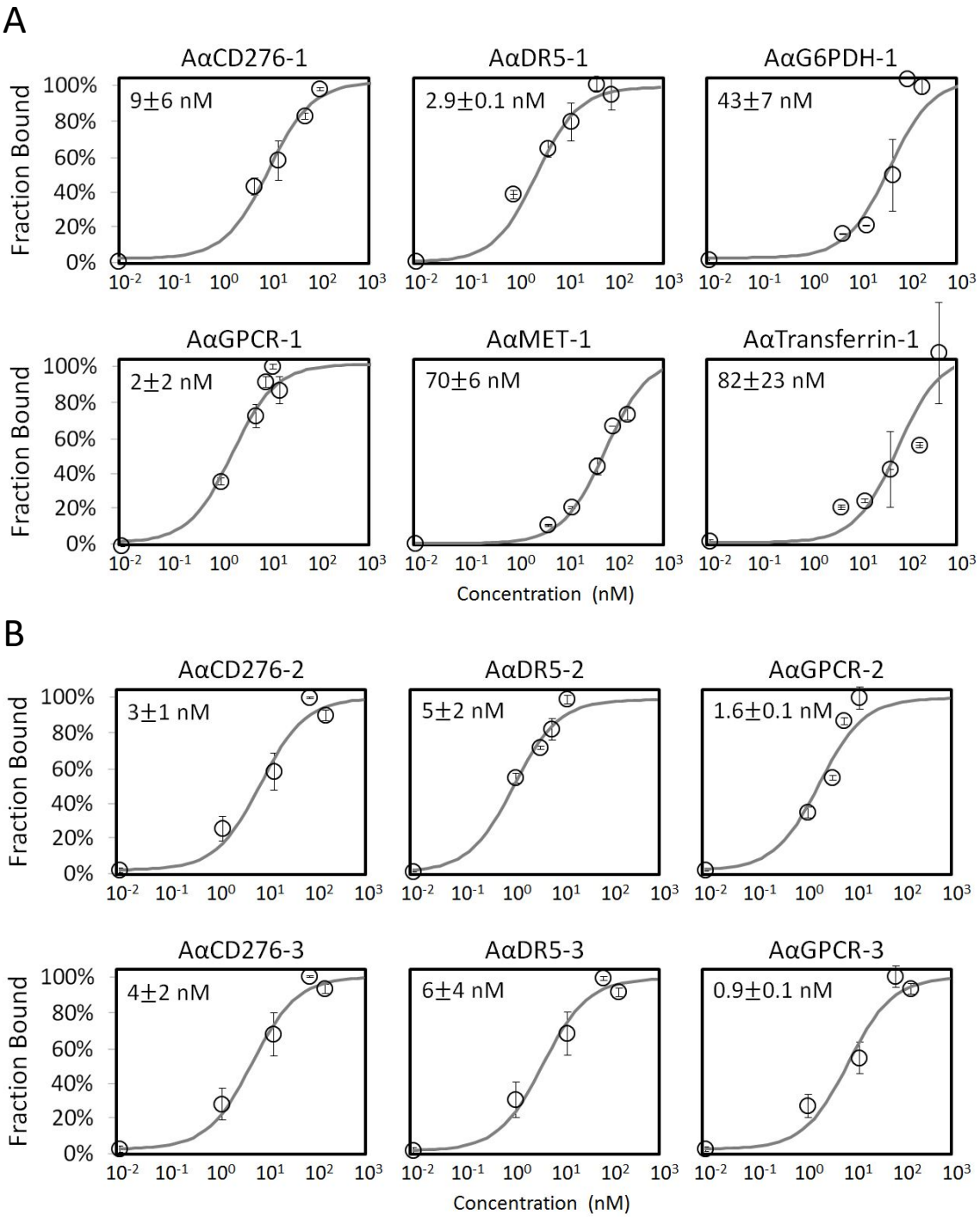


Figure S4-7: Library of origin comparison for high and ultra-high stringency variants. The posterior probabilities of individual strong binding variants having originated from the NNK, GS, or GS_{LC} library were calculated. High-throughput sequencing analyzed the high stringency dataset while Sanger sequencing was used for the ultra-high stringency set. The change in frequency indicates the difference in likelihood of observing a library member within the evolved binding population versus the initial library, $f_{Binders} - f_{Initial}$. A positive change in frequency reflects the preference of a particular library among highly functional variants. Statistical significance (*, $p < 0.001$) is shown between the change in frequency observed for the GS and NNK libraries as well as the extent of GS preference between the high and ultra-high stringency populations.

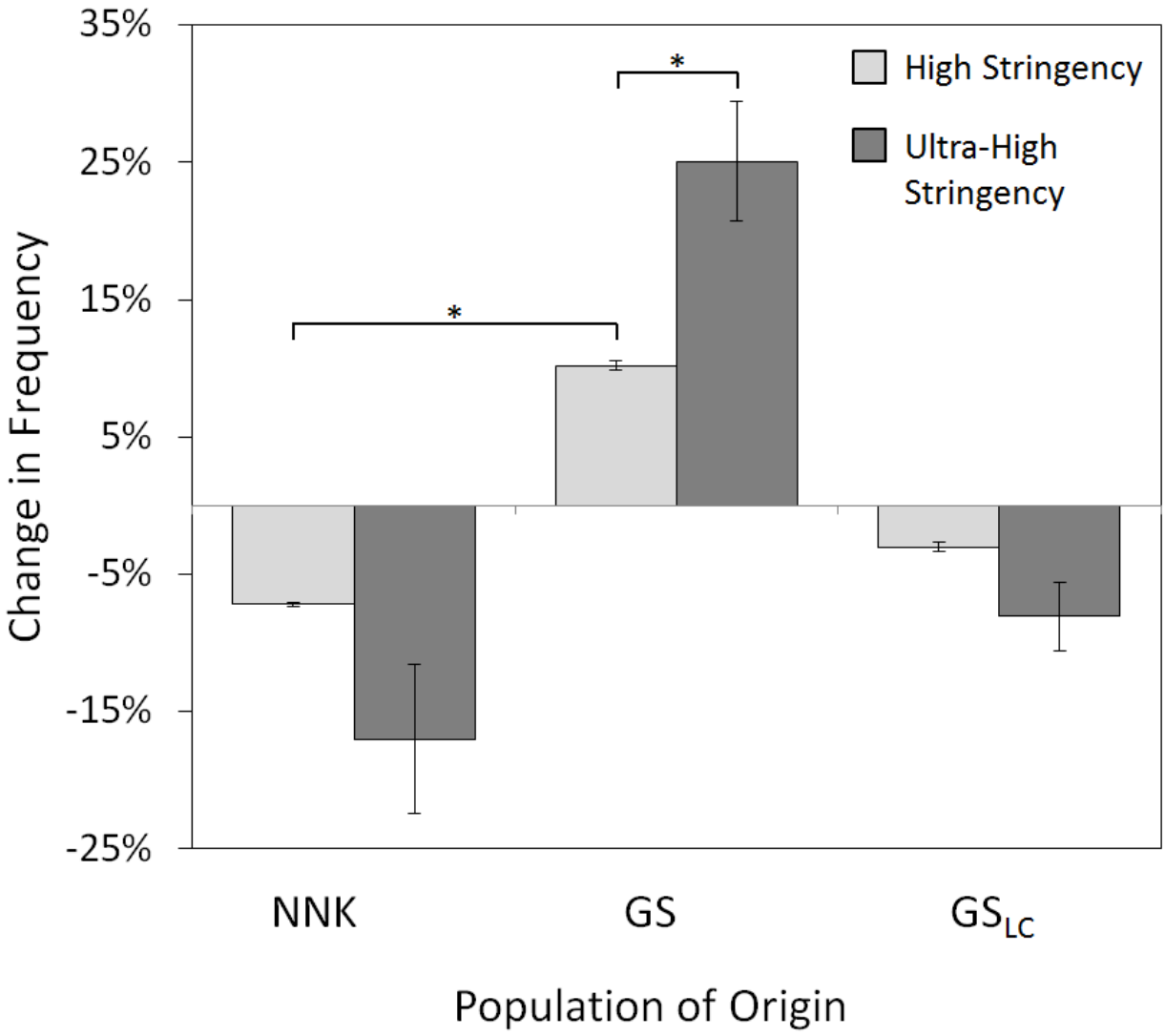


Figure S4-8: Specificity analysis for high stringency variants. Yeast displaying, separately, one representative evolved clone from each target binding campaign, were labeled with either the target of interest (50 nM) or non-target protein (100 nM). Binding was detected by flow cytometry. Note that the G6PDH binder was screened against only three non-target proteins. For calculating the fraction bound, maximum binding signal was established via the median fluorescence of strong binding variants under saturating conditions.

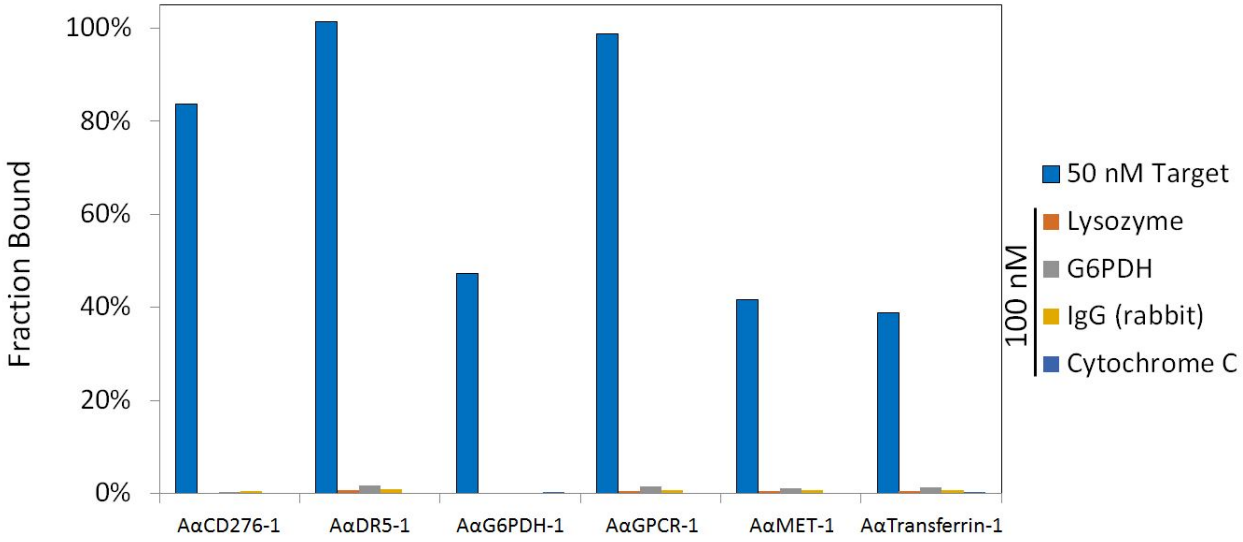
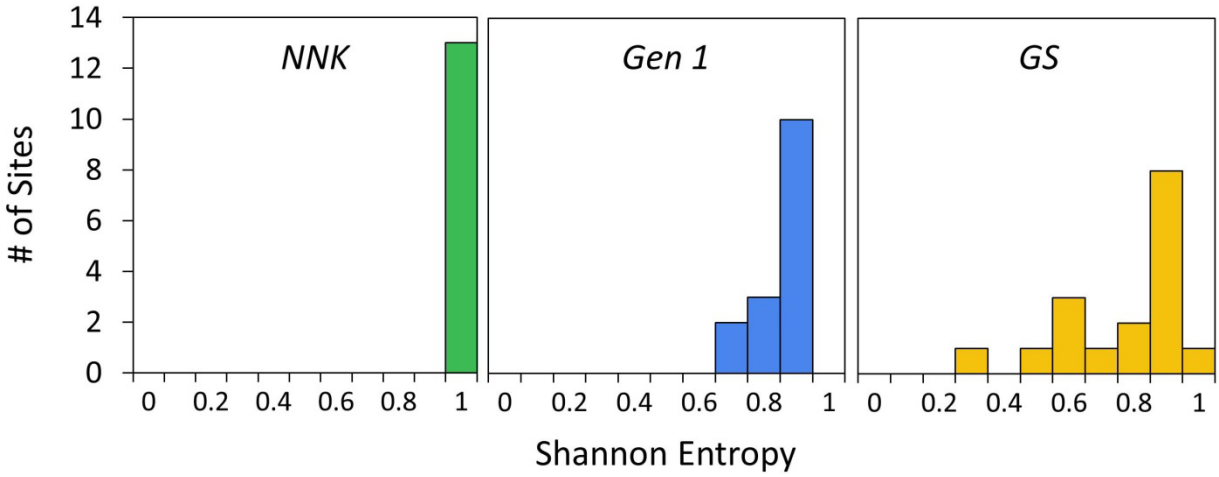


Figure S4-9: Diversity distributions. The diversity at each site is quantified as Shannon entropy for the three library designs. Traditional affibody libraries diversify thirteen sites, each with the NNK codon (Shannon entropy = 0.97). The first generation design offered a gradient of diversities at fifteen sites where Shannon entropy ranged from 0.68 to 0.87, averaging 0.82. The second generation (GS) design included a gradient of diversities at seventeen sites. Shannon entropy of sites within the GS design ranged from 0.23 to 0.92, averaging 0.74.



Chapter 5 – Concluding Remarks and Future Work

5.1. Designed constrained libraries are more efficient at protein discovery and yield more stable molecules

In the collection of work described here, the impact of sitewise diversification is thoroughly assessed by discovering and evolving high affinity binders in a high-throughput manner using combinatorial libraries. The generalizability of this approach is further validated by using multiple small protein scaffolds: 10 kDa beta-sandwich fibronectin with diversified solvent exposed loops and a 7 kDa three-helix bundle protein with diversified helices. In each case, a gradient of sitewise diversity was observed within the enriched binding populations. These results demonstrating constrained diversity are consistent with the binding interfaces found within natural ligand libraries *(i.e.* antibody repertoires), yet have been mostly absent from library design schemes of synthetic proteins prior to this work. Our results indicate that a constrained library design, compared to broad, uniform diversity, generates a larger variety of lead clones able to specifically bind a broad collection of targets. Within an under constrained and overly destabilized library setting, while the average library member performs suboptimally, rare sequences with stabilizing features such as disulfide bonds are dominantly enriched during selection. To overcome the need for disulfide-rich variants, several data sets (high-throughput sequences from evolved populations, computed stability, phylogenetic analysis, etc.) were collectively used to guide amino acid diversity at individual sites. The resulting library, consisting of sitewise amino acid bias, yields evolved variants with higher thermal stability. This provides evidence in support of the hypothesis that reduced destabilization improves evolvability. That is, a higher fraction of folded variants is achievable using a combinatorial

library containing variants that are less destabilized. Based on the ability to discover a broader collection of lead variants and the tendency of those variants to have superior thermal stability, we conclude that a gradient of sitewise diversity is more efficient at evolving high-affinity binding ligands.

Our high-throughput discovery approach was able to demonstrate that a single protein target can yield hundreds of diverse families of binders (i.e. subpopulation of sequences sharing a common motif with >80% sequence homology among members of the cluster). The distinct features between each sequence family suggest that a variety of epitopes are being offered by the target. Having several options for lead clones that interact with unique target epitopes improves the likelihood of being able to elicit the intended effect.[264] This presumption of diverse epitopes would benefit from further characterization. This could be accomplished using epitope mapping of the target protein to determine where binding interactions are occurring. With this method, a library of target variants containing point mutations is displayed on the surface of yeast. Analysis of mutations that disrupt binding interactions can be used to identify the precise epitope location.[265–267] An alternative approach could rely on FRET to determine the binding event configuration. However, with a typical detection range of 1–10 nm, this method would make it challenging to discern closely adjacent epitopes.[268–270]

Building on the progress discussed in this thesis, many future directions would further benefit protein engineering and improve our ability to discover highly functional proteins. Within the diverse sequence data sets, epistatic interactions between multiple sites should be explored further and incorporated into the design process. Pair-wise and higher order cooperative cohorts can be identified using metrics such as mutual information[271],

maximum-entropy probability models[272,273], and statistical-coupling analysis[169]. Diversification of site clusters can then be designed on the DNA level to ensure that correlated amino acids are implemented in tandem.

This thesis work addresses fundamental concepts involving an efficient search through sequence space in pursuit of stable proteins with novel function. It is, therefore, reasonable to expect that the conclusions derived from discovering binding ligands can be extended to enzyme engineering as well. Combinatorial libraries are widely used for increasing stability and *improving* enzymatic activity[274] but examples of engineering truly novel function are limited.[275] The field would benefit from implementing our approach to constrained sitewise library design in the context of *discovering* new catalytic function.

The field of protein engineering has brought forth a wealth of knowledge with far reaching consequences. Society has gained therapeutic ligands for clinical treatments, molecular recognition agents for detecting and monitoring diseases, as well as a much deeper understanding of fundamental molecular interactions. Nevertheless, there remains a tremendous scope of disease treatments and novel applications that have not yet been realized. Inevitably, the efficiency at which the field is able to identify problems and discover solutions will determine the pace of future progress. The methodology and accomplishments outlined in this thesis make direct contributions to improving the efficiency at which we discover proteins with novel function to address these challenges.

References

1. James ML, Gambhir SS. A Molecular Imaging Primer: Modalities, Imaging Agents, and Applications. *Physiol Rev*. 2012;92(2):897-965. doi:10.1152/physrev.00049.2010.
2. Carter PJ. Introduction to current and future protein therapeutics: A protein engineering perspective. *Exp Cell Res*. 2011;317(9):1261-1269. doi:10.1016/j.yexcr.2011.02.013.
3. Lagassé HAD, Alexaki A, Simhadri VL, et al. Recent advances in (therapeutic protein) drug development. *F1000Research*. 2017;6:113. doi:10.12688/f1000research.9970.1.
4. Kinch MS. An overview of FDA-approved biologics medicines. *Drug Discov Today*. 2015;20(4):393-398. doi:10.1016/j.drudis.2014.09.003.
5. Leader B, Baca QJ, Golan DE. Protein therapeutics: a summary and pharmacological classification. *Nat Rev Drug Discov*. 2008;7(1):21-39. doi:10.1038/nrd2399.
6. Stern LA, Case BA, Hackel BJ. Alternative Non-Antibody Protein Scaffolds for Molecular Imaging of Cancer. *Curr Opin Chem Eng*. 2013;2(4):425-432. doi:10.1016/j.coche.2013.08.009.
7. Weissleder R, Pittet MJ. Imaging in the era of molecular oncology. *Nature*. 2008;452(7187):580-589. doi:10.1038/nature06917.
8. Ellenbroek SIJ, van Rheenen J. Imaging hallmarks of cancer in living mice. *Nat Rev Cancer*. 2014;14(6):406-418. doi:10.1038/nrc3742.
9. Grönwall C, Ståhl S. Engineered affinity proteins-Generation and applications. *J Biotechnol*. 2009;140(3-4):254-269. doi:10.1016/j.jbiotec.2009.01.014.
10. Milo R, Last RL. Achieving Diversity in the Face of Constraints: Lessons from Metabolism. *Science (80-)*. 2012;336(6089):1663-1667. doi:10.1126/science.1217665.
11. Bar-Even A, Noor E, Savir Y, et al. The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*. 2011;50(21):4402-4410. doi:10.1021/bi2002289.
12. Currin A, Swainston N, Day PJ, Kell DB. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev*. 2015;44(5):1172-1239. doi:10.1039/C4CS00351A.
13. Moretti R, Fleishman SJ, Agius R, et al. Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins*. 2013;81(11):1980-1987. doi:10.1002/prot.24356.
14. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein Eng Des Sel*. 2009;22(9):553-560. doi:10.1093/protein/gzp030.
15. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*. 2006;7:61-80. doi:10.1146/annurev.genom.7.080505.115630.
16. Pucci F, Bourgeas R, Rooman M. Predicting protein thermal stability changes

upon point mutations using statistical potentials: Introducing HoTMuSiC. *Sci Rep*. 2016;6(February):23257. doi:10.1038/srep23257.
17. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol*. 2009;10(12):866-876. doi:10.1038/nrm2805.
18. Daugherty PS, Chen G, Iverson BL, Georgiou G. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc Natl Acad Sci U S A*. 2000;97(5):2029-2034. doi:10.1073/pnas.030527597.
19. Guo HH, Choe J, Loeb LA. Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A*. 2004;101(25):9205-9210. doi:10.1073/pnas.0403255101.
20. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*. 2006;444(7121):929-932. doi:10.1038/nature05385.
21. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*. 2009;19(5):596-604.
22. Pace CN. The stability of globular proteins. *CRC Crit Rev Biochem*. 1975;3(1):1-43. http://www.ncbi.nlm.nih.gov/pubmed/238787.
23. Fersht AR, Serrano L. Principles of protein stability derived from protein engineering experiments. *Curr Opin Struct Biol*. 1993;3(1):75-83. doi:10.1016/0959-440X(93)90205-Y.
24. Taverna DM, Goldstein RA. Why are proteins marginally stable? *Proteins Struct Funct Genet*. 2002;46(1):105-109. doi:10.1002/prot.10016.
25. Williams PD, Pollock DD, Goldstein R a. Functionality and the evolution of marginal stability in proteins: inferences from lattice simulations. *Evol Bioinform Online*. 2006;2:91-101.
26. Anfinsen CB. Principles that Govern the Folding of Protein Chains. *Science (80-).* 1973;181(4096):223-230. doi:10.1126/science.181.4096.223.
27. Aharoni A, Gaidukov L, Khersonsky O, McQ Gould S, Roodveldt C, Tawfik DS. The “evolvability” of promiscuous protein functions. *Nat Genet*. 2005;37(1):73-76. doi:10.1038/ng1482.
28. Nikolova P V., Woong K-P, DeDecker B, Henckel J, Fersht a R. Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations. *EMBO J*. 2000;19(3):370-378. doi:10.1093/emboj/19.3.370.
29. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA*. 2005;102. doi:10.1073/pnas.0406744102.
30. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A*. 2006;103(15):5869-5874. doi:10.1073/pnas.0510098103.
31. Chen R, Greer a, Dean a M. A highly active decarboxylating dehydrogenase with rationally inverted coenzyme specificity. *Proc Natl Acad Sci U S A*. 1995;92(25):11666-11670. doi:10.1073/pnas.92.25.11666.
32. Chaparro-Riggers JF, Polizzi KM, Bommarius AS. Better library design: Data-driven protein engineering. *Biotechnol J*. 2007;2(2):180-191. doi:10.1002/biot.200600170.
33. Rollence ML, Filpula D, Pantoliano MW, Bryan PN. Engineering thermostability

in subtilisin BPN' by in vitro mutagenesis. *Crit Rev Biotechnol*. 1988;8(3):217-224. doi:10.3109/07388558809147558.
34. Estell DA, Graycar TP, Wells JA. Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J Biol Chem*. 1985;260(11):6518-6521.
35. Steipe B, Schiller B, Pluckthun A, Steinbacher S. Sequence statistics reliably predict stabilizing mutations in a protein domain. *J Mol Biol*. 1994;240(3):188-192.
36. Wang XX, Shusta E V. The use of scFv-displaying yeast in mammalian cell surface selections. *J Immunol Methods*. 2005;304(1-2):30-42. doi:10.1016/j.jim.2005.05.006.
37. Ackerman M, Levary D, Tobon G, Hackel BJ, Orcutt KD, Wittrup KD. Highly avid magnetic bead capture: an efficient selection method for de novo protein engineering utilizing yeast surface display. *Biotechnol Prog*. 2009;25(3):774-783. doi:10.1021/bp.174.
38. Dodevski I, Markou GC, Sarkar CA. Conceptual and methodological advances in cell-free directed evolution. *Curr Opin Struct Biol*. 2015;33:1-7. doi:10.1016/j.sbi.2015.04.008.
39. Packer MS, Liu DR. Methods for the directed evolution of proteins. *Nat Rev Genet*. 2015;16(7):379-394. doi:10.1038/nrg3927.
40. Gray JJ, Moughon S, Wang C, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*. 2003;331(1):281-299. doi:10.1016/S0022-2836(03)00670-3.
41. Tinberg CE, Khare SD, Dou J, et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. 2013;501(7466):212-216. doi:10.1038/nature12443.
42. Smith CA, Kortemme T. Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. *PLoS One*. 2011;6(7):e20451.
43. Au L, Green DF. Direct Calculation of Protein Fitness Landscapes through Computational Protein Design. *Biophys J*. 2016;110(1):75-84. doi:10.1016/j.bpj.2015.11.029.
44. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J Mol Biol*. 2002;320(2):369-387. doi:10.1016/S0022-2836(02)00442-4.
45. Magliery TJ. Protein stability: Computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol*. 2015;33:161-168. doi:10.1016/j.sbi.2015.09.002.
46. Seeger MA, Zbinden R, Flütsch A, et al. Design, construction, and characterization of a second-generation DARPin library with reduced hydrophobicity. *Protein Sci*. 2013;22(9):1239-1257.
47. Fellouse FA, Wiesmann C, Sidhu SS. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci U S A*. 2004;101(34):12467-12472. doi:10.1073/pnas.0401786101.
48. Fellouse FA, Li B, Compaan DM, Peden AA, Hymowitz SG, Sidhu SS. Molecular recognition by a binary code. *J Mol Biol*. 2005;348(5):1153-1162.

doi:10.1016/j.jmb.2005.03.041.
49. Koide S, Sidhu SS. The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem Biol*. 2009;4(5):325-334.
50. Wojcik J, Hantschel O, Grebien F, et al. A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. *Nat Struct Mol Biol*. 2010;436(4):519-527.
51. Hackel BJ, Ackerman ME, Howland SW, Wittrup KD. Stability and CDR Composition Biases Enrich Binder Functionality Landscapes. *J Mol Biol*. 2010;401(1):84-96.
52. Mian IS, Bradwell AR, Olson AJ. Structure, function and properties of antibody binding sites. *J Mol Biol*. 1991;217(1):133-151. doi:10.1016/0022-2836(91)90617-F.
53. Fellouse FA, Esaki K, Birtalan S, et al. High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J Mol Biol*. 2007;373(4):924-940.
54. Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J Mol Biol*. 2008;377(5):1518-1528.
55. Birtalan S, Fisher RD, Sidhu SS. The functional capacity of the natural amino acids for molecular recognition. *Mol Biosyst*. 2010;6(7):1186-1194. doi:10.1039/b927393j.
56. Lee C V, Liang W-C, Dennis MS, Eigenbrot C, Sidhu SS, Fuh G. High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J Mol Biol*. 2004;340(5):1073-1093.
57. Woldring DR, Holec P V, Zhou H, Hackel BJ. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. Gill AC, ed. *PLoS One*. 2015;10(9):e0138956. doi:10.1371/journal.pone.0138956.
58. Zemlin M, Klinger M, Link J, et al. Expressed Murine and Human CDR-H3 Intervals of Equal Length Exhibit Distinct Repertoires that Differ in their Amino Acid Composition and Predicted Range of Structures. *J Mol Biol*. 2003;334(4):733-749. doi:10.1016/j.jmb.2003.10.007.
59. Binz HK, Amstutz P, Kohl A, et al. High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat Biotechnol*. 2004;22(5):575-582.
60. Schilling J, Schöppe J, Plückthun A. From DARPins to LoopDARPins: novel LoopDARPin design allows the selection of low picomolar binders in a single round of ribosome display. *J Mol Biol*. 2014;426(3):691-721.
61. Woldring DR, Holec P V, Zhou H, Hackel BJ. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One*. 2015;10(9):e0138956. doi:10.1371/journal.pone.0138956.
62. Woldring DR, Holec P V., Hackel BJ. ScaffoldSeq: Software for characterization of directed evolution populations. *Proteins Struct Funct Bioinforma*. 2016;84(7):869-874. doi:10.1002/prot.25040.
63. Arnold FH. Fancy footwork in the sequence space shuffle. *Nat Biotechnol*.

2006;24(3):328-330. doi:10.1038/nbt0306-328.
64. Karanicolas J, Corn JE, Chen I, et al. A de novo protein binding pair by computational design and directed evolution. *Mol Cell*. 2011;42(2):250-260.
65. Fleishman SJ, Whitehead T a, Ekiert DC, et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*. 2011;332(6031):816-821. doi:10.1126/science.1202617.
66. Khoury G a., Smadbeck J, Kieslich C a., Floudas C a. Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol*. 2014;32(2):99-109. doi:10.1016/j.tibtech.2013.10.008.
67. Dellus-Gur E, Toth-Petroczy A, Elias M, Tawfik DS. What makes a protein fold amenable to functional innovation? fold polarity and stability trade-offs. *J Mol Biol*. 2013;425(14):2609-2621. doi:10.1016/j.jmb.2013.03.033.
68. Stern LA, Case BA, Hackel BJ. Alternative non-antibody protein scaffolds for molecular imaging of cancer. *Curr Opin Chem Eng*. 2013;2(4):425-432. doi:10.1016/j.coche.2013.08.009.
69. Chen J, Sawyer N, Regan L. Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci*. 2013;22(4):510-515. doi:10.1002/pro.2230.
70. Zhai W, Glanville J, Fuhrmann M, et al. Synthetic antibodies designed on natural sequence landscapes. *J Mol Biol*. 2011;412(1):55-71.
71. Knappik A, Ge L, Honegger A, et al. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol*. 2000;296(1):57-86. doi:10.1006/jmbi.1999.3444.
72. Prassler J, Thiel S, Pracht C, et al. HuCAL PLATINUM, a synthetic fab library optimized for sequence diversity and superior performance in mammalian expression systems. *J Mol Biol*. 2011;413(1):261-278. doi:10.1016/j.jmb.2011.08.012.
73. Sidhu SS, Li B, Chen Y, Fellouse FA, Eigenbrot C, Fuh G. Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J Mol Biol*. 2004;338(2):299-310.
74. Fellouse FA, Esaki K, Birtalan S, et al. High-throughput Generation of Synthetic Antibodies from Highly Functional Minimalist Phage-displayed Libraries. *J Mol Biol*. 2007;373(4):924-940. doi:10.1016/j.jmb.2007.08.005.
75. Grönwall C, Jonsson A, Lindström S, Gunneriusson E, Ståhl S, Herne N. Selection and characterization of Affibody ligands binding to Alzheimer amyloid beta peptides. *J Biotechnol*. 2007;128(1):162-183.
76. Correa A, Pacheco S, Mechaly AE, et al. Potent and specific inhibition of glycosidases by small artificial binding proteins (Affitins). *PLoS One*. 2014;9(5). doi:10.1371/journal.pone.0097438.
77. Béhar G, Bellinzoni M, Maillasson M, et al. Tolerance of the archaeal Sac7d scaffold protein to alternative library designs: characterization of anti-immunoglobulin G Affitins. *Protein Eng Des Sel*. 2013;26(4):267-275. doi:10.1093/protein/gzs106.
78. Getz J a., Rice JJ, Daugherty PS. Protease-resistant peptide ligands from a knottin scaffold library. *ACS Chem Biol*. 2011;6(8):837-844. doi:10.1021/cb200039s.

79. Moore SJ, Cochran JR. Engineering knottins as novel binding agents. *Methods Enzymol*. 2012;503:223-251. doi:10.1016/B978-0-12-396962-0.00009-4.
80. Gebauer M, Schiefner A, Matschiner G, Skerra A. Combinatorial Design of an Anticalin Directed against the Extra-Domain B for the Specific Targeting of Oncofetal Fibronectin. *J Mol Biol*. 2013;425(4):780-802.
81. Schlatter D, Brack S, Banner DW, et al. Generation, characterization and structural data of chymase binding proteins based on the human Fyn kinase SH3 domain. *MAbs*. 2012;4(4):497-508. doi:10.4161/mabs.20452.
82. Gera N, Hussain M, Wright RC, Rao BM. Highly stable binding proteins derived from the hyperthermophilic Sso7d scaffold. *J Mol Biol*. 2011;409(4):601-616.
83. Steemson JD, Baake M, Rakonjac J, Arcus VL, Liddament MT. Tracking Molecular Recognition at the Atomic Level with a New Protein Scaffold Based on the OB-Fold. Dübel S, ed. *PLoS One*. 2014;9(1):e86050. doi:10.1371/journal.pone.0086050.
84. Barbas CF, Bain JD, Hoekstra DM, Lerner RA. Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc Natl Acad Sci U S A*. 1992;89(10):4457-4461. doi:10.1073/pnas.89.10.4457.
85. Koide A, Bailey CW, Huang X, Koide S. The fibronectin type III domain as a scaffold for novel binding proteins. *J Mol Biol*. 1998;284(4):1141-1151. doi:10.1006/jmbi.1998.2238.
86. Lipovsek D. Adnectins: engineered target-binding protein therapeutics. *Protein Eng Des Sel*. 2011;24(1-2):3-9.
87. Koide A, Wojcik J, Gilbreth RN, Hoey RJ, Koide S. Teaching an old scaffold new tricks: Monobodies constructed using alternative surfaces of the FN3 scaffold. *J Mol Biol*. 2012;415(2):393-405. doi:10.1016/j.jmb.2011.12.019.
88. Diem MD, Hyun L, Yi F, et al. Selection of high-affinity Centyrin FN3 domains from a simple library diversified at a combination of strand and loop positions. *Protein Eng Des Sel*. 2014;27(10):419-429. doi:10.1093/protein/gzu016.
89. Wojcik J, Hantschel O, Grebien F, et al. A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. *Nat Struct Mol Biol*. 2010;17(4):519-527. doi:10.1038/nsmb.1793.
90. Koide A, Wojcik J, Gilbreth RN, Reichel A, Piehler J, Koide S. Accelerating phage-display library selection by reversible and site-specific biotinylation. *Protein Eng Des Sel*. 2009;22(11):685-690. doi:10.1093/protein/gzp053.
91. Hackel BJ, Ackerman ME, Howland SW, Wittrup KD. Stability and CDR Composition Biases Enrich Binder Functionality Landscapes. *J Mol Biol*. 2010;401(1):84-96. doi:10.1016/j.jmb.2010.06.004.
92. Lipovšek D, Lippow SM, Hackel BJ, et al. Evolution of an Interloop Disulfide Bond in High-Affinity Antibody Mimics Based on Fibronectin Type III Domain and Selected by Yeast Surface Display: Molecular Convergence with Single-Domain Camelid and Shark Antibodies. *J Mol Biol*. 2007;368(4):1024-1041. doi:10.1016/j.jmb.2007.02.029.
93. Hackel BJ, Kapila A, Dane Wittrup K. Picomolar Affinity Fibronectin Domains Engineered Utilizing Loop Length Diversity, Recursive Mutagenesis, and Loop Shuffling. *J Mol Biol*. 2008;381(5):1238-1252. doi:10.1016/j.jmb.2008.06.051.
94. Sullivan M, Brooks L, Weidenborner P. Anti-Idiotypic Monobodies Derived from

a Fibronectin Scaffold. *Biochemistry*. 2013;52(10):1802-1813. doi:10.1021/bi3016668.
95. Liao H-I, Olson CA, Hwang S, et al. mRNA display design of fibronectin-based intrabodies that detect and inhibit sars-cov N protein. *J Biol Chem*. 2009;284(26):M901547200. doi:10.1074/jbc.M901547200.
96. Gilbreth RN, Truong K, Madu I, et al. Isoform-specific monobody inhibitors of small ubiquitin-related modifiers engineered using structure-guided library design. *Proc Natl Acad Sci U S A*. 2011;108(19):7751-7756. doi:10.1073/pnas.1102294108.
97. Tamaskovic R, Simon M, Stefan N, Schwill M, Plückthun A. Designed ankyrin repeat proteins (DARPins): From research to therapy. *Methods Enzymol*. 2012;503:101-134. doi:10.1016/B978-0-12-396962-0.00005-7.
98. Grimm S, Yu F, Nygren PÅ. Ribosome display selection of a murine IgG1 fab binding affibody molecule allowing species selective recovery of monoclonal antibodies. *Mol Biotechnol*. 2011;48(3):263-276. doi:10.1007/s12033-010-9367-1.
99. Gebauer M, Skerra A. Anticalins: Small engineered binding proteins based on the lipocalin scaffold. *Methods Enzymol*. 2012;503:157-188. doi:10.1016/B978-0-12-396962-0.00007-0.
100. Reichmann D, Cohen M, Abramovich R, et al. Binding Hot Spots in the TEM1-BLIP Interface in Light of its Modular Architecture. *J Mol Biol*. 2007;365:663-679. doi:10.1016/j.jmb.2006.09.076.
101. Schreiber G, Fersht AR. Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles. *J Mol Biol*. 1995;248(2):478-486. doi:10.1016/S0022-2836(95)80064-6.
102. Dall'Acqua W, Goldman ER, Eisenstein E, Mariuzza RA. A Mutational Analysis of the Binding of Two Different Proteins to the Same Antibody. *Biochemistry*. 1996;35(30):9667-9676. doi:10.1021/bi960819i.
103. Cunningham BC, Wells JA. Comparison of a structural and a functional epitope. *J Mol Biol*. 1993;234(3):554-563. doi:10.1006/jmbi.1993.1611.
104. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. *Science (80-)*. 1995;267(5196):383-386. doi:10.1126/science.7529940.
105. Jones JT. Binding Interaction of the Heregulinbeta egf Domain with ErbB3 and ErbB4 Receptors Assessed by Alanine Scanning Mutagenesis. *J Biol Chem*. 1998;273(19):11667-11674. doi:10.1074/jbc.273.19.11667.
106. Hackel BJ, Sathirachinda A, Gambhir SS. Designed hydrophilic and charge mutations of the fibronectin domain: Towards tailored protein biodistribution. *Protein Eng Des Sel*. 2012;25(10):639-647. doi:10.1093/protein/gzs036.
107. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol*. 1998;280(1):1-9. doi:10.1006/jmbi.1998.1843.
108. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol*. 2002;12(1):14-20. doi:10.1016/S0959-440X(02)00283-X.
109. Whitehead TA, Chevalier A, Song Y, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol*. 2012;30(6):543-548. doi:10.1038/nbt.2214.
110. Ernst A, Gfeller D, Kan Z, et al. Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol Biosyst*.

2010;6(10):1782-1790. doi:10.1039/c0mb00061b.
111. Deng Z, Huang W, Bakkalbasi E, et al. Deep sequencing of systematic combinatorial libraries reveals ??-lactamase sequence constraints at high resolution. *J Mol Biol*. 2012;424(3-4):150-167. doi:10.1016/j.jmb.2012.09.014.
112. Ravn U, Gueneau F, Baerlocher L, et al. By-passing in vitro screening - Next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res*. 2010;38(21). doi:10.1093/nar/gkq789.
113. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol*. 1997;15(6):553-557. doi:10.1038/nbt0697-553.
114. Benatuil L, Perez JM, Belk J, Hsieh C-M. An improved yeast transformation method for the generation of very large human antibody libraries (supplementary info). *Protein Eng Des Sel*. 2010;23(4):9-10.
115. Ackerman M, Levary D, Tobon G, Hackel B, Orcutt KD, Wittrup KD. Highly avid magnetic bead capture: An efficient selection method for de novo protein engineering utilizing yeast surface display. *Biotechnol Prog*. 2009;25(3):774-783. doi:10.1002/btpr.174.
116. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD. Isolating and engineering human antibodies using yeast surface display. *Nat Protoc*. 2006;1(2):755-768. doi:10.1038/nprot.2006.94.
117. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res*. 2014;322(1):12-20. doi:10.1016/j.yexcr.2014.01.008.
118. Dabney J, Meyer M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*. 2012;52(2):87-94. doi:10.2144/000113809.
119. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*. 2012;13(1):31. doi:10.1186/1471-2105-13-31.
120. Sanner M. Python: a programming language for software integration and development. *J Mol Graph Model*. 1999;17:57-61.
121. Greenfield NJ. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc*. 2006;1(6):2876-2890. doi:10.1038/nprot.2006.202.
122. Xu L, Aha P, Gu K, et al. Directed evolution of high-affinity antibody mimics using mRNA display. *Chem Biol*. 2002;9(8):933-942. doi:10.1016/S1074-5521(02)00187-4.
123. Sonnhammer ELL, Eddy SR, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins Struct Funct Genet*. 1997;28(3):405-420. doi:10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L.
124. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res*. 2005;33(Web Server issue):W382-8. doi:10.1093/nar/gki387.
125. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. 2000;28(1):235-242.

126. Cota E, Hamill SJ, Fowler SB, Clarke J. Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. *J Mol Biol*. 2000;302:713-725. doi:10.1006/jmbi.2000.4053.
127. Fellouse FA, Barthelemy PA, Kelley RF, Sidhu SS. Tyrosine Plays a Dominant Functional Role in the Paratope of a Synthetic Antibody Derived from a Four Amino Acid Code. *J Mol Biol*. 2006;357(1):100-114. doi:10.1016/j.jmb.2005.11.092.
128. Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS. The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies. *J Mol Biol*. 2008;377(5):1518-1528. doi:10.1016/j.jmb.2008.01.093.
129. Koide A, Jordan MR, Horner SR, Batori V, Koide S. Stabilization of a Fibronectin Type III Domain by the Removal of Unfavorable Electrostatic Interactions on the Protein Surface. *Biochemistry*. 2001;40(34):10326-10333. doi:10.1021/bi010916y.
130. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(4):623-656.
131. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992;89(22):10915-10919. doi:10.1073/pnas.89.22.10915.
132. Clark JM. Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. *Nucleic Acids Res*. 1988;16(20):9677-9686. doi:10.1093/nar/16.20.9677.
133. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem*. 1993;17(2):149-163. doi:10.1016/0097-8485(93)85006-X.
134. Parker MH, Chen Y, Danehy F, et al. Antibody mimics based on human fibronectin type three domain engineered for thermostability and high-affinity binding to vascular endothelial growth factor receptor two. *Protein Eng Des Sel*. 2005;18(9):435-444. doi:10.1093/protein/gzi050.
135. Traxlmayr MW, Hasenhindl C, Hackl M, et al. Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. *J Mol Biol*. 2012;423(3):397-412. doi:10.1016/j.jmb.2012.07.017.
136. Finn RD, Bateman A, Clements J, et al. Pfam: The protein families database. *Nucleic Acids Res*. 2014;42(D1):D290-D301. doi:10.1093/nar/gkt1223.
137. Fraczkiewicz R, Braun W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comput Chem*. 1998;19(3):319-333. doi:10.1002/(SICI)1096-987X(199802)19:3<319::AID-JCC6>3.0.CO;2-W.
138. Pal G, Kossiakoff AA, Sidhu SS. The functional binding epitope of a high affinity variant of human growth hormone mapped by shotgun alanine-scanning mutagenesis: insights into the mechanisms responsible for improved affinity. *J Mol Biol*. 2003;332(1):195-204.
139. Ma B, Wolfson HJ, Nussinov R. Protein functional epitopes: Hot spots, dynamics and combinatorial libraries. *Curr Opin Struct Biol*. 2001;11:364-369. doi:10.1016/S0959-440X(00)00216-5.

140. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*. 2002;99:14116-14121. doi:10.1073/pnas.202485799.
141. Daugherty PS, Chen G, Iverson BL, Georgiou G. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc Natl Acad Sci U S A*. 2000;97(5):2029-2034. doi:10.1073/pnas.030527597.
142. Shafikhani S, Siegel RA, Ferrari E, Schellenberger V. Generation of large libraries of random mutants in Bacillus subtills by PCR-based plasmid multimerization. *Biotechniques*. 1997;23(2):304-310.
143. Walensky LD, Kung AL, Escher I, et al. Activation of apoptosis in vivo by a hydrocarbon-stapled BH3 helix. *Science*. 2004;305(5689):1466-1470. doi:10.1126/science.1099191.
144. Roux KH, Greenberg AS, Greene L, et al. Structural analysis of the nurse shark (new) antigen receptor (NAR): molecular convergence of NAR and unusual mammalian immunoglobulins. *Proc Natl Acad Sci U S A*. 1998;95(September):11804-11809. doi:10.1073/pnas.95.20.11804.
145. Muyldermans S, Atarhouch T, Saldanha J, Barbosa JA, Hamers R. Sequence and structure of VH domain from naturally occurring camel heavy chain immunoglobulins lacking light chains. *Protein Eng*. 1994;7(9):1129-1135. doi:10.1093/protein/7.9.1129.
146. DeKosky BJ, Ippolito GC, Deschner RP, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol*. 2013;31(2):166-169. doi:10.1038/nbt.2492.
147. Reddy ST, Ge X, Miklos AE, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol*. 2010;28(9):965-969. doi:10.1038/nbt.1673.
148. Tse E, Lobato MN, Forster A, Tanaka T, Chung GTY, Rabbitts TH. Intracellular antibody capture technology: application to selection of intracellular antibodies recognising the BCR-ABL oncogenic proteina. *J Mol Biol*. 2002;317(1):85-94. doi:10.1006/jmbi.2002.5403.
149. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods*. 2014;11(8):801-807. doi:10.1038/nmeth.3027.
150. Tripathi A, Varadarajan R. Residue specific contributions to stability and activity inferred from saturation mutagenesis and deep sequencing. *Curr Opin Struct Biol*. 2014;24:63-71.
151. Kruziki MA, Bhatnagar S, Woldring DR, Duong VT, Hackel BJ. A 45-Amino-Acid Scaffold Mined from the PDB for High-Affinity Ligand Engineering. *Chem Biol*. 2015;22(7):946-956. doi:10.1016/j.chembiol.2015.06.012.
152. Fowler DM, Araya CL, Gerard W, Fields S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*. 2011;27(24):3430-3431. doi:10.1093/bioinformatics/btr577.
153. Kim T, Tyndel MS, Huang H, et al. MUSI: An integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic Acids Res*. 2012;40(6):e47. doi:10.1093/nar/gkr1294.
154. Bloom JD. Software for the analysis and visualization of deep mutational scanning

data. *BMC Bioinformatics*. 2015;16(1):168. doi:10.1186/s12859-015-0590-4.
155. Matochko WL, Chu K, Jin B, Lee SW, Whitesides GM, Derda R. Deep sequencing analysis of phage libraries using Illumina platform. *Methods*. 2012;58(1):47-55. doi:10.1016/j.ymeth.2012.07.006.
156. Jolma A, Kivioja T, Toivonen J, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. 2010;20:861-873. doi:10.1101/gr.100552.109.
157. Alam KK, Chang JL, Burke DH. FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Mol Ther Acids*. 2015;4(August 2014):e230. doi:10.1038/mtna.2015.4.
158. Ravn U, Didelot G, Venet S, et al. Deep sequencing of phage display libraries to support antibody discovery. *Methods*. 2013;60(1):99-110. doi:10.1016/j.ymeth.2013.03.001.
159. Dickson RJ, Gloor GB. Bioinformatics Identification of Coevolving Residues. In: *Methods in Molecular Biology (Clifton, N.J.) Molecular Biology*. ; 2014:223-243. doi:10.1007/978-1-62703-968-0_15.
160. Feldwisch J, Tolmachev V, Lendel C, et al. Design of an optimized scaffold for affibody molecules. *J Mol Biol*. 2010;398(2):232-247. doi:10.1016/j.jmb.2010.03.002.
161. Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A. Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J Mol Biol*. 2003;332(3):489-503. doi:10.1016/S0022-2836(03)00896-9.
162. Mahon CM, Lambert M a, Glanville J, et al. Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. *J Mol Biol*. 2013;425(10):1712-1730. doi:10.1016/j.jmb.2013.02.015.
163. Lamminmäki U, Paupério S, Westerlund-Karlsson a, et al. Expanding the conformational diversity by random insertions to CDRH2 results in improved anti-estradiol antibodies. *J Mol Biol*. 1999;291(3):589-602. doi:10.1006/jmbi.1999.2981.
164. Parmley SF, Smith GP. Antibody-selectable filamentous fd phage vectors: affinity purification of target genes. *Gene*. 1988;73:305-318. doi:10.1016/0378-1119(88)90495-7.
165. Seelig B. mRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries. *Nat Protoc*. 2011;6(4):540-552. doi:10.1038/nprot.2011.312.
166. Mattheakis LC, Bhatt RR, Dower WJ. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc Natl Acad Sci U S A*. 1994;91(September):9022-9026. doi:10.1073/pnas.91.19.9022.
167. Marks JD, Ouwehand WH, Bye JM, et al. Human antibody fragments specific for human blood group antigens from a phage display library. *Bio/Technology*. 1993;11(10):1145-1149.
168. Quail M, Smith ME, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*. 2012;13(1):1. doi:10.1186/1471-2164-13-341.

169. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*. 2011;108(49):E1293-301. doi:10.1073/pnas.1111471108.
170. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. *Protein Sci*. 1992;1(3):409-417. doi:10.1002/pro.5560010313.
171. Hinkley T, Martins J, Chappey C, et al. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet*. 2011;43(5):487-489. doi:10.1038/ng.795.
172. Gong LI, Bloom JD. Epistatically Interacting Substitutions Are Enriched during Adaptive Protein Evolution. *PLoS Genet*. 2014;10(5). doi:10.1371/journal.pgen.1004328.
173. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet*. 2013;14(4):249-261. doi:10.1038/nrg3414.
174. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*. 2005;21(22):4116-4124. doi:10.1093/bioinformatics/bti671.
175. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins Struct Funct Genet*. 2004;56(2):211-221. doi:10.1002/prot.20098.
176. Brown CA, Brown KS. Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PLoS One*. 2010;5(6). doi:10.1371/journal.pone.0010779.
177. Durani V, Magliery TJ. *Protein Engineering and Stabilization from Sequence Statistics: Variation and Covariation Analysis.* Vol 523. 1st ed. Elsevier Inc.; 2013. doi:10.1016/B978-0-12-394292-0.00011-4.
178. Little DY, Chen L. Identification of Coevolving Residues and Coevolution Potentials Emphasizing Structure, Bond Formation and Catalytic Coordination in Protein Evolution. Shiu S-H, ed. *PLoS One*. 2009;4(3):e4762. doi:10.1371/journal.pone.0004762.
179. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*. 2008;24(3):333-340. doi:10.1093/bioinformatics/btm604.
180. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*. 2009;25(9):1125-1131. doi:10.1093/bioinformatics/btp135.
181. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol*. 2000;17(1):164-178. http://www.ncbi.nlm.nih.gov/pubmed/10666716.
182. Woldring DR, Holec P V, Hackel BJ. Hackel Lab GitHub. https://github.com/HackelLab-UMN. Published 2015.
183. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G. BLOSUM62 miscalculations improve search performance. *Nat Biotechnol*. 2008;26(3):274-275. doi:10.1038/nbt0308-274.
184. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by

analysis of correlated mutations. *Proteins Struct Funct Genet*. 2002;48(4):611-617. doi:10.1002/prot.10180.
185. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*. 2014;30(5):614-620. doi:10.1093/bioinformatics/btt593.
186. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27(21):2957-2963. doi:10.1093/bioinformatics/btr507.
187. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170.
188. Cole JR, Wang Q, Fish JA, et al. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res*. 2014;42(D1):1-10. doi:10.1093/nar/gkt1244.
189. Banta S, Dooley K, Shur O. Replacing antibodies: engineering new binding proteins. *Annu Rev Biomed Eng*. 2013;15:93-113.
190. Beerli RR, Rader C. Mining human antibody repertoires. *MAbs*. 2010;2(4).
191. Alder MN. Diversity and Function of Adaptive Immune Receptors in a Jawless Vertebrate. *Science (80-)*. 2005;310(5756):1970-1973. doi:10.1126/science.1119420.
192. Binz HK, Amstutz P, Plückthun A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol*. 2005;23(10):1257-1268.
193. Nord K, Gunneriusson E, Ringdahl J, Ståhl S, Uhlén M, Nygren P-Å. Binding proteins selected from combinatorial libraries of an α-helical bacterial receptor domain. *Nat Biotechnol*. 1996;15(8):772-777. doi:10.1038/nbt0897-772.
194. Löfblom J, Feldwisch J, Tolmachev V, Carlsson J, Ståhl S, Frejd FY. Affibody molecules: engineered proteins for therapeutic, diagnostic and biotechnological applications. *FEBS Lett*. 2010;584(12):2670-2680.
195. Lendel C, Dincbas-Renqvist V, Flores A, et al. Biophysical characterization of Z SPA-1 -A phage-display selected binder to protein A. *Protein Sci*. 2004;13(8):2078-2088. doi:10.1110/ps.04728604.
196. Feldwisch J, Tolmachev V, Lendel C, et al. Design of an optimized scaffold for affibody molecules. *J Mol Biol*. 2010;398:232-247. doi:10.1016/j.jmb.2010.03.002.
197. Lundberg E, Brismar H, Gräslund T. Selection and characterization of Affibody ligands to the transcription factor c-Jun. *Biotechnol Appl Biochem*. 2009;52(Pt 1):17-27. doi:10.1042/BA20070178.
198. Nygren PÅ. Alternative binding proteins: Affibody binding proteins developed from a small three-helix bundle scaffold. *FEBS J*. 2008;275(11):2668-2676. doi:10.1111/j.1742-4658.2008.06438.x.
199. Lindborg M, Cortez E, Höidén-Guthenberg I, et al. Engineered high-affinity affibody molecules targeting platelet-derived growth factor receptor β in vivo. *J Mol Biol*. 2011;407(2):298-315.
200. Kronqvist N, Malm M, Göstring L, et al. Combining phage and staphylococcal surface display for generation of ErbB3-specific Affibody molecules. *Protein Eng Des Sel*. 2011;24(4):385-396.

201. Gunneriusson E, Nord K, Uhlén M, Nygren P. Affinity maturation of a Taq DNA polymerase specific affibody by helix shuffling. *Protein Eng*. 1999;12(10):873-878.
202. Eklund M, Axelsson L, Uhlén M, Nygren P-Å. Anti-idiotypic protein domains selected from protein A-based affibody libraries. *Proteins*. 2002;48(3):454-462.
203. Grimm S, Lundberg E, Yu F, et al. Selection and characterisation of affibody molecules inhibiting the interaction between Ras and Raf in vitro. *N Biotechnol*. 2010;27(6):766-773. doi:10.1016/j.nbt.2010.07.016.
204. Grimm S, Salahshour S, Nygren P-Å. Monitored whole gene in vitro evolution of an anti-hRaf-1 affibody molecule towards increased binding affinity. *N Biotechnol*. 2012;29(5):534-542.
205. Wernérus H, Samuelson P, Ståhl S. Fluorescence-activated cell sorting of specific affibody-displaying staphylococci. *Appl Environ Microbiol*. 2003;69(9):5328-5335.
206. Kronqvist N, Löfblom J, Jonsson A, Wernérus H, Ståhl S. A novel affinity protein selection system based on staphylococcal cell surface display and flow cytometry. *Protein Eng Des Sel*. 2008;21(4):247-255.
207. Lindberg H, Johansson A, Härd T, Ståhl S, Löfblom J. Staphylococcal display for combinatorial protein engineering of a head-to-tail affibody dimer binding the Alzheimer amyloid-β peptide. *Biotechnol J*. 2013;8(1):139-145.
208. Case BA, Hackel BJ. Synthetic and natural consensus design for engineering charge within an affibody targeting epidermal growth factor receptor. *Biotechnol Bioeng*. 2016:epub ahead of print.
209. Chen J, Sawyer N, Regan L. Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci*. 2013;22(4):510-515. doi:10.1002/pro.2230.
210. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A*. 2006;103(15):5869-5874. doi:10.1073/pnas.0510098103.
211. Tokuriki N, Stricher F, Serrano L, Tawfik DS. How protein stability and new functions trade off. *PLoS Comput Biol*. 2008;4(2):35-37. doi:10.1371/journal.pcbi.1000002.
212. Nagatani RA, Gonzalez A, Shoichet BK, Brinen LS, Babbitt PC. Stability for Function Trade-Offs in the Enolase Superfamily "Catalytic Module" [†, ‡]. *Biochemistry*. 2007;46(23):6688-6695. doi:10.1021/bi700507d.
213. Mukaiyama A, Haruki M, Ota M, Koga Y, Takano K, Kanaya S. A hyperthermophilic protein acquires function at the cost of stability. *Biochemistry*. 2006;45(42):12673-12679. doi:10.1021/bi060907v.
214. Koide A, Wojcik J, Gilbreth RN, Hoey RJ, Koide S. Teaching an Old Scaffold New Tricks: Monobodies Constructed Using Alternative Surfaces of the FN3 Scaffold. *J Mol Biol*. 2012;415(2):393-405.
215. Diem MD, Hyun L, Yi F, et al. Selection of high-affinity Centyrin FN3 domains from a simple library diversified at a combination of strand and loop positions. *Protein Eng Des Sel*. 2014;27(10):419-429. doi:10.1093/protein/gzu016.
216. Araya CL, Fowler DM. Deep mutational scanning: Assessing protein function on a massive scale. *Trends Biotechnol*. 2011;29(9):435-442.

doi:10.1016/j.tibtech.2011.04.003.
217. Fowler DM, Stephany JJ, Fields S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat Protoc*. 2014;9(9):2267-2284. doi:10.1038/nprot.2014.153.
218. Wrenbeck EE, Faber MS, Whitehead TA. Deep sequencing methods for protein engineering and design. *Curr Opin Struct Biol*. 2017;45:36-44. doi:10.1016/j.sbi.2016.11.001.
219. Rockah-Shmuel L, Tóth-Petróczy Á, Tawfik DS. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Comput Biol*. 2015;11(8):1-28. doi:10.1371/journal.pcbi.1004421.
220. Au L, Green DF. Direct Calculation of Protein Fitness Landscapes through Computational Protein Design. *Biophys J*. 2016;110(1):75-84. doi:10.1016/j.bpj.2015.11.029.
221. Brender JR, Zhang Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS Comput Biol*. 2015;11(10):1-25. doi:10.1371/journal.pcbi.1004494.
222. Xiong P, Zhang C, Zheng W, Zhang Y. BindProfX: Assessing mutation-induced binding affinity change by protein interface profiles with pseudo counts. *J Mol Biol*. 2016. doi:10.1016/j.jmb.2016.11.022.
223. Fellouse FA, Barthelemy PA, Kelley RF, Sidhu SS. Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *J Mol Biol*. 2006;357(1):100-114.
224. Engfeldt T, Renberg B, Brumer H, Nygren P-Å, Karlström AE. Chemical synthesis of triple-labelled three-helix bundle binding proteins for specific fluorescent detection of unlabelled protein. *Chembiochem*. 2005;6(6):1043-1050.
225. Friedman M, Nordberg E, Höidén-Guthenberg I, et al. Phage display selection of Affibody molecules with specific binding to the extracellular domain of the epidermal growth factor receptor. *Protein Eng Des Sel*. 2007;20(4):189-199.
226. Friedman M, Orlova A, Johansson E, et al. Directed evolution to low nanomolar affinity of a tumor-targeting epidermal growth factor receptor-binding affibody molecule. *J Mol Biol*. 2008;376(5):1388-1402.
227. Grimm S, Yu F, Nygren P-Å. Ribosome display selection of a murine IgG1 Fab binding affibody molecule allowing species selective recovery of monoclonal antibodies. *Mol Biotechnol*. 2011;48(3):263-276. doi:10.1007/s12033-010-9367-1.
228. Grimm S. Ribosome display for selection and evolution of affibody molecules. 2011.
229. Grönwall C, Jonsson A, Lindström S, Gunneriusson E, Ståhl S, Herne N. Selection and characterization of Affibody ligands binding to Alzheimer amyloid β peptides. *J Biotechnol*. 2007;128(1):162-183. doi:10.1016/j.jbiotec.2006.09.013.
230. Grönwall C, Snelders E, Palm AJ, Eriksson F, Herne N, Ståhl S. Generation of Affibody® ligands binding interleukin-2 receptor α/CD25. *Biotechnol Appl Biochem*. 2008;50(2):97. doi:10.1042/BA20070261.
231. Hansson M, Ringdahl J, Robert a, et al. An in vitro selected binding protein (affibody) shows conformation-dependent recognition of the respiratory syncytial virus (RSV) G protein. *Immunotechnology*. 1999;4(3-4):237-252. doi:#.

232. Högbom M, Eklund M, Nygren P-Å, Nordlund P. Structural basis for recognition by an in vitro evolved affibody. *Proc Natl Acad Sci U S A*. 2003;100(6):3191-3196.
233. Jonsson A. Development of molecular recognition by rational and combinatorial engineering. *PhD Thesis*. 2009.
234. Jonsson A, Wållberg H, Herne N, Ståhl S, Frejd FY. Generation of tumour-necrosis-factor-alpha-specific affibody molecules capable of blocking receptor binding in vitro. *Biotechnol Appl Biochem*. 2009;54(2):93-103. doi:10.1042/BA20090085.
235. Lendel C. Molecular principles of protein stability and protein-protein interactions. 2005.
236. Lendel C, Dogan J, Härd T. Structural Basis for Molecular Recognition in an Affibody:Affibody Complex. *J Mol Biol*. 2006;359(5):1293-1304. doi:10.1016/j.jmb.2006.04.043.
237. Li J, Lundberg E, Vernet E, Larsson B, Hoiden-Guthenberg I, Graslund T. Selection of affibody molecules to the ligand-binding site of the insulin-like growth factor-1 receptor. *Biotechnol Appl Biochem*. 2010;55(2):99-109. doi:10.1042/BA20090226.
238. Löfdahl P, Nygren P. Affinity maturation of a TNFα-binding Affibody molecule by Darwinian survival selection. *Biotechnol Appl Biochem*. 2010;55(3):111-120. doi:10.1042/BA20090274.
239. Löfdahl P-Å, Nord O, Janzon L, Nygren P-Å. Selection of TNF-α binding affibody molecules using a β-lactamase protein fragment complementation assay. *N Biotechnol*. 2009;26(5):251-259. doi:10.1016/j.nbt.2009.06.980.
240. Nord K, Nord O, Uhlén M, Kelley B, Ljungqvist C, Nygren PA. Recombinant human factor VIII-specific affinity ligands selected from phage-displayed combinatorial libraries of protein A. *Eur J Biochem*. 2001;268(15):4269-4277. doi:10.1046/j.1432-1327.2001.02344.x.
241. Orlova A, Magnusson M, Eriksson TLJ, et al. Tumor imaging using a picomolar affinity HER2 binding Affibody molecule. *Cancer Res*. 2006;66(8):4339-4348. doi:10.1158/0008-5472.CAN-05-3521.
242. Rönnmark J, Grönlund H, Uhlén M, Nygren P-Å. Human immunoglobulin A (IgA)-specific ligands from combinatorial engineering of protein A. *Eur J Biochem*. 2002;269(11):2647-2655.
243. Sandström K, Xu Z, Forsberg G, Nygren P. Inhibition of the CD28-CD80 co-stimulation signal by a CD28-binding affibody ligand developed by combinatorial protein engineering. *Protein Eng*. 2003;16(9):691-697. doi:10.1093/protein/gzg086.
244. Wahlberg E, Lendel C, Helgstrand M, et al. An affibody in complex with a target protein: structure and coupled folding. *Proc Natl Acad Sci U S A*. 2003;100(6):3185-3190. doi:10.1073/pnas.0436086100.
245. Wållberg H, Löfdahl P-Å, Tschapalda K, et al. Affinity recovery of eight HER2-binding affibody variants using an anti-idiotypic affibody molecule as capture ligand. *Protein Expr Purif*. 2011;76(1):127-135. doi:10.1016/j.pep.2010.10.008.
246. Wikman M, Steffen A-C, Gunneriusson E, et al. Selection and characterization of HER2/neu-binding affibody ligands. *Protein Eng Des Sel*. 2004;17(5):455-462.

247. Wikman M, Rowcliffe E, Friedman M, et al. Selection and characterization of an HIV-1 gp120-binding affibody ligand. *Biotechnol Appl Biochem*. 2006;45(2):93. doi:10.1042/BA20060016.
248. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res*. 2005;33(Web Server):W382-W388.
249. Myers JK, Pace CN, Scholtz JM. A direct comparison of helix propensity in proteins and peptides. *Proc Natl Acad Sci U S A*. 1997;94(7):2833-2837. doi:10.1073/pnas.94.7.2833.
250. Blaber M, Zhang X-J, Matthews BW. Structural basis of amino acid α-helix propensity. *Science (80-)*. 1993;260(5114):1637-1640. doi:10.1126/science.8503008.
251. Horovitz A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold Des*. 1996;1(6):R121-R126. doi:10.1016/S1359-0278(96)00056-9.
252. Lyu P, Liff M, Marky L, Kallenbach N. Side chain contributions to the stability of alpha-helical structure in peptides. *Science (80-)*. 1990;250(4981):669-673. doi:10.1126/science.2237416.
253. Williams RW, Chang A, Juretić D, Loughran S. Secondary structure predictions and medium range interactions. *Biochim Biophys Acta - Protein Struct Mol Enzymol*. 1987;916(2):200-204. doi:10.1016/0167-4838(87)90109-9.
254. O'Neil KT, DeGrado WF. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*. 1990;250(4981):646-651. doi:10.1126/science.2237415.
255. Benatuil L, Perez JM, Belk J, Hsieh C-M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng Des Sel*. 2010;23(4):155-159.
256. Ackerman M, Levary D, Tobon G, Hackel B, Orcutt KD, Wittrup KD. Highly avid magnetic bead capture: An efficient selection method for de novo protein engineering utilizing yeast surface display. *Biotechnol Prog*. 2009;25(3):774-783. doi:10.1002/btpr.174.
257. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD. Isolating and engineering human antibodies using yeast surface display. *Nat Protoc*. 2006;1(2):755-768.
258. Hackel BJ, Kapila A, Wittrup KD. Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. *J Mol Biol*. 2008;381(5):1238-1252.
259. Woldring DR, Holec P V, Hackel BJ. ScaffoldSeq: Software for characterization of directed evolution populations. *Proteins*. 2016;84(7):869-874.
260. Eigenbrot C, Ultsch M, Dubnovitsky A, Abrahmsén L, Härd T. Structural basis for high-affinity HER2 receptor binding by an engineered protein. *Proc Natl Acad Sci U S A*. 2010;107(34):15039-15044. doi:10.1073/pnas.1005025107.
261. Searle MS, Williams DH. The cost of conformational order: entropy changes in molecular associations. *J Am Chem Soc*. 1992;114(27):10690-10697. doi:10.1021/ja00053a002.
262. Cole C, Warwicker J. Side-chain conformational entropy at protein-protein

interfaces. *Protein Sci*. 2002;11:2860-2870. doi:10.1110/ps.0222702.or.
263. Bloom JD, Wilke CO, Arnold FH, Adami C. Stability and the evolvability of function in a model protein. *Biophys J*. 2004;86(5):2758-2764. doi:10.1016/S0006-3495(04)74329-5.
264. Hackel BJ, Neil JR, White FM, Wittrup KD. Epidermal growth factor receptor downregulation by small heterodimeric binding proteins. *Protein Eng Des Sel*. 2012;25(2):47-57. doi:10.1093/protein/gzr056.
265. Chao G, Cochran JR, Dane Wittrup K. Fine epitope mapping of anti-epidermal growth factor receptor antibodies through random mutagenesis and yeast surface display. *J Mol Biol*. 2004;342(2):539-550. doi:10.1016/j.jmb.2004.07.053.
266. Rosenfeld L, Shirian J, Zur Y, Levaot N, Shifman JM, Papo N. Combinatorial and computational approaches to identify interactions of macrophage colony-stimulating factor (M-CSF) and its receptor c-FMS. *J Biol Chem*. 2015;290(43):26180-26193. doi:10.1074/jbc.M115.671271.
267. Boersma YL, Chao G, Steiner D, Wittrup KD, Plučkthun A. Bispecific Designed Ankyrin Repeat Proteins (DARPins) targeting epidermal growth factor receptor inhibit A431 cell proliferation and receptor recycling. *J Biol Chem*. 2011;286(48):41273-41285. doi:10.1074/jbc.M111.293266.
268. Martin SF, Tatham MH, Hay RT, Samuel IDW. Quantitative analysis of multi-protein interactions using FRET: application to the SUMO pathway. *Protein Sci*. 2008;17(4):777-784. doi:10.1110/ps.073369608.
269. Sridharan R, Zuber J, Connelly SM, Mathew E, Dumont ME. Fluorescent approaches for understanding interactions of ligands with G protein coupled receptors. *Biochim Biophys Acta - Biomembr*. 2014;1838(1 PARTA):15-33. doi:10.1016/j.bbamem.2013.09.005.
270. Taraska JW, Puljung MC, Olivier NB, Flynn GE, Zagotta WN. Mapping the structure and conformational movements of proteins with transition metal ion FRET. *Nat Methods*. 2009;6(7):532-537. doi:10.1038/nmeth.1341.
271. Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci*. 2000;97(7):3288-3291. doi:10.1073/pnas.97.7.3288.
272. Stein RR, Marks DS, Sander C. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Comput Biol*. 2015;11(7):1-22. doi:10.1371/journal.pcbi.1004182.
273. Wu NC, Olson CA, Sun R. High-throughput identification of protein mutant stability computed from a double mutant fitness landscape. *Protein Sci*. 2015;0:n/a-n/a. doi:10.1002/pro.2840.
274. Goldsmith M, Tawfik DS. *Enzyme Engineering by Targeted Libraries.* Vol 523. 1st ed. Elsevier Inc.; 2013. doi:10.1016/B978-0-12-394292-0.00012-6.
275. Seelig B, Szostak JW. Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature*. 2007;448(7155):828-831. doi:10.1038/nature06032.

Appendix

Quantification of MET expression in pseudo-metastatic tumors

High expression of hepatocyte growth factor receptor (MET) is associated with certain invasive cancer types. Developing small protein ligands that specifically bind to hepatocyte growth factor receptor (MET) with high affinity will provide opportunities for detecting cancerous tumors and metastasis. We aim to establish a model system for characterizing engineered proteins that bind to MET in the context of pseudo-metastatic tumors localized in the lung. Several cell lines were analyzed for their potential utility as model systems of high MET expression (Figure 1). For future comparative studies, identifying a cell line with low expression of MET was also necessary.

Pseudo-metastatic tumors were generated via mice tail vein injection of each cell type and subsequently characterized for MET expression. Cell line characterization consisted of excising lungs that contained tumor growth, preparing a single cell suspension of the tumor cells, and quantifying the level of MET expression per cell using flow cytometry. Thus far, we have determined the expression level for multiple cell lines ranging from 1.6 – 50k MET per cell in the context of pseudo-metastatic tumors (Figure 2).

In addition, MET expression has been assessed in the context of multiple growth environments. Comparing the extent that MET is expressed by metastasized cells, xenograft tumors or cells having been cultured in petri dishes alone will elucidate the impact of cellular micro-environment on surface protein content (Figure 1). This work would benefit from further investigation of possible relationships between MET expression and the prevalence of tumor vasculature, extracellular matrix, lymphocytes, and immune cells.

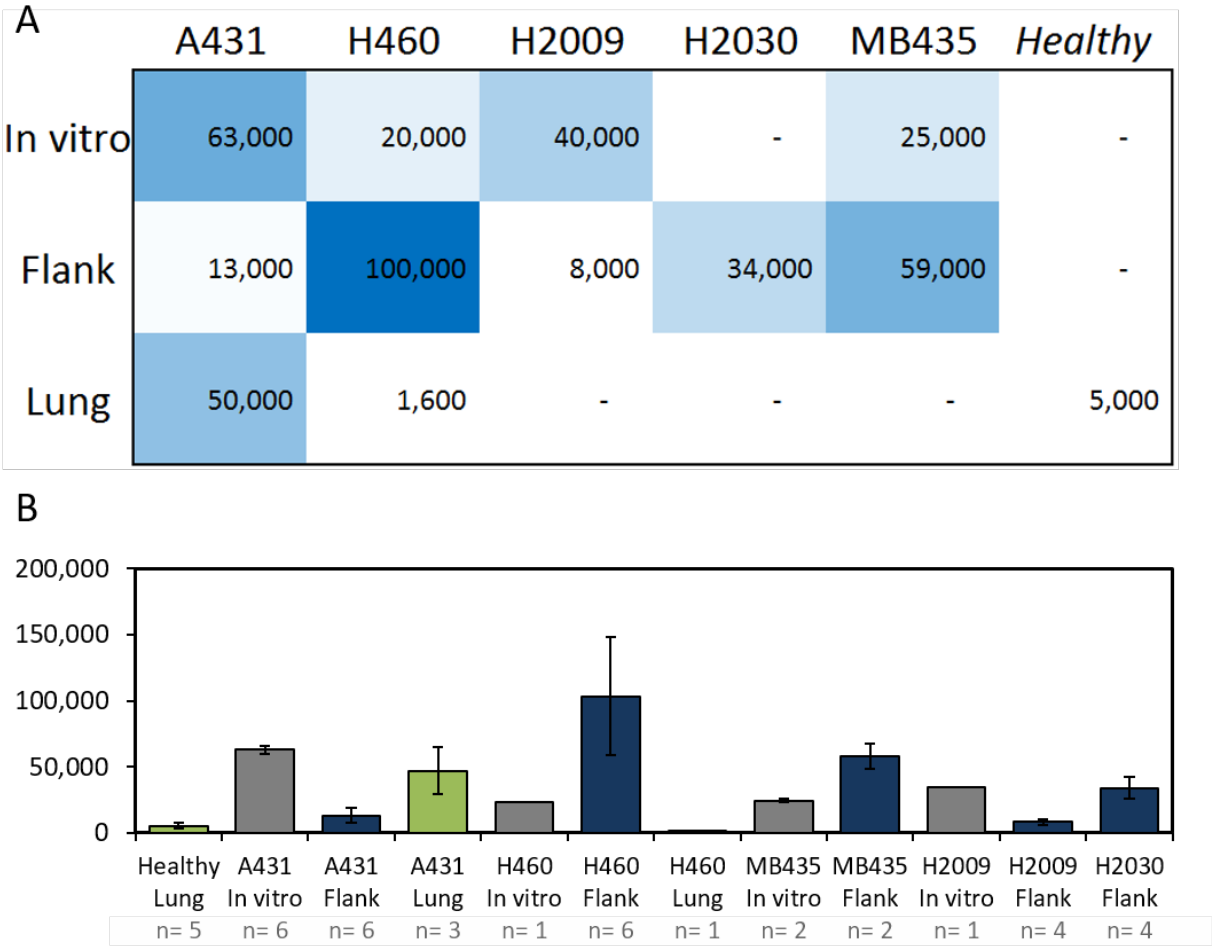


Figure 1. ***Summary of MET expression in multiple cell lines and growth environments***. The median number of MET expressed per cell are shown for as a heatmap (A) and bar graph (B) with error bars indicating standard deviation. Healthy and tumorous lung samples are shown as green bars. Xenograft tumors and cells grown in a petri dish alone are shown in blue and gray bars, respectively.

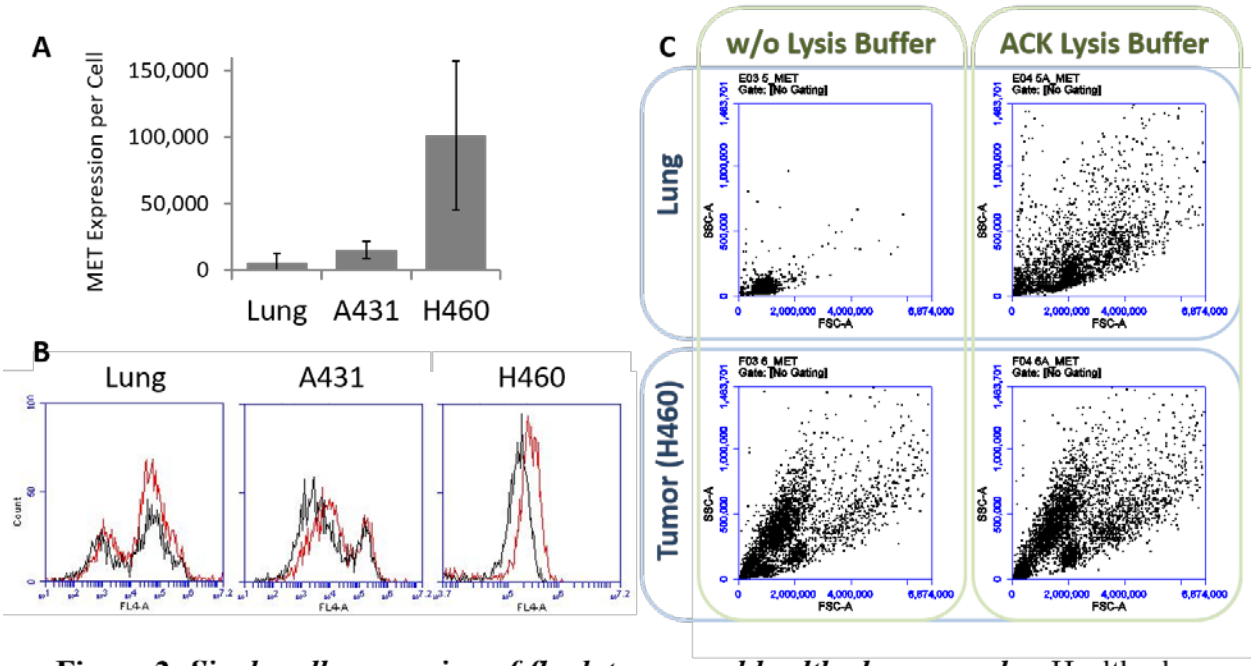


Figure 2. ***Single cell suspension of flank tumor and healthy lung samples.*** Healthy lung and xenograft flank A431 and H460 tumors (n=5,6,6, respectively) were assessed for MET expression via cytometry. Mean expression levels are shown in panel **A**. Panel **B** shows fluorescence corresponding to anti-MET antibody. Samples not receiving anti-MET Ab incubation are in black; anti-MET Ab labeled samples are in red. During the preparation of the single cell suspensions, two treatment methods were done in parallel where ACK red blood cell lysis buffer was either included or excluded prior to final wash step. The lysis buffer had a significant impact on the fraction of very small cells, especially in the lung samples, likely corresponding to red blood cells. One representative healthy lung and flank tumor sample is shown in panel **C** with and without the ACK lysis buffer.

Cell isolation and preparation

Protocol adapted from Blake Jacobson and Jeremy Drees

Tumor cell isolation

1. Remove tumor sample from receptacle and place into dry 60mm petri dish.
2. Mince tumor using clean razor blade until ~1mm pieces.
    a. Note: only 1g of tumor or less should be homogenized at this point.
3. Use razor blade to "scoop" minced tissue and add to labeled gentlemacs C tube
4. Add 2.5mL of tumor digestion enzyme mix to gentlemacs C tube
   Note: Use this 2.5 mL to rinse the razor blade above the C-tube to leave as little sample behind as possible Note: FBS inactivates the enzymes in the solution, this is why if possible the tissues should be collected in FBS-free medium.
5. Add Gentlemacs tubes containing sample to a GentleMacs Tissue dissociator, run program m_impTumor_02.
6. Incubate with gentle mixing for 45 minutes at 37^{o}C
    a. *If possible, use the GentleMacs rotator mixer; however only 4 C-tubes will fit on in this apparatus. If more than 4 samples need incubation, a 50mL conical vial rack works just as well: make sure the tissue is not "stuck" in the tube blades, face all of the tubes in the same orientation in the rack, tape the tubes in place in the rack, lay the rack on its side in an incubator shaker, and shake on a slow rotation for the incubation period.*
    b. During this incubation time, move on to prepping other tissues in steps 2-4 (if applicable).
7. Following incubation, add gentlemacs tubes again to tissue dissociator and run program Mouse m_impTumor_03. Add tubes to ice following this program.
8. Place nylon strainer onto labeled 50ml conical vial and pour homogenized sample into strainer. Add 5mL of ice cold RMPI+10%FBS, to gentlemacs tube, recap the tube and mix the medium around to wash remaining tissue homogenate from the C-tube. Pour this wash through the strainer as well.
   Note: At this point the homogenate and wash will most likely be close to overflowing in the strainer, especially if larger tumors were homogenized. To help the cell homogenate through the strainer, slowly lift the strainer out of the tube by the lip. This will usually force the homogenate through. Replacing and lifting the strainer multiple times will usually work with particularly large or troublesome tumor samples.
9. (Optional) Transfer the homogenate to a 15ml conical vial.
   Note: Usually enough samples are being processed where it is more practical to do subsequent processing in 15mL tubes instead of the 50ml conical vial, at the expense of more plasticware.
10. Spin the samples at 4^{o}C at 300xg for 5 minutes.
11. Aspirate supernatant.
    a. *Be sure to get as much supernatant as possible, FBS will prevent lysis of RBS in the next step.*
12. Resuspend in 2mL of **room temp (RT)** ACK lysis buffer. Incubate at RT for 2 minutes.
    Note: Red blood cells can interfere with flow analysis. See figure below. Lysing the red blood cells is sometimes not necessary for more poorly vascularized tumors, but it cleans up the sample well. However, it is recommended for every flank or lung tumor sample.

13. Add 13mL of ice cold RPMI+10% FBS to cells.
   a. *This inactivates the lysis buffer and prevents the lysis of leukocytes.*
   b. *At this point, run the samples through a new cell strainer held above a 15ml conical vial to remove any more debris that may be in the cell suspension. If no debris is seen floating in the suspension, it is not necessary. To do this, hold the strainer tightly against the 15ml conical vial centered below it, slowly pour the liquid into the center of the strainer. Sample will be lost if this is done too quickly or if it is not poured directly above the 15ml conical vial.*

14. Spin again at 4ºC at 300xg for 5 minutes.

15. Resuspend in 3mL of RPMI+FBS. Cells are now ready to be stained. Keep vials on ice until other organs are prepped (if any).
Note: Cells should only be kept on ice in suspension for a maximum of 4 hours. If cells are going to be on ice for a while (i.e. many other organ samples still need to be prepped before staining at this point), lay vials horizontally on ice to prevent cell pelleting. Gently swirl periodically to keep cells in suspension as much as possible to maintain viability).

Quantification of Cell Expression via Beads

Primary labeling

Beads

1. In new 1.7mL tube, add 2.5 µL of each bead → Bead labels: **B, 1, 2, 3, 4**
2. Add 1 µL 9E10 to the beads
3. Add 36.5 µL PBSA (total V = 50 uL)

Mammalian Cells

4. Cells: Add ~100,000 cells to each tube (Count using hemocytometer)
5. Label cells at 0 µg/mL (negative control) and 10 µg/mL Target Ab
   a. Anti-hHGF R/cMET clone 95106 (~150kDa) → 10ug/mL in 50 µL
6. Incubate Primary mixes for ~20 min, 4 C; wash with PBSA

Secondary labeling

7. Label beads and cells with 1:1000 dilution of Goat anti-Mouse-647
8. Incubate 5-10 min, 4 C; wash with PBSA
9. Analyze via cytometry (Accuri) – compare histogram peaks of beads and sample.