ACS Combinatorial Science

Research Article

Cite This: ACS Comb. Sci. 2019, 21, 323−335

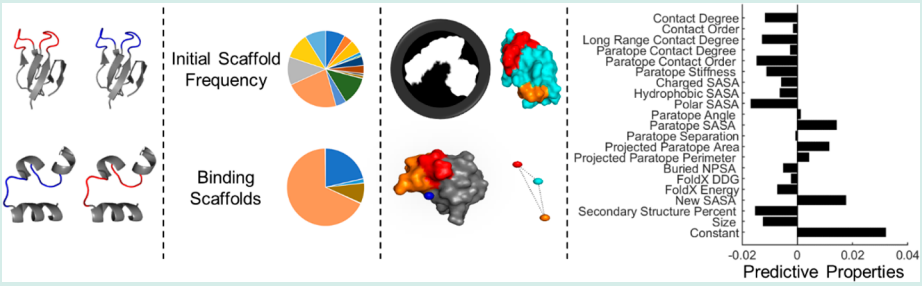
pubs.acs.org/acscombsci

Biophysical Characterization Platform Informs Protein Scaffold Evolvability

Alexander W. Golinski, Patrick V. Holec, Katelynn M. Mischler, and Benjamin J. Hackel*

Department of Chemical Engineering and Materials Science, University of Minnesota−Twin Cities, 421 Washington Avenue Southeast, 356 Amundson Hall, Minneapolis, Minnesota 55455, United States

Supporting Information



ABSTRACT: Evolving specific molecular recognition function of proteins requires strategic navigation of a complex mutational landscape. Protein scaffolds aid evolution via a conserved platform on which a modular paratope can be evolved to alter binding specificity. Although numerous protein scaffolds have been discovered, the underlying properties that permit binding evolution remain unknown. We present an algorithm to predict a protein scaffold's ability to evolve novel binding function based upon computationally calculated biophysical parameters. The ability of 17 small proteins to evolve binding functionality across seven discovery campaigns was determined via magnetic activated cell sorting of 10^{10} yeast-displayed protein variants. Twenty topological and biophysical properties were calculated for 787 small protein scaffolds and reduced into independent components. Regularization deduced which extracted features best predicted binding functionality, providing a 4/6 true positive rate, a 9/11 negative predictive value, and a 4/6 positive predictive value. Model analysis suggests a large, disconnected paratope will permit evolved binding function. Previous protein engineering endeavors have suggested that starting with a highly developable (high producibility, stability, solubility) protein will offer greater mutational tolerance. Our results support this connection between developability and evolvability by demonstrating a relationship between protein production in the soluble fraction of *Escherichia coli* and the ability to evolve binding function upon mutation. We further explain the necessity for initial developability by observing a decrease in proteolytic stability of protein mutants that possess binding functionality over nonfunctional mutants. Future iterations of protein scaffold discovery and evolution will benefit from a combination of computational prediction and knowledge of initial developability properties.

KEYWORDS: *protein scaffolds, predictive algorithm, protein evolvability*

INTRODUCTION

Proteins have evolved to empower a broad array of functionality. While minimal amino acid mutations can yield dramatic enhancements in functional performance via evolution,[1,2] discovery of completely new function typically requires greater leaps in sequence.[3] Given the relative barrenness and tortuosity of sequence space,[2] efficient strategies are needed to achieve successful de novo discovery. One strategy to facilitate discovery is the use of a protein scaffold[4,5] comprising a conserved framework to provide biophysical robustness and a variable active site to provide diverse function. One particular function, molecular recognition via binding ligands, has ubiquity in natural biology and broad technological utility in targeted molecular therapies[6] and diagnostics.[7] A functional protein ligand scaffold must be able to evolve new, specific binding function upon mutation of the paratope[8] and possess optimal developability properties (e.g., stability, solubility, and expression) for downstream use.[9] To date, numerous protein scaffolds have been engineered to obtain strong affinity toward clinically relevant targets,[10,11] while some have entered clinical trials.[12−15] Protein scaffolds offer novel topologies and differential size, allowing for unique binding interfaces and tunable pharmacokinetic properties.[16,17] The diversity of topologies and physicochemistries of published scaffolds and the paucity of data on unsuccessful scaffolds preclude an understanding of the biophysical features that allow the development of binding functionality. Thus, to advance the understanding of de novo protein discovery and evolution, as well as to advance technological capability for ligand engineering, we sought to develop a platform to

Received: November 30, 2018
Revised: January 19, 2019
Published: January 25, 2019

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

elucidate the factors that dictate scaffold performance and to identify new scaffolds.

Previously established scaffolds have been discovered based on an evolutionary or mechanically themed hypothesis. The use of antibodies,[6] antibody fragments,[18] and leucine-rich repeats[19] presumed that their natural function for high affinity binding will serve as a starting point for scaffold engineering. Fibronectin type III "monobodies"[20] and designed ankyrin repeat proteins[21] are structurally similar to these immune scaffolds. Lipocalins,[22] three-helix bundle affibodies,[23] fynomers,[24] and others[17] offer unique topologies with native binding ability. Alternatively, multiple scaffolds are chosen for their strong structural stability, including cystine knots[25] and thermophilic affitins[26] and homologues. Similarly, a host of other scaffolds have provided compelling performance in ligand development, while others have been tested without the same level of success.[21] A comparison of potential scaffolds was recently performed, which identified the Gp2 scaffold for its small size, adjacent, solvent-exposed loops with significant surface area, and mutational tolerance.[10] However, a rigorous evaluation of the properties that permit protein scaffold function, now enabled by advances in high-throughput screening and sequencing, has yet to be performed.

Herein, we propose an iterative discovery and evaluation platform for new protein scaffolds in which we computationally characterize biophysical properties of scaffold topologies and experimentally evaluate binder evolution (Figure 1). Parameter

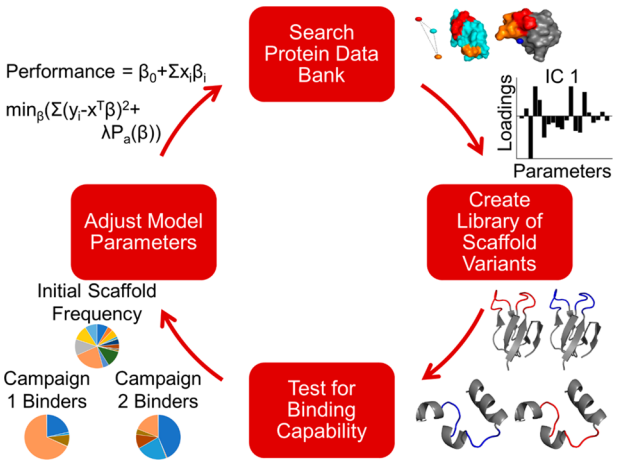


Figure 1. Algorithm for protein scaffold discovery. Small proteins deposited in the Protein Data Bank are analyzed for structural, chemical, and predicted stability parameters. Proteins for experimental evaluation are chosen via a proposed model to predict binding performance. Protein scaffold libraries consisting of millions of unique variants are expressed with diversified binding interfaces. Binding function is evaluated against several molecular targets to determine which proteins evolve specific binding variants. The observed binding performance is then used to adjust the predictive model. Iterative evaluation can be performed.

selection techniques are then employed to assess predictive characteristics of evolvable scaffolds. In this Research Article, computationally derived stability and topology parameters were used to identify the first predictive model of protein scaffold function, which can be used to identify future successful protein scaffold candidates. Additionally, experimental characterization of scaffold developability suggests stable and producible proteins yield improved binder evolution to combat a trade-off between stability and new binding function. The findings in the study suggest a combination of developability and biophysical metrics should be used to identify future protein scaffolds.

RESULTS AND DISCUSSION

Computational Scaffold Analysis. We hypothesize that not all proteins possess the characteristics to robustly and efficiently evolve novel binding function upon mutation. To advance the understanding of scaffold properties that dictate evolvability, and to reduce the experimental burden of identifying new scaffolds or improving existing scaffolds, we aim to advance a computational/experimental framework to evaluate binding evolvability of candidates. We hypothesize that a combination of topological and biophysical parameters can be used to provide insight on performance.

We focused the current study on small (<65 amino acids), single-domain proteins for multiple reasons. Small proteins provide improved physiological transport and rapid clearance of unbound molecules for enhanced selectivity.[27] Small, single-domain architecture eases fusion and site-specific conjugation for multifunctional constructs. The small size reduces exposed surface area that may lead to undesired nonspecific interactions. Moreover, small size heightens the challenge to simultaneously balance evolution of intramolecular stability and intermolecular binding,[28,29] which makes it a strong test case for evolution. Multiple types of protein structure can be used for diversification of a binding paratope including loops,[20,30] α-helices,[31,32] β-strands,[33] and mixed topologies.[21,22] Although the impact of entropic cost upon binding,[34,35] relative to more constrained paratope structures, remains difficult to accurately access, the conformational flexibility of loops suggests this secondary structure will be most accepting of mutagenesis.[36] Thus, we sought proteins with at least two enclosed loop regions each with at least four residues for diversification.

The >100 000 proteins in the Protein Data Bank (PDB) were (i) filtered for size (30−65 AA pretrimming) and the presence of two loops with at least four residues. 787 unique protein scaffolds were (ii) demarcated into conserved frameworks and diversifiable paratopes and (iii) characterized by 20 parameters describing geometrical, chemical, and stability properties (summarized in Table 1 and the following text and described in depth in Experimental Procedures). (1) *Protein Connectivity*. We hypothesized that the connectivity of residues would impact protein stability, leading to the calculation of inter-residue contact degree (total and long-range) and contact order.[37] (2) *Paratope Connectivity*. Paratope connectivity and flexibility, the latter via normal-mode analysis,[38] was also calculated as we believed spatially removed diversifications will be less destabilizing to the remainder of the protein. (3) *Conserved Surface Area Chemical Nature*. As for the conserved framework, the amount and chemical nature of exposed residues are likely to affect the ability of proteins to withstand destabilizing mutations. PyMOL[39] was used to model the protein surface and calculate the chemical nature of the solvent accessible surface area (SASA). (4) *Paratope Size and Topology*. Paratope orientation was parametrized by spatial and angular separation to capture the potential additivity of the two paratope loops. Paratope size and shape were described by measuring the properties of the 2D and 3D binding interface. (5) *Computational Stability*. It is proposed that scaffolds must be stable and have mutational stability to maintain structural

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

Table 1. Evaluated Descriptors of Protein Scaffolds

factor	description	mean ± SD (n = 787)
	protein connectivity	
contact degree	total number of residue contacts within 8 Å	920 ± 270 AU
contact order	sum of contact sequence separation divided by size and contact degree	0.38 ± 0.01 AU
long-range contact degree	number of residue contacts with sequence separation >12 divided by size	11.8 ± 3.1 AU
	paratope connectivity	
paratope contact degree	total number of residue contacts within 8 Å between a paratope and conserved residue	430 ± 140 AU
paratope contact order	sum of paratope contacts sequence separation divided by paratope size and contact degree	1.2 ± 0.4 AU
paratope stiffness	average stiffness of the paratope in an anisotropic network model	−0.28 ± 0.39 AU
	conserved surface area chemical nature	
charged SASA	conserved solvent accessible surface area of D, E, K, R	980 ± 430 Å^2
hydrophobic SASA	conserved solvent accessible surface area of A, F, G, I, L, M, P	790 ± 340 Å^2
polar SASA	conserved solvent accessible surface area of C, H, N, Q, S, T, W, Y	780 ± 360 Å^2
	paratope size and topology	
paratope angle	[paratope 1: entire scaffold: paratope 2] angle based upon centers of volume	110 ± 30°
paratope SASA	solvent-exposed surface area of an alanine-scanned paratope region	780 ± 360 Å^2
paratope separation	distance between the center of volumes of the paratopes	16 ± 6 Å
projected paratope area	two-dimensional projected area of the paratope in the orientation of maximum area	74 ± 25 AU
projected paratope perimeter	perimeter of the projected area of the paratope in the orientation of maximum area	1.2 ± 0.4 AU
	computational stability	
buried NPSA	amount of buried nonpolar surface area upon folding	2700 ± 900 Å^2
FoldX DDG	mean difference in stability from parental across 50 variants	17 ± 12 kJ/mol
FoldX energy	mean energy of 50 NNK variants using FoldX's forcefield	35 ± 25 kJ/mol
	general scaffold properties	
new SASA	amount of solvent exposed area created when removing unstructured termini	320 ± 260 Å^2
secondary structure percent	percent of residues in an α-helix or β-sheet	51 ± 12%
size	total number of residues in the scaffold	47 ± 7 AA

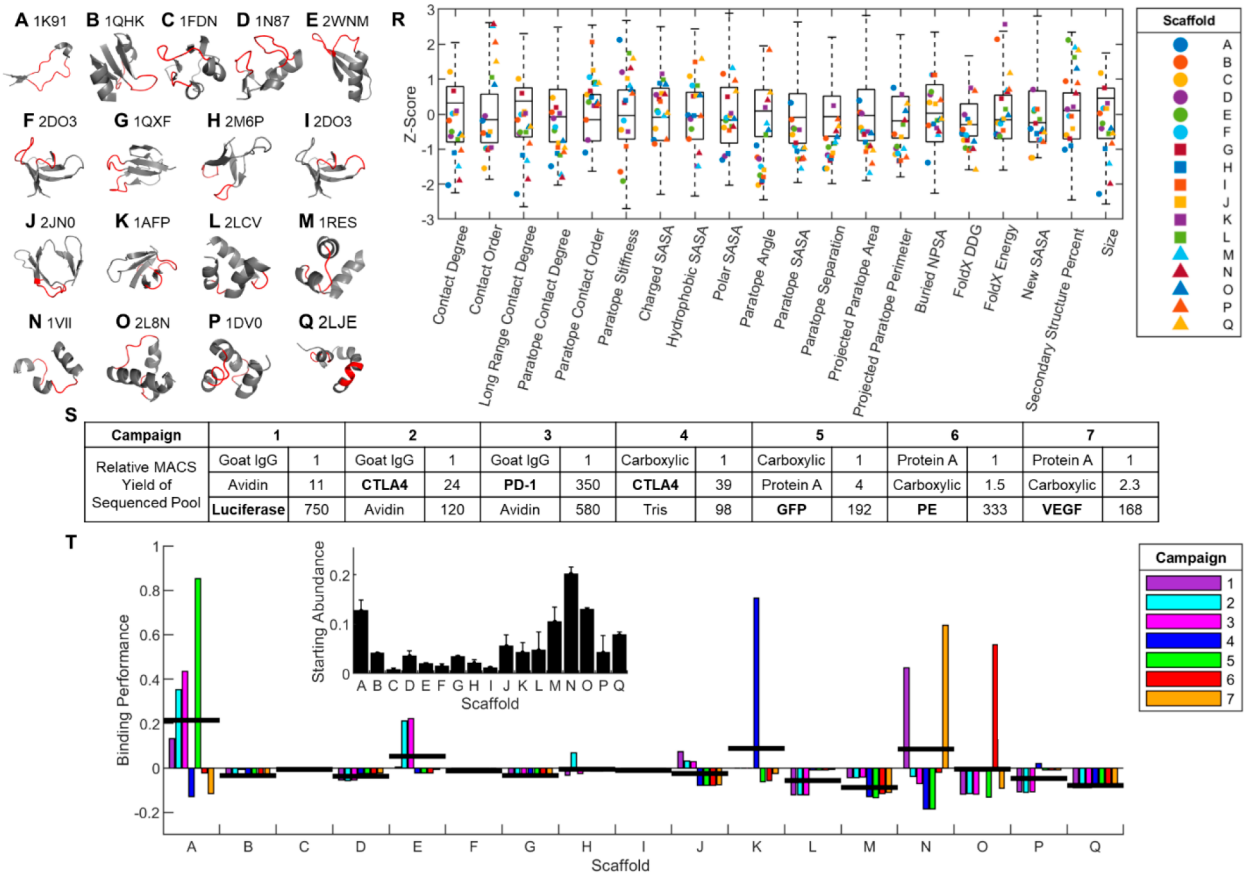
integrity when obtaining binding function. The FoldX empirical force field was used to estimate mutational destabilization and overall stability.[40] The amount of buried nonpolar surface area was also estimated as a relationship with stability was recently observed for small proteins.[41] (6) *General Scaffold Properties*. We propose the amount of new SASA introduced by cleaving termini may introduce destabilizing exposed surfaces. Termini without secondary structure were removed from experimental and computational analysis except in the calculation of new SASA. We also included descriptions of the amount of common secondary structure and total residues. Small protein topologies exhibit a broad range of values for these 20 parameters (Figure 2R), which provides potential utility for scaffold differentiation. Seventeen candidate scaffolds (Figure 2A−Q), which provide a range of characteristics (Figure 2R), were chosen for experimental evaluation.

Scaffold Binding Evaluation. To evaluate scaffold evolvability, we performed de novo discovery of binding ligands from a merged combinatorial library of all 17 scaffolds. Combinatorial libraries were genetically synthesized in which the two paratope loops were diversified with 8−17 (mean 11.3) "NNK" degenerate codons, which enable all 20 natural amino acids. The gene libraries were transformed into a yeast surface display system to robustly produce scaffold variants, which yielded 3−9 × 10^8 variants per scaffold. The 17 scaffold libraries were mixed resulting in a total diversity of 1 × 10^{10} protein variants. Deep sequencing revealed that the synthesized library matched design with only 1.2% median deviation from NNK diversity and a 1.1% framework mutation rate.

The pooled library was sorted to identify specific binding ligands to a panel of diverse proteins: luciferase, CTLA4, avidin, PD-1, green fluorescent protein, R-phycoerythrin, and vascular endothelial growth factor. Four to five rounds of magnetic activated cell sorting were used to deplete nonspecific binders and enrich selective binders. Maximum diversity of the sequenced population, estimated by the lowest-yielding sort with each cell containing a unique variant, ranged from 3500 to 715 000 per campaign. Enriched populations exhibited selective binding (Figure 2S) and were deep sequenced to characterize scaffold variants. 280 000 (range = 1250−115 000 per campaign) full-length reads were obtained yielding 21 000 (range = 160−9000 per campaign) unique binding variants. Individual campaign sorting and sequencing statistics are summarized in Table S1. With oversampled sorting, enrichment is correlated with binding affinity.[42] MACS sorts were performed with at least 10-fold diversity of yeast, allowing for differential recovery among clones of various binding strength. While our depth of sequencing did not fully sample the theoretical diversity, the differential frequencies of obtained variant reads suggests the obtained results reflect the differential affinities of the assayed scaffold variants. The overall binding performance of a scaffold was calculated as the mean difference in normalized abundance between the final and initial binding populations after transforming (quartic-root dampening[43]) sequence frequencies to combine the binding strength and the number of unique binding variants. It should be acknowledged that the binding performance metric in this study is dependent on the performances of the other tested scaffolds, and only provides a relative comparison between scaffolds. To define a threshold value of performance, a binding performance of −0.006 was determined to best classify experimental binding performance by the ability to develop a strong binding variant (Figure S1).

The assayed protein scaffolds possessed a range of ability to evolve novel binding function upon paratope mutations

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335



Campaign	1		2		3		4		5		6		7	
Relative MACS Yield of Sequenced Pool	Goat IgG	1	Goat IgG	1	Goat IgG	1	Carboxylic	1	Carboxylic	1	Protein A	1	Protein A	1
	Avidin	11	**CTLA4**	24	**PD-1**	350	**CTLA4**	39	Protein A	4	Carboxylic	1.5	Carboxylic	2.3
	Luciferase	750	Avidin	120	Avidin	580	Tris	98	**GFP**	192	**PE**	333	**VEGF**	168

Figure 2. Protein scaffold candidates show varying binding performance. (A−Q) The 17 assayed protein scaffolds with conserved region colored gray and variable paratope colored red. (R) 787 protein scaffolds of 30−65 amino acids with two solvent-exposed loops were computationally analyzed for 20 topological and biophysical factors (Table 1). The *z*-score distributions across all scaffolds are depicted by the box plots (box, 25−75th percentile; center bar, median; whiskers, 1.5 × interquartile range). The plotted values for each of the 17 assayed scaffolds indicate a diversity of proteins were assayed. (S) A pooled sample of 1×10^{10} variants across 17 scaffolds was enriched for binding variants in seven campaigns. MACS sorting was performed until seven binding populations were identified toward diverse molecular targets. Positive selection sorts (bold molecular target) were completed after two depletion sorts of the other listed targets. Binding functionality, quantified here as increased relative yield over control beads, was observed in all campaigns. (T) The relative binding performance for each scaffold against each molecular target as determined by the difference in scaffold abundance from the initial population to the binding populations. Scaffold abundance combines unique variants and variant binding strength using exponential dampening of sequence counts. Inset: The initial abundance of each scaffold. Error bars represent standard error (n = 3).

(Figure 2T). Five scaffold libraries failed to contain binding variants in any campaign: scaffolds C, F, and I maintained a near-neutral score as the starting abundance was rare, whereas scaffolds G and Q performed comparatively worse as each sequence had more potential to find binding variants. Scaffolds D and L produced binders to a single target. Yet, the binding was not strong relative to other binders, which rendered the scaffolds' overall performances as poor. Libraries of scaffolds A, B, E, H, J, K, M, N, O, and P contained binders to more than one target, with A, E, H, J, K, N, and O producing binders with sequences that occupied ≥1% of the reads for a campaign (Figure S2). Scaffolds J, H, O, and P increased abundance in at least one campaign but overall yielded a negative performance (i.e., depletion in frequency upon evolution).

Four scaffolds (A, E, K, and N) yielded an increased abundance across the study (Figure 3). Scaffolds A, E, and N had an increase in normalized abundance above 0.1 in two or more campaigns. Scaffold A, a binding subunit of the chaperone protein calreticulin with a relatively extended fold exposing both diversified loop regions, was found in all binding campaigns. Scaffold E, an RNA polymerase inhibitor, presents a pair of solvent-exposed loops on one end of a scaffold in which a single α-helix packs across from a β-sheet. This topology, recently identified via scaffold mining,[10] has been validated as a protein scaffold and serves as a positive control for this experiment. Scaffold N, an actin-binding protein presenting a pair of loops between three relatively small helices, obtained binding function in six campaigns with only 9 diversified sites. Scaffold K, an antifungal protein, dominated the fourth binding campaign and comprises three interacting β-sheets. These scaffolds offer diverse options for ligand evolution and provide, along with analysis of the other scaffolds, a means by which to evaluate the impact of topological and biophysical parameters on scaffold evolvability.

We would like to acknowledge a few limitations in the analysis of scaffold performance using the employed methodology in the experiment. Scaffold libraries may under- or overperform their overall evolvability for multiple reasons. The diversified sites may not be optimal as evolution can be aided by conservation of loop sites[44] and diversification of sites with secondary structure adjacent to paratope.[32,44] Full amino acid diversity is not optimal for evolution at many sites.[32,44] Yet the

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

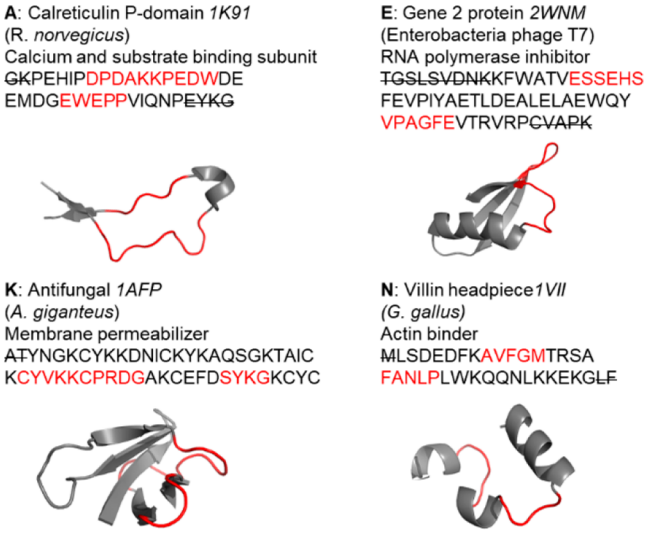


Figure 3. Successful protein scaffolds have diverse topologies. The identity, natural function, structure, and sequence of the top performing scaffolds are presented. The top proteins have various amounts and types of secondary structure. Diversified paratope residues are colored red in both the primary sequence and PyMOL rendering of the protein. Strikethroughs in the sequence represent residues present in the solved structure that were removed in our experimental analysis (as unstructured termini).

library designs that optimally balance intramolecular stability and intermolecular binding potential are not evident a priori. Thus, for consistency of scaffold evaluation, this common diversification strategy was employed. Additionally, assessing binding functionality via multivalent MACS with multivalent yeast display only requires moderate affinity. As our ability to identify functional scaffolds increases, modifying the selection stringency may modify scaffold performance and associated predictive parameters. There are several potential sources of variability in the experiments. Illumina preparation could have PCR bias;[45] however, initial library sequencing identified all scaffolds and our evolvability metric accounts for differences in initial abundance, which mitigates this issue. Additional differences in initial abundance could be explained by differential library construction efficiency. Severe undersampling of the theoretical 10^{16} variants yields potential stochasticity; however, the depth and breadth of evolved binders (21 000 unique sequences) provides a generalizable result. Finally, it is observed that not all scaffolds perform equally for all targets. The use of seven campaigns addresses this concern, and future experiments may benefit from further increasing campaign breadth.

Identifying Evolvable Scaffold Properties. To evaluate a generalizable impact of topological and biophysical parameters on scaffold evolvability, a tandem independent component analysis (ICA) and elastic net regularization protocol was performed. Given the extensive resources required to evaluate numerous scaffold performances, we sought to predict performance from our limited data set while avoiding overfitting. Briefly, the 20 calculated factors for 787 potential scaffolds were z-transformed and subsequently whitening transformed by principal component analysis to determine orthogonal metavariables, which describe variability between scaffolds in lower dimensional space and remove correlation (Figure S3). Six scaffold features were then reconstructed using ICA to identify underlying independent features describing protein scaffolds (Figure S4). The six independent components for the 17 assayed scaffolds were then fed into an elastic net regularization to determine predictive descriptions of scaffold binding performance. Regularization penalizes the norm of term coefficients, removing terms which do not aid predictive power. The technique isolated two components which best reduced a leave-one-out (LOO) root mean squared error (RMSE) in predicting scaffold performance (Figures 4A and S4). The final model was composed of a constant term, to account for bias in the definition of scaffold performance, and two independent components. The most predictive model successfully identifies 4 of the 6 functional scaffolds above the determined threshold. Nine of the 11 scaffolds predicted to be less evolvable indeed fit that description. Yet the model does result in false positives for 2 of 6 scaffolds.

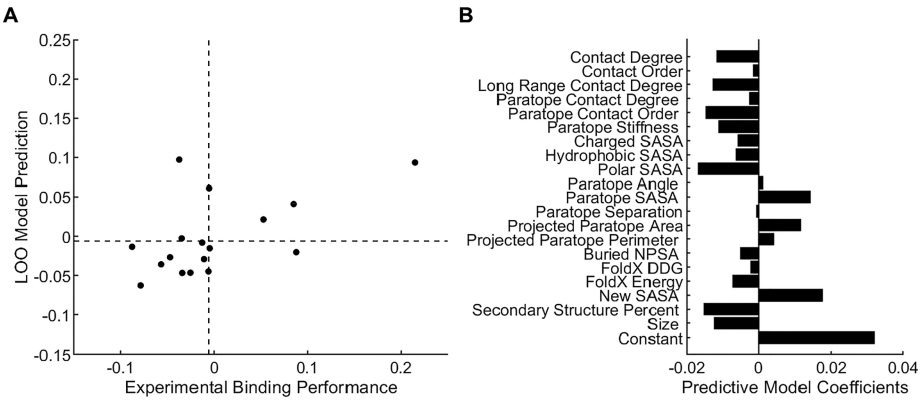


Figure 4. Large disconnected paratopes are associated with increased binding performance. ICA analysis was completed to describe the independent features of protein scaffolds. Elastic net regularization was performed to determine which of the features predicted binding performance. The resulting linear model was composed of two independent components and a constant term yielding a LOO RMSE of 0.06. (A) The LOO prediction of scaffold binding performance obtained a 4/6 true positive rate, a 9/11 negative predictive value, and a precision (positive predictive value) of 4/6. Classification threshold was determined by ability to evolve a strong binding variant. (B) The predictive model is a linear combination of the 20 calculated parameters and a constant term. The coefficients describe which parameters to modify to improve binding performance of a small protein scaffold.

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

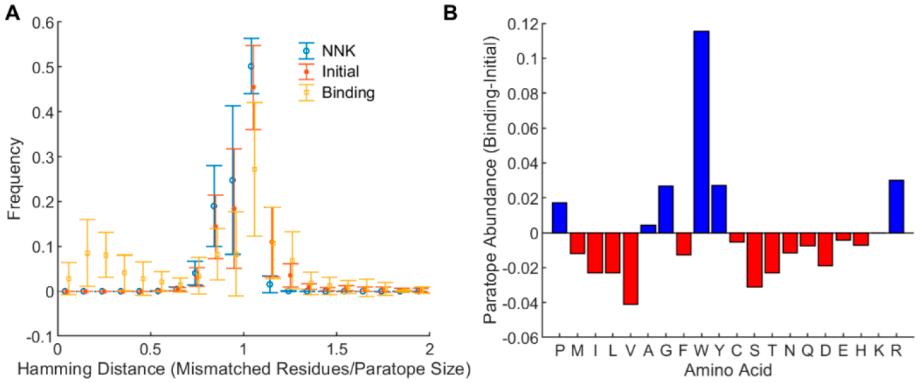


Figure 5. Binding variants describe functional amino acid space. (A) The diversity of sequenced variants based upon matched residues per position. NNK distribution was estimated via 5000 random NNK paratope-diversified sequences with a 1/1000 chance of framework mutations (Q30). The Hamming distance was then summarized by 20 bins based upon the number of mismatched residues per paratope size. Error bars represent standard deviation of Hamming distance frequencies across scaffolds (n = 17 for NNK and initial, n = 12 for binding). (B) The change in amino acid frequencies of binding variants relative to the initial library for all paratope sites across all scaffolds.

By distributing the weights of the independent components in the model back onto the calculated biophysical parameters, we can hope to obtain a physical understanding of what predicts scaffold success. On the basis of the linear model term coefficients, the predicted model suggests generally decreasing scaffold connectivity, paratope connectivity, conserved exposed surface area, buried nonpolar surface area, FoldX energy, secondary structure, and size (Figure 4B). It also suggests increasing paratope 2D and 3D surface area, 2D perimeter, and exposing new surface area upon removal of unstructured termini. While an exact interpretation of the model is complex, a general trend appears to suggest a large, disconnected paratope may predict increased binding performance. The distribution of binding performance of all predicted scaffolds can be found in Figure S5.

While several approaches to identify predictive biophysical parameters could have been utilized, we identified what we believe to be the most compelling approach using underlying features of protein scaffolds. For thoroughness, we also tested a similar approach utilizing principal components, which best describe differences between scaffolds, yielding a comparable outcome in terms of predictability and parameter insight (Figure S6). Both models agree on reducing protein and paratope contacts, minimizing conserved SASA, and increasing paratope SASA yet differ in the impact of paratope stiffness, FoldX energy, and new SASA. In a third approach, each individual parameter was analyzed to determine predictive performance. The top two predictive models in terms of minimizing LOO RMSE also suggest a decrease in conserved polar SASA or an increase in paratope SASA.

Paratope Analysis. We sought to analyze the characteristics of the evolved scaffold variants to illuminate any trends which may aid in future paratope design. We first asked if the binding variants for each scaffold were closely related in sequence space by plotting the distribution of pairwise Hamming distances for each scaffold. (Figure 5A). A paratope size normalized Hamming distance of 1 represents a completely unique paratope by position. A distance less than 1 represents variants with more similar paratope motifs. On the basis of the Hamming distance, only 2 of 12 binding scaffolds significantly reduced the sequence space from their initial distribution (P < 0.05, one-tailed Kolmogorov−Smirnov Test with Bonferroni correction for multiple comparisons). The similar Hamming distance distribution between the initial and binding populations provides evidence that the populations have roughly the same extent of diversity. The decreased distance for some scaffolds suggests that not all sequence space is functional in evolving novel binding function for some scaffolds but proves the results of our assay are not dominated by single binding motifs. Additionally, the mutational rate of the conserved residues of the binding proteins was 5% (relative to 1.1% in the naïve library), suggesting some mutations outside of the paratope may benefit binding evolution.

We then analyzed the evolution of paratope composition to assess the impact of particular amino acids on the creation of binding function (Figure 5B). Tryptophan and tyrosine, increased by 12% and 3%, respectively, have been previously reported to interact specifically across many interfaces because of the ability to partake in different bonds including π-stacking, hydrogen-bonding, and cation−π interactions.[46−48] Arginine, which often serves as a hot-spot residue for key interactions but has also been previously associated with nonspecific interactions, increased by 3%.[46−48] Glycine increased abundance by 3% perhaps by adding flexibility to the loop regions.[49] Proline increased in abundance by 2%, perhaps by improving scaffold stability by reducing the conformational entropy of the unfolded state.[49] Interestingly, serine has previously shown to be upregulated in binding variants but was greatly reduced in this study.[46−48] The raw abundance for each residue in the various sequencing populations is depicted in Figure S7.

Developability Impacts Scaffold Performance. In addition to evolving novel binding function upon mutation, the developability of a protein scaffold is also important for utility as a molecular targeting agent. We define a developable protein to possess high producibility, stability, solubility, and other usability factors. While the preceding experimental evolution did not directly select for developability, we sought to provide an introductory analysis of developability metrics of the studied scaffolds. We produced protein scaffold variants recombinantly in *Escherichia coli* to determine if recombinant yield was predictive of scaffold performance (Figure 6). Parental proteins, evolved binding variants, and random variants from the naïve library were expressed via pET plasmids in T7 Express *E. coli*. The identification of soluble protein was performed via PAGE gel analysis, FPLC purification, and anti-His tag ELISA. We found that modifying temperature and time of induction impacted protein yield for

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

		Parental Protein Producibility (Variant Producibility)	
		+	—
Ability to Evolve Strong Binding Variant	+	A (2/25) J (4/6) K (2/2) O (1/2)	E (0/2) H (1/1) N (4/6)
	—	D G	B C F I L (1/3) M P (0/3) Q

Figure 6. Limited protein producibility highlights the importance of scaffold developability. Each scaffold is classified by the ability to develop a strong binder (abundance > 1% in at least one campaign) and the parental protein producibility (ability to produce in T7 *E. coli* in detectable soluble yields). If applicable, the producibility of scaffold variants are shown as no. produced/no. attempted.

producible clones but did not recover any poorly produced proteins.

On the basis of the detection of parental protein in the soluble fraction of T7 *E. coli*, scaffolds whose parental protein is effectively produced in the soluble fraction have a higher probability of evolving a strong binding variant (one-tailed two-sample proportion test, $p = 0.057$). Under the hypothesis that proteins expressed must be stable, have low aggregation propensity, and readily fold, this data suggests that well-behaved proteins will serve as a better starting point for scaffold discovery. Additionally, the data recommend that protein scaffolds should be derived from highly developable proteins, rather than engineering developable parameters postidentification of binding functionality. Interestingly, the ability of a parental clone to produce was not indicative of variant producibility ($p = 0.3$).

Proteolytic Stability. We then sought to characterize the stability of scaffold variants on the surface of yeast, where binding function was observable and more complex protein production machinery exists. Using proteinase K, flow cytometry, and deep sequencing, the relative proteolytic stability of 1300 unique scaffold variants were determined by analyzing the amount of protease required to cleave the distal epitope tag on a yeast surface displayed scaffold variant (Figure 7A). The method could be influenced by protein aggregation protecting variants from cleavage. Notably, the scaffold A parental variant was resistant to cleavage yet found in multimeric states on PAGE gels and mass spectrometry upon recombinant soluble expression. Nevertheless, this high-throughput analysis informs on stability as recently validated.[41]

We first examined the stability of the parental variants for each scaffold and observed a positive correlation with the scaffold's binding performance during MACS sorting (Spearman's $\rho = 0.56$, $p < 0.05$; Figure 7B). The shape appears to suggest a threshold of stability is required to obtain high binding performance. We then tested the hypothesis that the stability of random diversified variants could correlate to parental protein stability. We measured the stability of an average of 60 variants per scaffold (range = 14−73; Figure S8). A large range of stabilities were observed among the naive variants without any evident correlation with parental stability (Spearman's $\rho = 0.43$, $p = 0.1$). This outcome could be

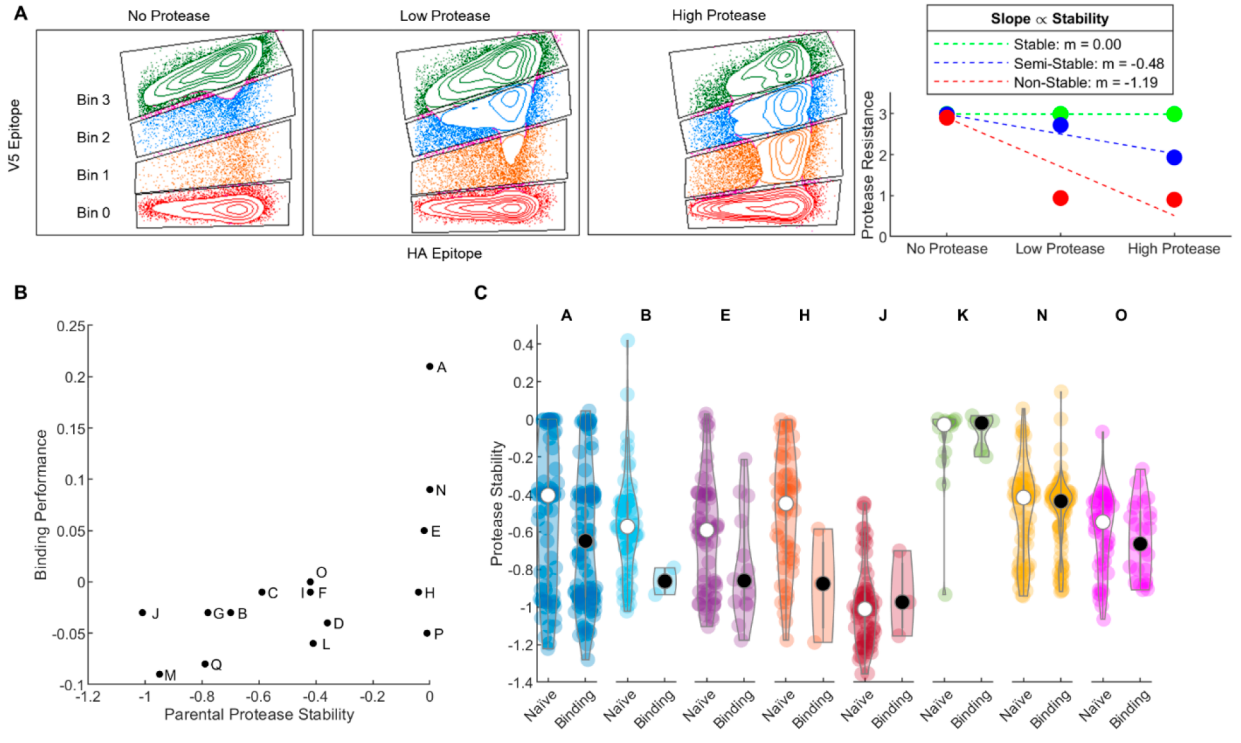


Figure 7. Proteolytic stability assay identifies stability requirement for binding. (A) Protein scaffold variants were exposed to various levels of proteinase K and sorted based on degree of cleavage on the surface of yeast. The slope of the protease resistance (i.e., collection bin) versus protease concentration is correlated to protein stability. (B) The proteolytic stability of the parental scaffold is correlated to the binding performance of the scaffold. (Note: n.d. for Scaffold K.) (C) Violin plot comparing stabilities of naïve variants and binding variants. A Wilcoxon one-tailed signed rank test indicates that binding variants are less stable than naïve variants ($p = 0.034$).

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

explained by the substantial diversification of the initial pool, which is likely to contain variants both close and far from the parental clone.

A final comparison was performed between stabilities of naïve variants and binding variants for each scaffold. Interestingly, the protease stability of binding variants is significantly lower than that of nonbinding variants (one-tailed Wilcoxon signed-rank test on set medians, $p = 0.034$; Figure 7C). This suggests there is a trade-off between binding functionality and stability, as previously hypothesized.[50,51]

Paired with the relationship between parental protease stability and scaffold binding function, we hypothesize that protein scaffolds with high protease stability will more efficiently evolve binding variants because they can "sacrifice" stability while remaining folded. This suggests that the search for future protein scaffolds should first involve a comprehensive study of protein stabilities and expression. This additional test may aid in the differentiation of proteins with otherwise similar biophysical properties when predicting evolvability as protein scaffolds.

CONCLUSION

The current study develops a computational-experimental platform to identify successful protein scaffolds and provides insight on the topological and biophysical parameters that dictate evolvability. However, the ability to develop specific binding function is not enough for a scaffold to be useful in downstream applications. The stability and producibility of the proteins also determine scaffold utility. Interestingly, these developability factors also correlate to binding evolvability of the protein scaffold. Future work in this field should combine the predictive biophysical model and the observed relationship between protein stability and scaffold functionality to narrow the assayed candidates.

We also note that this method of computationally calculating biophysical parameters of proteins to relate to desired functionality is applicable beyond protein scaffold identification. A similar analysis could be completed to determine predictive performances of protein developability metrics, enzyme efficacy, and antimicrobial peptide activity. The current limitation in such studies is the collection of a sufficiently rich data set to build a robust computational model.

EXPERIMENTAL PROCEDURES

Scaffold Parameter Calculation. Protein Data Bank files were obtained for files containing a protein chain ranging from 30 and 65 amino acids. Chains were then parsed for unique sequence and secondary structure as determined by the depositor. Paratope loop regions were assigned as continuous stretches of at least four amino acids without secondary structure. Terminal amino acids were removed if located at 3 or more residues from the outermost secondary structure. Homemade Python scripts were then used to calculate 20 parameters. Scripts are available online on GitHub: https://github.com/HackelLab-UMN.

Protein Connectivity. We hypothesize that a more connected protein is correlated to increased stability but decreased mutational stability. The distances between residue β-carbons (or α-carbon for glycine) are measured for all residues in the terminal-trimmed protein. Residues with Euclidian distances of ≤8 Å are considered contacts, consistent with ranges found in literature.[37] Three parameters are calculated: (1) contact degree, the total number of contacts;

$$\text{contact degree} = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\begin{cases}1 & \text{AA}_i, \text{AA}_j \le 8\ \text{Å}\\ 0 & \text{else}\end{cases}$$

(2) contact order, the sum across all contacts of the difference in primary sequence index, normalized by contact degree and the total number of residues;

$$\text{contact order} = \frac{\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\begin{cases}j-i & \text{AA}_i, \text{AA}_j \le 8\ \text{Å}\\ 0 & \text{else}\end{cases}\right)}{N \times \text{contact degree}}$$

and (3) long-range contact degree, the number of contacts with difference in primary sequence index greater than 12, normalized by the total number of residues.

$$\text{long-range contact degree} = \frac{\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\begin{cases}1 & \text{AA}_i, \text{AA}_j \le 8\ \text{Å and } j-i>12\\ 0 & \text{else}\end{cases}\right)}{N}$$

Paratope Connectivity. We hypothesize that less connected and more flexible paratopes will be more accepting of diversification required to obtain binding function by limiting the destabilization of the entire protein. Contacts were calculated between paratope residues and conserved residues within 8 Å. Normal mode analysis[52,53] was used to estimate the flexibility of the paratope as determined by its connectivity to the remainder of the protein. Three parameters are calculated: (4) paratope contact degree, the number of contacts between a paratope residue and a conserved residue;

$$\text{paratope contact degree} = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\begin{cases}1 & \left\|\text{AA}_i, \text{AA}_j\right\|_2 \le 8\ \text{Å and } \text{AA}_i \oplus \text{AA}_j \in \text{paratope}\\ 0 & \text{else}\end{cases}$$

(5) paratope contact order, the sum of paratope contacts' difference in primary sequence index, normalized by paratope contact degree and the number of paratope residues;

$$\text{paratope contact order} = \frac{\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\begin{cases}j-i & \left\|\text{AA}_i, \text{AA}_j\right\|_2 \le 8\ \text{Å and } \text{AA}_i \oplus \text{AA}_j \in \text{paratope}\\ 0 & \text{else}\end{cases}\right)}{\text{size of paratope} \times \text{paratope contact degree}}$$

(6) paratope stiffness, the average of the z-score transformed mean mechanical stiffness spring constant of paratope residues' α-carbon calculated by an anisotropic network model[38]—high stiffness suggests a less flexible and more connected residue.

Conserved Surface Area Chemical Nature. We hypothesize that the type of conserved exposed surface area will affect protein scaffold stability. The solvent accessible surface area (SASA), as determined by the radius of a water molecule in PyMOL, was summed for each residue based upon chemical nature. Chemical categorization led to three parameters: (7) charged (D, E, K, R) SASA, which may aid in protein stability by creating surface intramolecular salt bridges; (8) hydrophobic (A, F, G, I, L, M, P, V) SASA, which is likely

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

destabilizing because of the entropic cost of solvation; (9) polar (C, H, N, Q, S, T, W, Y) SASA, which may contribute to stabilization in polar solvents.

Paratope Size and Topology. We hypothesize that two large and spatially close paratope regions will maximize the binding surface and increase the total energetics of binding toward the molecular target. Three parameters were based upon 3D structural data: (10) paratope angle, the [paratope 1: entire protein: paratope 2] angle based upon the atomic center of volume; (11) paratope SASA, calculated after mutating all paratope residues to alanine in PyMOL; (12) paratope separation, the distance between atomic center of volumes of the paratopes. A 2D projection, created by modifying PyMOL's depth cue, fog, and lighting, was also used for two 2D parameters: (13) projected paratope area, the sum of the pixels containing the paratope residues' projection and (14) projected paratope perimeter, the number of paratope pixels bordered by a non paratope pixel. To obtain the 2D projections, the protein was rotated to determine the projection with the maximum area of the paratope. The background and conserved residues are colored black with the epitope colored white. A ray-traced image is populated, and the pixel intensity is counted using Python's Image Library. Both area and perimeter were normalized by the pixel area of a pseudoatom placed at the center of the paratope regions.

Computational Stability. We hypothesize that protein stability will impact mutational tolerance[50] and sought to computationally estimate stability based upon existing correlations. Three parameters were calculated: (15) buried nonpolar surface area (buried NPSA),[41] the sum of solvent exposed nonpolar amino acids in Gly-X-Gly[54] minus the sum of solvent exposed nonpolar amino acids in the folded protein; (16) FoldX DDG, the mean difference in force field energy between mutant and parental variants; and (17) FoldX Energy, the mean force field energy of predicted scaffold mutants. For FoldX calculations, 50 variants randomly selected from an NNK distribution were simulated by FoldX 4,[40] which is sufficient to obtain a 5.1% average coefficient of variation (n = 3 sets of 50 variants).

General Scaffold Properties. We hypothesize that additional factors, which are not explicitly included in categories above, may also impact scaffold performance. Three factors were included: (18) new SASA, the amount of new SASA of scaffold residues after unstructured tails are removed; (19) secondary structure percent, the percentage of scaffold residues categorized as part of an α-helix or a β-sheet; and (20) size, the number of residues in the scaffold after removal of non-secondary structured termini.

Binder Discovery. We first sought to select proteins with small size, strong computed mutational stability, large and spatially proximal paratopes, minimal newly exposed SASA upon terminal trimming, and a small ratio of perimeter^2 to area for the projected paratope. The weights assigned to each factor were randomly assigned and 24 scaffolds were selected for testing from the 619 initial candidates: 8 containing α-helices, 8 containing β-sheets, and 8 containing both secondary structures. Twenty-four scaffolds were chosen to balance breadth of parental proteins and experimentally achievable depth of scaffold variants. Seven of the 24 synthesized libraries had less than 3/10 clones match design and were removed from the study. Genetic combinatorial libraries were synthesized to encode for the 17 scaffolds with full amino acid diversity at the paratope sites encoded via NNK codons. Oligonucleotides for these libraries were purchased from LabGenius. Genes were amplified via PCR (200 μL, 1 μM primers, 200 μM dNTPs, 10 U Taq Polymerase, 1× ThermoPol Buffer, 0.5 μM template gene, 30 cycles) and concentrated via ethanol precipitation with PelletPaint (Millipore Sigma). Yeast display plasmid providing an N-terminal Aga2p, an HA epitope, a flexible $(G_4S)_3$ polypeptide linker, and a C-terminal AU5 epitope (pCT-AU5), was produced in NEB5α *E. coli* (New England Biolabs) and purified via silica spin column (Epoch Life Science) according to manufacturer's protocol. The vector was linearized via restriction digest with *Nde*I, *Pst*I-HF, and *Bam*HI-HF (New England Biolabs). Digested vector was ethanol precipitated and resuspended in deionized water. For each scaffold, 6 μg digested vector and all ethanol concentrated genes were transformed into *Saccharomyces cerevisiae* yeast (EBY100) via homologous recombination. Transformation followed previously described protocols,[55] with the addition of 30% v/v PEG 8000 in step 39, which was found to increase transformation efficacy.[56] Transformed sequence diversity was estimated by dilution plating onto selective media assuming all transformants were unique. Anti-AU5 antibodies failed to isolate full length display constructs; thus, nonsense sequences were obtained during sequencing, but omitted from analysis.

The 17 scaffold yeast libraries were grown and induced as previously described,[55] and 10× the transformed diversity of each sublibrary was mixed to create a pooled library. For each round of magnetic-activated cell sorting (MACS), induced yeast were rotated with magnetic beads for 2 h at 4 °C and placed on a magnet for 5 min to isolate binding variants. Each round of MACS consisted of depletion sorts on two negative targets followed by enrichment on positive target beads. For depletion sorts, nonbinding yeast were collected for the next sort and binding yeast were plated for quantification. For enrichment sorts, the bound yeast were collected and grown for subsequent rounds. Yeast binding to both positive and negative target beads were washed with 1 mL of PBSA (1× phosphate buffed saline with 1 g/L bovine serum albumin, once for the first two rounds and thrice for additional rounds), and resuspended in selective growth media. A diluted fraction was plated for quantification. Positive selectivity (more yeast binding to positive target beads relative to negative target beads) was found after four to five rounds of MACS based upon plated recovery.

A variety of protein targets were used to represent the diversity of potential molecular targets of protein scaffolds. Biotinylated green florescent protein (GFP), and *Gaussia princeps* luciferase (luciferase) were purchased from Avidity. Biotinylated human PD-1 extracellular domain and human CTLA4 extracellular domain were purchased from G&P Biosciences. Biotinylated R-phycoerythrin (PE) was purchased from AssayPro. Biotinylated human VEGF121 was purchased from ACROBiosystems. Protein targets were either added to Dynabeads Biotin Binder (ThermoFisher) or Dynabeads M-270 Carboxylic Acid beads, as described below. For selections on carboxylic acid beads, counter-sorts included bare carboxylic acid beads, tris(hydroxymethyl)aminomethane (Tris)-quenched carboxylic acid beads, or Dynabeads Protein A (ThermoFisher). For selections on avidin-coated Biotin Binder beads, counter-sorts included bare avidin beads and biotinylated goat IgG (Rockland Immunochemical) on avidin beads.

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

Campaigns 1−3 were completed with 16.5 pmol/bead biotinylated protein targets conjugated to avidin beads. Campaigns 4−7 were completed with 33 pmol/bead targets conjugated to avidin beads for the first and third round and to carboxylic acid beads for the second and fourth rounds (and fifth round for campaign 4). Campaigns 1, 5, 6, and 7 isolated binders toward luciferase, GFP, PE, and VEGF121, respectively. Campaigns 2, 3, and 4 isolated binders toward CTLA4/Avidin, PD 1/Avidin, and CTLA4/Tris. Though binding was observed toward two molecules, the specificity over a third negative target signifies an enriched population with binding functionality. For avidin-based sorts, 10 μL of beads were mixed with 5 or 10 μL of 3.3 μM target in 100 μL of PBSA; beads were rotated at room temperature for 1 h, isolated via magnet, aspirated, and washed with 1 mL of PBSA before cells were added to the tube. For carboxylic acid sorts, manufacturer's two-step coating protocol (without NHS) was followed except for the following modification: 2 μL of beads were used for each target to match total beads to avidin sorts.

Evaluation of Binder Performance via Deep Sequencing. DNA encoding for scaffolds was isolated from yeast using Zymolyase (Zymo Research). Briefly, 1×10^8 cells are incubated in 200 μL of lysis solution (50 mM phosphate buffer, 1 M sorbitol, 10 mM β-mercaptoethanol, and 75 U/mL zymolyase longlife) for 30 min at 37 °C after which DNA is extracted via silica spin column. PCR addition of Illumina adapters was performed to sequence scaffold genes in the initial and binding pools using Illumina MiSeq. Sequences were filtered using PANDASeq[57] with a confidence threshold value of 0.9 for primer and assembled reads. Scaffold identification was completed via homemade MATLAB scripts available on GitHub. Briefly, sequencing reads were translated, and filtered for sequences matching 70% of the $(G_4S)_3$ linker and AU5 tag. The scaffold was identified by sequences of the same length and 70% match of conserved residues. Unique sequence counts were based upon translated sequences.

Three independent sequencing runs of the initial unsorted pool were completed, with at least 10 000 scaffold variants identified in each sample. The distribution of paratope residues reasonably matched the intended NNK diversity (median absolute deviation = 1.2%, Figure S7). The conserved residues had a mutational rate of 1.1%. To determine the distribution of sequences analyzed, the Hamming distance was calculated between all observed sequences. Comparison to computationally simulated NNK sequences indicated diverse sequence sampling with 15 of 17 libraries not significantly more clustered in sequence space than designed (Figure 4, $P > 0.05$, one-tailed Kolmogorov−Smirnov test with Bonferroni correction for multiple comparisons).

Binding populations were individually barcoded and sequenced, yielding 280 000 full length reads across the seven binding populations. The binding performance of each scaffold is a function of the number of unique binders and the strength of binders. However, utilizing the raw read counts leads to descriptions of binding pools dominated by the strongest binding variants. One such method of combining diversity and binding functionality is exponential dampening.[43] Therefore, the number of reads for each unique sequence was quartic root dampened (a subjective balance to reward clonal performance, while dampening dominant clones to provide information from diverse clones), and the abundance of a scaffold is the total fraction of dampened reads per molecular target.

$$\text{abundance (scaffold X)} = \frac{\sum_{\text{unique sequence } i=1}^{\text{sequences for scaffold X}} \text{reads of sequence}_i^{1/4}}{\sum_{\text{unique sequence } i=1}^{\text{sequences for all scaffolds}} \text{reads of sequence}_i^{1/4}}$$

To account for differences in starting abundance, the final binding performance metric was calculated as the mean difference in abundance for the seven scaffolds. It should be noted the binding performance metric is dependent on the other scaffolds assayed, yet it still provides a relative performance between scaffolds. To estimate a threshold value of useful binding performance, scaffolds were classified by the ability to develop a high affinity binding variant with >1% campaign abundance (A, E, H, J, K, N, O). A receiver operating characteristic curve was used to determine a binding performance threshold of −0.006 (Figures S1 and S2).

Evolutionary Model. With more calculated parameters than experimental data points (i.e., scaffolds), we sought to reduce the scaffold parameter space and avoid overfitting of a predictive model. We believe that some calculated parameters may be correlated and hypothesized we could describe the scaffolds using a smaller dimensional space of underlying features. Reconstructive independent component analysis (ICA) attempts to identify features by separating the data set into mutually independent latent variables.[58] ICA requires a whitening transformation of data to remove correlation, which was achieved via principal component analysis (PCA). PCA can be used to reduce dimensionality by describing scaffolds with orthogonal metavariables, which removes low order correlations.[59] Broadly, ICA describes features of protein scaffolds, whereas PCA describes features that best differentiate protein scaffolds.

The calculation of the parameters was finalized and calculated for 787 protein scaffold candidates via scripts available on GitHub. All parameters were calculated via a deterministic algorithm with a singular result per scaffold, except for FoldX calculations described above which were performed on random library variants. Principal components were then calculated via singular value decomposition using the *pca* function in MATLAB's Statistics and Machine Learning Toolbox. The first six components, which individually explained at least 5% of the variance in scaffold parameters with a sum of 80% total explained variation, were retained to predict scaffold performance (Figure S3). Independent components were then obtained via a modification of ICA with a reconstructive cost using the *rica* function in MATLAB (Figure S4).

We then sought to determine which of the independent components best predicted scaffold binding performance. Regularization is a technique used to remove parameters which are not predictive of a desired characteristic.[60] A penalty term included in the objective function, associated with the norm of term coefficients, prevents overfitting of data by driving the coefficients of noisy inputs to zero. The six independent components for the 17 experimentally tested scaffolds were used to predict the observed binding performance using the MATLAB regularization function *lassoglm* with leave-one-out estimation of deviance. Elastic net regularization was performed with various penalty calculations of the L1/L2 norm (α = 0.01, 0.1 0.25, 0.5, 0.75, 1) and maximum number of model terms allowed (DFmax = 1−6). The performance of the regularization output was tested via leave-one-out

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

prediction of the assayed scaffolds. The model with the lowest root-mean-squared-error of binding performance prediction was identified. MATLAB scripts for ICA/PCA analysis and regularization can be found on GitHub. The ability of the predictive model to identify functional scaffolds was based upon the threshold determined by the ability to develop strong binding variants.

Protein Production. Genes encoding for observed and parental scaffold variants were obtained from Twist Bio-Science. Genes were ligated into pET production plasmids with a C-terminal His_6 tag and transformed into T7 Express Competent *E. coli* (New England Biolabs) following manufacturer's protocol. Cells were induced at 37 °C for 2 h with 0.5 mM isopropyl β-D-1-thiogalactopyranoside, pelleted, and frozen. The cells were then lysed in (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) (HEPES) lysis buffer (50 mM HEPES, 5 mM CHAPS, 25 mM imidazole, 2 mM $MgCl_2$, 20 mM NaCl, 7 U/μL benzonase, 50 mg/mL lysozyme, EDTA-free protease inhibitor, and 5% v/v glycerol) and incubated at 37 °C for 30 min before centrifugation and isolation of the soluble fraction. Protein purification was performed using HisTrap HP columns on an ÄKTAprime plus (GE Healthcare) with wash buffer (20 mM HEPES, 500 mM NaCl, 20 mM imidazole, pH 7.4) and elution buffer (20 mM HEPES, 500 mM NaCl, 500 mM imidazole) flowed at 1 mL/min.

To quantify protein via ELISA, 100 μL of soluble lysate fraction was incubated in a 96-well plate overnight at 4 °C, washed 4× with 0.05% v/v Tween 20 in PBS via squirt bottle and patted dry. Plates were incubated in 100 μL of 0.1 μg/mL Anti-6X His tag HRP antibody (ab1187, Abcam) in PBS for 1 h at room temperature, washed 4×, treated with 100 μL of 3,3′,5,5′-tetramethylbenzidine (TMB) for 15 min, followed by 100 μL of TMB Stop Solution (ThermoFisher). His-tagged protein abundance was measured via absorbance at 450 nM using a plate reader. Known purified biotinylated protein was spiked into lysate without His-tagged protein to quantify the limit of detection: 2 mg of protein per liter of bacterial culture.

Identification of produced protein was obtained via PAGE gel with and without nickel column purification or an Anti-His6 ELISA performed compared to a non-His tagged control protein. NuPAGE Bis-Tris Gels were used to identify the addition of a protein at the expected molecular weight based upon protein standard following manufacture's protocol.

Proteolytic Resistance. Genes encoding for observed and parental scaffolds were transformed into a yeast surface display construct with N-terminal HA and C-terminal V5 epitope tags (PCT-V5) as described above, except gene preparation was performed via 400 μL PCR using Phusion polymerase (New England Biolabs). One $\times 10^6$ yeast induced to display protein were incubated in 50 μL of PBSA with 0, 4×10^{-6}, or 22×10^{-6} U/μL proteinase K at 37 °C for 10 min, and immediately washed with cold PBSA. Epitope tags were labeled with chicken anti-HA antibody (ab9111, Abcam) and mouse anti-V5 antibody (ab27671, Abcam), followed by AlexaFluor488-conjugated goat antichicken IgY (H+L) (Thermo Fisher Scientific) and AlexaFluor647-conjugated goat antimouse IgG (H+L) (Thermo Fisher Scientific). Labeling was performed as follows: 1×10^6 cells were rotated for 30 min at room temperature in 50 μL of PBSA with 1 ng/μL primary antibodies, pelleted at 8000*g* for 1 min, aspirated, washed with 1 mL of PBSA, incubated for 20 min at 4 °C in 50 μL of PBSA with 1 ng/μL secondary antibody; pelleted, washed, and resuspended at 2×10^7 cells/mL in PBSA for florescence activated cell sorting (FACS). Cells were sorted into four gates (bins) based upon C-terminal: N-terminal epitope signal ratio, with a low ratio suggesting full cleavage of the protein. Collection bin 3 corresponds to intact protein, and collection bin 0 corresponds to fully cleaved protein.

Scaffold plasmids were extracted with Zymolase and PCR amplified with extension to add Illumina adapters as described above. Two experimental replicates were sorted and separately sequenced using Illumina HiSeq and processed using USearch[61] by filtering for a maximum 5% error rate per read and matching to ordered proteins. The mean collection bin of each protein was calculated for all three protease concentrations. For fully displayed proteins without protease, a line was fit with a fixed intercept corresponding to the no-protease collection bin. A zero slope indicates no decrease in mean collection bin (epitope signal ratio) with increasing protease concentration and suggests protease stability. The normalized deviation (magnitude trial difference average/range) across trials is 0.11 (Figure S9).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscombsci.8b00182.

Sorting and sequencing summary, calibration of binding performance, bubble plot of scaffold performance against each molecular target, principal component analysis, independent component analysis, predicted scaffold performance, alternative predictive models, amino acid abundance across all protein scaffold paratopes, proteolytic stability comparison, and proteolytic stability of yeast-displayed proteins (PDF)

Scaffold parameters (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 612.624.7102. E-mail: hackel@umn.edu.

ORCID

Benjamin J. Hackel: 0000-0003-3561-9463

Author Contributions

A.W.G., P.V.H., and B.J.H. conceived and designed the experiments, A.W.G. and K.M.M. performed the experiments, and A.W.G. and B.J.H. cowrote the manuscript and Supporting Information. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the National Institutes of Health (R01 EB023339) and a National Science Foundation Graduate Research Fellowship (to A.W.G.). We appreciate assistance from the University of Minnesota Flow Cytometry Core, University of Minnesota Genomics Center, and the Minnesota Supercomputing Institute (MSI) at the University of Minnesota.

■ REFERENCES

(1) Bershtein, S.; Tawfik, D. S. Advances in Laboratory Evolution of Enzymes. *Curr. Opin. Chem. Biol.* **2008**, *12* (2), 151−158.

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

(2) Romero, P. A.; Arnold, F. H. Exploring Protein Fitness Landscapes by Directed Evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10* (12), 866−876.
(3) Tokuriki, N.; Stricher, F.; Serrano, L.; Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **2008**, *4* (2), No. e1000002.
(4) Škrlec, K.; Štrukelj, B.; Berlec, A. Non-Immunoglobulin Scaffolds: A Focus on Their Targets. *Trends Biotechnol.* **2015**, *33* (7), 408−418.
(5) Banta, S.; Dooley, K.; Shur, O. Replacing Antibodies: Engineering New Binding Proteins. *Annu. Rev. Biomed. Eng.* **2013**, *15*, 93−113.
(6) Scott, A. M.; Wolchok, J. D.; Old, L. J. Antibody Therapy of Cancer. *Nat. Rev. Cancer* **2012**, *12* (4), 278−287.
(7) Stern, L.; Case, B.; Hackel, B. Alternative Non-Antibody Scaffolds for Molecular Imaging of Cancer. *Curr. Opin. Chem. Eng.* **2013**, *2* (4), 425−432.
(8) Packer, M. S.; Liu, D. R. Methods for the Directed Evolution of Proteins. *Nat. Rev. Genet.* **2015**, *16* (7), 379−394.
(9) Jain, T.; Sun, T.; Durand, S.; Hall, A.; Houston, N. R.; Nett, J. H.; Sharkey, B.; Bobrowicz, B.; Caffry, I.; Yu, Y.; et al. Biophysical Properties of the Clinical-Stage Antibody Landscape. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (5), 944−949.
(10) Kruziki, M. A.; Bhatnagar, S.; Woldring, D. R.; Duong, V. T.; Hackel, B. J. A 45-Amino-Acid Scaffold Mined from the PDB for High-Affinity Ligand Engineering. *Chem. Biol.* **2015**, *22* (7), 946−956.
(11) Frejd, F. Y.; Kim, K.-T. Affibody Molecules as Engineered Protein Drugs. *Exp. Mol. Med.* **2017**, *49* (3), e306−e306.
(12) Binz, H. K.; Bakker, T. R.; Phillips, D. J.; Cornelius, A.; Zitt, C.; Göttler, T.; Sigrist, G.; Fiedler, U.; Ekawardhani, S.; Dolado, I.; et al. Design and Characterization of MP0250, a Tri-Specific Anti-HGF/Anti-VEGF DARPin® Drug Candidate. *MAbs* **2017**, *9* (8), 1262−1269.
(13) Schiff, D.; Kesari, S.; de Groot, J.; Mikkelsen, T.; Drappatz, J.; Coyle, T.; Fichtel, L.; Silver, B.; Walters, I.; Reardon, D. Phase 2 Study of CT-322, a Targeted Biologic Inhibitor of VEGFR-2 Based on a Domain of Human Fibronectin, in Recurrent Glioblastoma. *Invest. New Drugs* **2015**, *33* (1), 247−253.
(14) Souied, E. H.; Devin, F.; Mauget-Faÿsse, M.; Kolár, P.; Wolf-Schnurrbusch, U.; Framme, C.; Gaucher, D.; Querques, G.; Stumpp, M. T.; Wolf, S. Treatment of Exudative Age-Related Macular Degeneration with a Designed Ankyrin Repeat Protein That Binds Vascular Endothelial Growth Factor: A Phase I/II Study. *Am. J. Ophthalmol.* **2014**, *158* (4), 724−732.
(15) Rothe, C.; Skerra, A. Anticalin® Proteins as Therapeutic Agents in Human Diseases. *BioDrugs* **2018**, *32* (3), 233−243.
(16) Stern, L. A.; Case, B. A.; Hackel, B. J. Alternative Non-Antibody Protein Scaffolds for Molecular Imaging of Cancer. *Curr. Opin. Chem. Eng.* **2013**, *2* (4), 425−432.
(17) Vazquez-Lombardi, R.; Phan, T. G.; Zimmermann, C.; Lowe, D.; Jermutus, L.; Christ, D. Challenges and Opportunities for Non-Antibody Scaffold Drugs. *Drug Discovery Today* **2015**, *20* (10), 1271−1283.
(18) Holliger, P.; Hudson, P. J. Engineered Antibody Fragments and the Rise of Single Domains. *Nat. Biotechnol.* **2005**, *23* (9), 1126−1136.
(19) Kobe, B.; Deisenhofer, J. The Leucine-Rich Repeat: A Versatile Binding Motif. *Trends Biochem. Sci.* **1994**, *19* (10), 415−421.
(20) Koide, A.; Bailey, C. W.; Huang, X.; Koide, S. The Fibronectin Type III Domain as a Scaffold for Novel Binding Proteins. *J. Mol. Biol.* **1998**, *284* (4), 1141−1151.
(21) Binz, H. K.; Amstutz, P.; Kohl, A.; Stumpp, M. T.; Briand, C.; Forrer, P.; Grütter, M. G.; Plückthun, A. High-Affinity Binders Selected from Designed Ankyrin Repeat Protein Libraries. *Nat. Biotechnol.* **2004**, *22* (5), 575−582.
(22) Beste, G.; Schmidt, F. S.; Stibora, T.; Skerra, A. Small Antibody-like Proteins with Prescribed Ligand Specificities Derived from the Lipocalin Fold. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (5), 1898−1903.
(23) 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.* **1997**, *15* (8), 772−777.
(24) Grabulovski, D.; Kaspar, M.; Neri, D. A Novel, Non-Immunogenic Fyn SH3-Derived Binding Protein with Tumor Vascular Targeting Properties. *J. Biol. Chem.* **2007**, *282* (5), 3196−3204.
(25) Kolmar, H. Alternative Binding Proteins: Biological Activity and Therapeutic Potential of Cystine-Knot Miniproteins. *FEBS J.* **2008**, *275* (11), 2684−2690.
(26) Correa, A.; Pacheco, S.; Mechaly, A. E.; Obal, G.; Béhar, G.; Mouratou, B.; Oppezzo, P.; Alzari, P. M.; Pecorari, F. Potent and Specific Inhibition of Glycosidases by Small Artificial Binding Proteins (Affitins). *PLoS One* **2014**, *9* (5), No. e97438.
(27) Orlova, A.; Wållberg, H.; Stone-Elander, S.; Tolmachev, V. On the Selection of a Tracer for PET Imaging of HER2-Expressing Tumors: Direct Comparison of a 124I-Labeled Affibody Molecule and Trastuzumab in a Murine Xenograft Model. *J. Nucl. Med.* **2009**, *50* (3), 417−425.
(28) 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.
(29) Engh, R. A.; Bossemeyer, D. Structural Aspects of Protein Kinase Control—role of Conformational Flexibility. *Pharmacol. Ther.* **2002**, *93* (2−3), 99−111.
(30) Novotný, J.; Bruccoleri, R.; Newell, J.; Murphy, D.; Haber, E.; Karplus, M. Molecular Anatomy of the Antibody Binding Site. *J. Biol. Chem.* **1983**, *258* (23), 14433−14437.
(31) Nord, K.; Nilsson, J.; Nilsson, B.; Uhlén, M.; Nygren, P. A. A Combinatorial Library of an Alpha-Helical Bacterial Receptor Domain. *Protein Eng., Des. Sel.* **1995**, *8* (6), 601−608.
(32) Woldring, D. R.; Holec, P. V.; Stern, L. A.; Du, Y.; Hackel, B. J. A Gradient of Sitewise Diversity Promotes Evolutionary Fitness for Binder Discovery in a Three-Helix Bundle Protein Scaffold. *Biochemistry* **2017**, *56* (11), 1656−1671.
(33) Koide, A.; Wojcik, J.; Gilbreth, R. N.; Hoey, R. J.; 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.
(34) Searle, M. S.; Williams, D. H. The Cost of Conformational Order: Entropy Changes in Molecular Associations. *J. Am. Chem. Soc.* **1992**, *114* (27), 10690−10697.
(35) Cole, C.; Warwicker, J. Side-Chain Conformational Entropy at Protein-Protein Interfaces. *Protein Sci.* **2002**, *11* (12), 2860−2870.
(36) Yu, H.; Yan, Y.; Zhang, C.; Dalby, P. A. Two Strategies to Engineer Flexible Loops for Improved Enzyme Thermostability. *Sci. Rep.* **2017**, *7*, 41212.
(37) Nagarajan, R.; Archana, A.; Thangakani, A. M.; Jemimah, S.; Velmurugan, D.; Gromiha, M. M. PDBparam: Online Resource for Computing Structural Parameters of Proteins. *Bioinf. Biol. Insights* **2016**, *10*, BBI.S38423.
(38) Eyal, E.; Bahar, I. Toward a Molecular Understanding of the Anisotropic Response of Proteins to External Forces: Insights from Elastic Network Models. *Biophys. J.* **2008**, *94* (9), 3424−3435.
(39) Schrödinger, L. L. C.. *The {PyMOL} Molecular Graphics System, Version ∼ 1.8*; 2015.
(40) 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*, W382.
(41) Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; et al. Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* **2017**, *357* (6347), 168−175.
(42) Kowalsky, C. A.; Faber, M. S.; Nath, A.; Dann, H. E.; Kelly, V. W.; Liu, L.; Shanker, P.; Wagner, E. K.; Maynard, J. A.; Chan, C.; et al. Rapid Fine Conformational Epitope Mapping Using

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335

Comprehensive Mutagenesis and Deep Sequencing. *J. Biol. Chem.* **2015**, *290* (44), 26457−26470.
(43) Woldring, D. R.; Holec, P. V.; Hackel, B. J. ScaffoldSeq: Software for Characterization of Directed Evolution Populations. *Proteins: Struct., Funct., Genet.* **2016**, *84* (7), 869−874.
(44) Woldring, D. R.; Holec, P. V.; Zhou, H.; Hackel, B. J. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One* **2015**, *10* (9), No. e0138956.
(45) Aird, D.; Ross, M. G.; Chen, W.-S.; Danielsson, M.; Fennell, T.; Russ, C.; Jaffe, D. B.; Nusbaum, C.; Gnirke, A. Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries. *Genome Biol.* **2011**, *12* (2), R18.
(46) Birtalan, S.; Fisher, R. D.; Sidhu, S. S. The Functional Capacity of the Natural Amino Acids for Molecular Recognition. *Mol. BioSyst.* **2010**, *6* (7), 1186.
(47) Birtalan, S.; Zhang, Y.; Fellouse, F. A.; Shao, L.; Schaefer, G.; Sidhu, S. S. The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies. *J. Mol. Biol.* **2008**, *377* (5), 1518−1528.
(48) Koide, S.; Sidhu, S. S. The Importance of Being Tyrosine: Lessons in Molecular Recognition from Minimalist Synthetic Binding Proteins. *ACS Chem. Biol.* **2009**, *4* (5), 325−334.
(49) Eijsink, V. G. H.; Bjørk, A.; Gåseidnes, S.; Sirevåg, R.; Synstad, B.; Van Den Burg, B.; Vriend, G. Rational Engineering of Enzyme Stability. *J. Biotechnol.* **2004**, *113*, 105−120.
(50) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (15), 5869−5874.
(51) Tokuriki, N.; Stricher, F.; Serrano, L.; Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **2008**, *4* (2), No. e1000002.
(52) Ma, J. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. *Structure* **2005**, *13* (3), 373−380.
(53) Skjaerven, L.; Hollup, S. M.; Reuter, N. Normal Mode Analysis for Proteins. *J. Mol. Struct.: THEOCHEM* **2009**, *898* (1−3), 42−48.
(54) Miller, S.; Janin, J.; Lesk, A. M.; Chothia, C. Interior and Surface of Monomeric Proteins. *J. Mol. Biol.* **1987**, *196* (3), 641−656.
(55) Chao, G.; Lau, W. L.; Hackel, B. J.; Sazinsky, S. L.; Lippow, S. M.; Wittrup, K. D. Isolating and Engineering Human Antibodies Using Yeast Surface Display. *Nat. Protoc.* **2006**, *1* (2), 755−768.
(56) Hood, M. T.; Stachow, C. Influence of Polyethylene Glycol on the Size of Schizosaccharomyces Pombe Electropores. *Appl. Environ. Microbiol.* **1992**, *58* (4), 1201−1206.
(57) Masella, A. P.; Bartram, A. K.; Truszkowski, J. M.; Brown, D. G.; Neufeld, J. D. PANDAseq: Paired-End Assembler for Illumina Sequences. *BMC Bioinf.* **2012**, *13* (1), 31.
(58) Le, Q. V; Karpenko, A.; Ngiam, J.; Ng, A. Y. ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning. *Proceedings of the 24th International Conference on Neural Information Processing Systems* **2011**, 1017−1025.
(59) Jolliffe, I. Principal Component Analysis. In *International Encyclopedia of Statistical Science*; Springer, 2011; pp 1094−1096.
(60) Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **2005**, *67* (2), 301−320.
(61) Edgar, R. C. Search and Clustering Orders of Magnitude Faster than BLAST. *Bioinformatics* **2010**, *26* (19), 2460−2461.

DOI: 10.1021/acscombsci.8b00182
ACS Comb. Sci. 2019, 21, 323−335