Data Driven Approach to Engineering Protein Evolvability and Developability

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

Alexander Wayne Golinski

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

Benjamin J. Hackel & Stefano Martiniani

August 2021

Acknowledgements

Success is often achieved not by what you know, but who you know. My family, pets, friends, advisors, undergraduates, and professional colleagues have enabled my ability to perform this research. I could not have asked for a better support system to allow me to achieve my goals. To those of you I have met and to those of you that continue to support me: thank you.

Abstract

Proteins can be engineered to perform a variety of functions ranging from diagnostics and therapeutics to industrial and commercial enzymes. The ability to computationally evaluate the performance of a protein from its amino acid sequence would increase the efficiency of discovery, expanding the impact of engineered proteins. However, the problem is plagued by the immensity, complexity, and barrenness of the amino acid sequence-function landscape. The following research is focused on predicting two nontraditional protein functions: 1) Evolvability - the ability to generate novel functionality based upon the mutation of a subset of amino acid positions, and 2) Developability - the ability to be efficiently manufactured and maintain primary functionality. Limited prior understanding of these functions was available across broad swaths of sequence space. This work advanced a hybrid experimental/computational platform to provide broad and deep experimental data on sequence-function relationship. Empowered by data analytics, the dataset enabled accurate predictions and provided mechanistic insight regarding protein evolvability and developability. The first story aimed to determine which computable biophysical properties drive evolvability. Utilizing high-throughput screens for evolving specific molecular targeting, the performance of seventeen protein scaffolds were obtained for seven molecular targets. A model predicting evolvability from biophysical properties was trained, with a focus on generalizability and interpretability. Achieving a 4/6 true positive rate, a 9/11 negative predictive value, and a 4/6 positive predictive value, the predictive model analysis suggests a large, disconnected paratope (location of sequence variation) will permit evolved binding function. The second story aimed to generate a model to predict protein developability, as determined by bacterial production, from amino acid sequence. As traditional metrics of developability

are often capacity limited (10^2 - 10^3), a set of three of high-throughput (10^5) assays were created to generate a sufficient dataset. The relevance of the assays to traditional metrics was certified by a model that predicts expression from assay performance 35% closer to the experimental variance and trains 80% more efficiently than a model predicting from sequence information alone. The validated assays offer the ability to identify developable proteins at unprecedented scales, reducing a bottleneck of protein commercialization. Neural networks were trained to generate a numeric developability representation (embedding) for each sequence from the high-throughput dataset and transfer the embedding to predict recombinant expression. Mimicking protein theory, our deep-learning model convolves machine-learned amino acid properties to predict expression 42% closer to the experimental variance compared to a traditional approach. Analysis of trained numeric encodings of the amino acids highlights the unique capability of cysteine, the importance of hydrophobicity and charge, and unimportance of aromaticity when aiming to improve developability of the protein scaffold Gp2. The completion of the studies supports the hypothesis that data-driven protein engineering can both accurately predict protein evolvability and developability while also providing meaningful insight into the properties driving functionality. The success of this approach is predicted to increase significantly as the capacity to parametrize protein function continues to grow. The research presents the increased ability to engineer proteins across their diverse sequence landscape using modern experimental techniques and data analytics.

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

1.1 Protein Diversity Enables Broad Functionality

Proteins are composed of twenty canonical amino acids of various size, charge, and polarity. Anfinsen's dogma states the order and type of amino acids determines the resulting three-dimensional structure, and thus the function, of the protein[1]. While simply stated, the combinatorics leads to an astronomical number of possibilities allowing proteins to serve numerous roles including: proteases (e.g. trypsin found in digestive tracts[2]), lipases (e.g. BSK-L found as an additive in laundry detergent[3]), antibiotics (e.g. antimicrobial peptides found as a promising new avenue in fighting infections[4]), and molecular targeting agents (e.g. antibodies found in immune systems[5,6]). The field of protein engineering has emerged to modify existing proteins, or develop novel proteins, to achieve several goals including: increasing performance (e.g. increasing reaction efficiency and speed[7]), increasing specificity (e.g. limiting the reactivity of undesired of substrates and formation of undesired products[8]), developing novel function (e.g. activating a pathway to invoke self-destruction of a cancerous cell[9]), or increasing robustness (e.g. improving the self-life of an antibody used to activate the immune system for various treatments[10]). Thus, the ability to predict the function of a protein from the amino acid sequence will enable universal advancements.

1.2 Protein Diversity Also Complicates Search for Optimal Sequences

It is estimated that the average protein length ranges from 250 residues (archaebacteria) to 450 residues (eukaryotes)[11]. Assuming every amino acid can be substituted at every position, this results in 20^{250} - 20^{450} (10^{325} - 10^{585}) unique combinations. Put into perspective, it is currently approximated that there are only 10^{11} stars in the Milky Way. Not only is it impractical to experimentally produce and assay every protein's

function, even if it were possible to computationally predict function in a single floating-point operation (FLOP), the world's current fastest super computer (IBM's Summit[12]) of 200 petaFLOPs ($2x10^{17}$ FLOPs) per second would take approximately 10^{430} years to compute all possibilities.

Making the situation more complicated, it is believed that the landscape describing the relationship between protein sequence and function is barren (most sequences lack desired function) and rugged (a single substitution could drastically modify function)[13]. This is likely due to a network of residue-residue interactions created when the protein folds, resulting in a highly complex dynamical system. Thus, it is vital to develop efficient experimental techniques and computational models to navigate protein candidates.

1.3 Rational Design Enables Protein Engineering with Limitations

A common first approach at modifying a protein's sequence is to design mutations utilizing some background knowledge of the protein and/or its function[14]. The most useful piece of information is often a three-dimensional representation (obtained by X-ray crystallography, nuclear magnetic resonance[15], or electron microscopy[16] experiments) of the protein in complex with the target or analogue that can allow visualization of potential mutations. As an example, Wells et al.[17] improved the catalytic efficiency and specificity of subtilisin though targeted substitutions of charged amino acids after noticing potential ion pairs between the enzyme and substrate from a crystal structure. This process is often limited by the nontrivial ability to obtain accurate representations[18], as well as the inability to fully predict how a substitution will affect other nearby residues. That being said, there has been success in creating ultra-stable proteins of various geometries through the use of rational design[19], commenting on the ability to design more complex properties.

1.4 Random Mutagenesis Paired with Directed Evolution Expands Possibilities

A more recent approach attempts to increase protein functionality by building in a feedback mechanism to select desired proteins and continue the search within a smaller vicinity. Championed by Frances Arnold[20] and utilized by many, directed evolution has shown promise in improving the function of numerous proteins spanning from biocatalysts for biofuels[21] to photoresponsive peptide ligands[22]. Direction evolution is comprised of three steps: 1) creating a library of protein variants though processes such as error-prone PCR or degenerate codon synthesis, 2) assaying the library for a desired function, and 3) isolating and amplifying the DNA encoding for the beneficial mutations. Depending on the performance of the isolated variants, this process can be repeated until the desired level of function is achieved. Compared to rational design, directed evolution does not require knowledge of protein structure or the mechanism of interactions. However, this process is heavily based upon the mutation strategy and can often result in local solutions rather than the ideal global solution[23]. This has led to the creation of more guided mutation strategies, including domain swapping and walking techniques[24–27]. But the benefit of not requiring knowledge for random mutation is also a limitation via the lack of ability to apply known information.

1.5 Mechanistic Models Offer Limited Success with Poor Scalability

The next approach to determine protein structure and function is to use first principal forces to calculate the most stable configuration. Software packages such as FoldX[28] and Rosetta[29,30] calculate properties such as electrostatic forces, entropic effects, and solvent interactions to determine the stability (ΔG) of the folded protein compared to the unfolded amino acid chain. However, to accurately calculate these properties, the packages must first find the most stable conformation of the backbone and the rotamer

position of the residue side chains. This task has a large geometric domain that often requires rounds of iteration and refinement leading to supercomputer-level requirements. Another approach is to use a homologue, or similar protein based upon sequence, to obtain an initial structure and then calculate the relative change in stability ($\Delta\Delta G$) by mutating positions of interest. Unfortunately, this approach has similar limitations as rational design by requiring a structure of the homologue and is limited in accuracy with increasing distance from the starting point. It is also hypothesized these software contain training/validation bias due to datasets containing unbalanced data largely in favor of destabilizing mutations[31].

1.6 Protein Evolvability and Developability Are Uniquely Difficult Functions to Engineer

The most common applications of the previously mentioned protein engineering strategies are to improve stability (specifically thermostability) and interaction strength (specifically binding affinity and catalytic power). These functions can often be rationalized by looking at the specific inter- or intramolecular interactions, possibility leading to higher success rates. However, not all protein functions are as straightforward. The work in this dissertation focuses on more abstract functionality including evolvability (the ability to create new function upon mutation of sequence) and developability (the ease of manufacturing and maintenance of function).

In nature, proteins must perform the desired function and possess the ability to adapt new functionality to match changing stimulus and environment. Regarding engineered proteins, the ability to switch specificity (either binding or catalytic) by mutating a relatively small part of the amino acid sequence is commonly referred to as evolvability[32]. The field distinguishes *innovability* and *evolvability* as the ability to

generate new function and the ability to modulate existing function, respectively. Within this thesis, we use *evolvability* to refer to the ability to modulate function, without distinction on functional novelty. A common approach has been protein scaffolds, where it is hypothesized that the conserved (unmodified) portion of the protein will aid the identification of a useful sequence by providing a stabilizing region and providing a scaffolding (shape) that has been known to be functional in other applications[33–38]. It has been shown that a tradeoff exists between the rigidity of the backbone that provides stability and the flexibility of backbone that permits numerous mutated regions that may be required to obtain a desired function[32,39,40]. This nontrivial tradeoff is the focus of Chapter 2 of this dissertation. The increased stability of a parental protein often increases success[41–43], due to most mutations being destabilizing[44,45]. However, as protein scaffolds exist is many sizes and shapes[46], it is nonobvious what biophysical properties aid in maximizing evolvability. Understanding the properties of proteins that are correlated to evolvability could allow for the selection of more ideal scaffold rigidities easing the engineering effort to achieving a desirable functionality.

Developability is an often overlooked property that describes the general ease of production, storage, and use of a protein without significant degradation[47–49]. While the previously mentioned stability is part of the developability equation, other properties such as solubility and production yield involve hard to predict interactions with the solvent and other cellular proteins. A study by Jain exemplified the importance of developability by noting a correlation between poor developability properties and the ability of the molecule to pass clinical trials.[48] Accordingly, a slew of experimental techniques (e.g. AC-SINS[50,51], and hydrophobic interaction chromatography[52,53]) and computational tools (Therapeutic

Antibody Profiler[49], CamSol[54], and Developability Index[55]) have been popularized. It is known that developability is best characterized by several properties[56], and computational tools have mixed success rates[57]. This function is particularly challenging because most assays have limited throughput (100's), which limits the confidence of any observed trends. Thus, it remains unknown if limited predictive capability is due to extreme complexity or due to limited ability to train a sufficient model. In this work, Chapter 3 is focused on identifying assays which increase assessment capacity while Chapter 4 is focused on utilizing the increased dataset to identify factors driving developability. The ability to engineer protein developability could remove a major hurdle in the commercialization pipeline and even increase the efficiency of the candidate selection process[58].

1.7 Data Science Aims to Identify Driving Properties

It is often difficult to determine which amino acid and protein properties are related to the function of interest, particularly with evolvability and developability. The field of data science has developed several tools to answer such questions including machine learning - which is aimed to narrow down a list of potential properties, and deep learning - which is aimed to teach itself important properties from the data[59]. Compared to the previously mentioned methods of protein engineering, the data science approach does not require as much structural information, sophisticated simulations, nor rely solely on random chance to discover beneficial mutations. Instead, this method of engineering requires a vast dataset which is then used to train a predictive model. The driving factors for function are then determined by analyzing model parameters and by analyzing predicted protein variants. For example, Alley and team trained a Unified Representation (UniRep) of natural proteins[60]. They assessed the trained parameters that directly

transformed each amino acid to a numeric vector and found residues located near each other with similar properties of charge and aromaticity - suggesting the importance of those properties. They also compared the numeric representation of proteins from various species and found that UniRep considered taxological data important when trying to predict protein sequences. While training of the models from data can be time consuming, the evaluation of proposed variants is often much faster than physical-based approaches. Data science has been applied to protein engineering to solve a myriad of functions including enzyme productivity and thermostability[61,62]. One major drawback in data driven studies is the data: as the utility of predictions and interpretations can only be as accurate and relevant as the input. Thus, the quality and quantity of the input dataset must always be questioned.

1.8 Success of Data Science Approaches Require Relevant and Sufficient Data

The leading limitation to applying machine learning and deep learning to protein engineering is the lack of useful training data. Depending on the type of question asked, models may need to fit 10^1 to 10^5 parameters to achieve sufficient accuracy. While some machine learning techniques are able to generalized when overparameterized[63], its generally believed the current literature of protein functionalities lack sufficient depth. One strategy to overcome this limitation is to use databases of alternative functions for training, such as a list of proteins found in nature or in the Protein Data Bank[64]. It can be hypothesized that the natural proteins must all possess sufficient developability to have been discovered and training a model on such information will allow sufficient accuracy in the prediction of the developability metric. Though this strategy has shown success in predicting functions related to requirements to be useful in nature like stability[60,65], it remains unclear if these training sets are relevant to evolvability and other metrics of

developability. Biswas and team was recently able to show the success of low-N protein engineering, but commented on the requirement of existing high-fidelity and throughput assays to measure protein fuction[66].

1.9 Contributions of Dissertation

This dissertation addresses the hypothesis that if a relevant dataset of protein evolvability and developability is obtained, then data science approaches can be used to predict and interpret critical factors of the respective functions.

1.9.1 Aim 1: Interpreting and Predicting Protein Evolvability

In Chapter 2, a series of seventeen protein scaffolds varying in twenty biophysical properties were tested in their ability to evolve binding functionality. Using existing yeast display technology[67], each scaffold was scored by the ability to create unique binders towards seven protein targets. A model was created, with appropriate techniques to limit redundancy and improve generalization, which revealed the importance of a large, spatially-independent paratope. We were then able to accurately assess an additional 700 potential scaffolds for evolvability potential. The completion of this aim provides evidence focused high-throughput experiments paired with machine learning techniques can be employed on engineering protein evolvability within the context of protein scaffolds.

1.9.2 Aim 2: Interpreting and Predicting Protein Developability

We next focused on creating an amino-acid based model to predict protein developability for paratope variants of a specific protein scaffold. As previously mentioned, the large protein combinatorics requires a highly-parameterized model which requires a sufficiently sized dataset to train. As protein developability lacked assays capable of reasonably scaling beyond 10^2 - 10^3, Chapter 3 develops and validates three

high-throughput (10^5) developability assays and creates a database of sequence-based developability information. These assays (an on-yeast protease assay, a split GFP assay, and a split β-lactamase assay) represent progress towards eliminating the barrier of variant developability quantification in the protein commercialization pipeline. Chapter 4 then determines if the high-throughput assay information can train a model capable of predicting a traditional developability metric from amino acid sequence that can also provide mechanistic insight. Through analysis of amino acid and protein numeric representations, the data demonstrate the unique impact of cysteine, which can form a stabilizing (and developability increasing) disulfide bond. Through analysis of conformational landscape exploration, we find a region of developability with a highly rugged landscape containing unique properties suggesting the existence of numerous beneficial sequence motifs. The completion of this aim provides both new methodology for developability analysis and evidence that deep-learning models can be trained for protein scaffold developability.

Chapter 2 - Biophysical Characterization Platform Informs Protein Scaffold Evolvability

Adapted from "Alexander W. Golinski, Patrick V. Holec, Katelynn M. Mischler, and Benjamin J Hackel. 'Biophysical Characterization Platform Informs Protein Scaffold Evolvability.' ACS Comb. Sci. 2019, 21, 323−335."

2.1 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 which permit binding evolution remain unknown. We present an algorithm to predict a protein scaffold's ability to obtain novel binding function based upon computationally calculated biophysical parameters. The ability of seventeen 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 feature 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 *E. 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 which possess binding functionality over non-functional mutants. Future iterations of protein scaffold discovery and evolution will benefit from a combination of computational prediction and knowledge of initial developability properties.

2.2 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[13,68], discovery of completely new function typically requires greater leaps in sequence[39]. Given the relative barrenness and tortuosity of sequence space[13], efficient strategies are needed to achieve successful *de novo* discovery. One strategy to facilitate discovery is the use of a protein scaffold[33,37] 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[69] and diagnostics[34]. A functional protein ligand scaffold must be able to create new, specific binding function upon mutation of the paratope[70] and possess optimal developability properties (e.g. stability, solubility, and expression) for downstream use.[48] To date, numerous protein scaffolds have been engineered to obtain strong affinity towards clinically relevant targets[71,72], while some have entered clinical trials.[73–76] Protein scaffolds offer novel topologies and differential size, allowing for unique binding interfaces and tunable pharmacokinetic properties.[46,77] The diversity of topologies and physicochemistries of published scaffolds, and the paucity of

data on unsuccessful scaffolds, preclude an understanding of the biophysical features which 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 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[69], antibody fragments[35], and leucine-rich repeats[78] presumed that their natural function for high affinity binding will serve as a starting point for scaffold engineering. Fibronectin type III 'monobodies'[79] and designed ankyrin repeat proteins[80] are structurally similar to these immune scaffolds. Lipocalins[81], three-helix bundle affibodies[82], fynomers[83], and others[46] offer unique topologies with native binding ability. Alternatively, multiple scaffolds are chosen for their strong structural stability including cystine knots[84] and thermophilic affitins[85] and homologs. Similarly, a host of other scaffolds have provided compelling performance in ligand development while others have been tested without the same level of success.[80] 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 stability tolerance.[71] 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 selection

techniques are then employed to assess predictive characteristics of functional scaffolds. In this paper, 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 tradeoff 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.

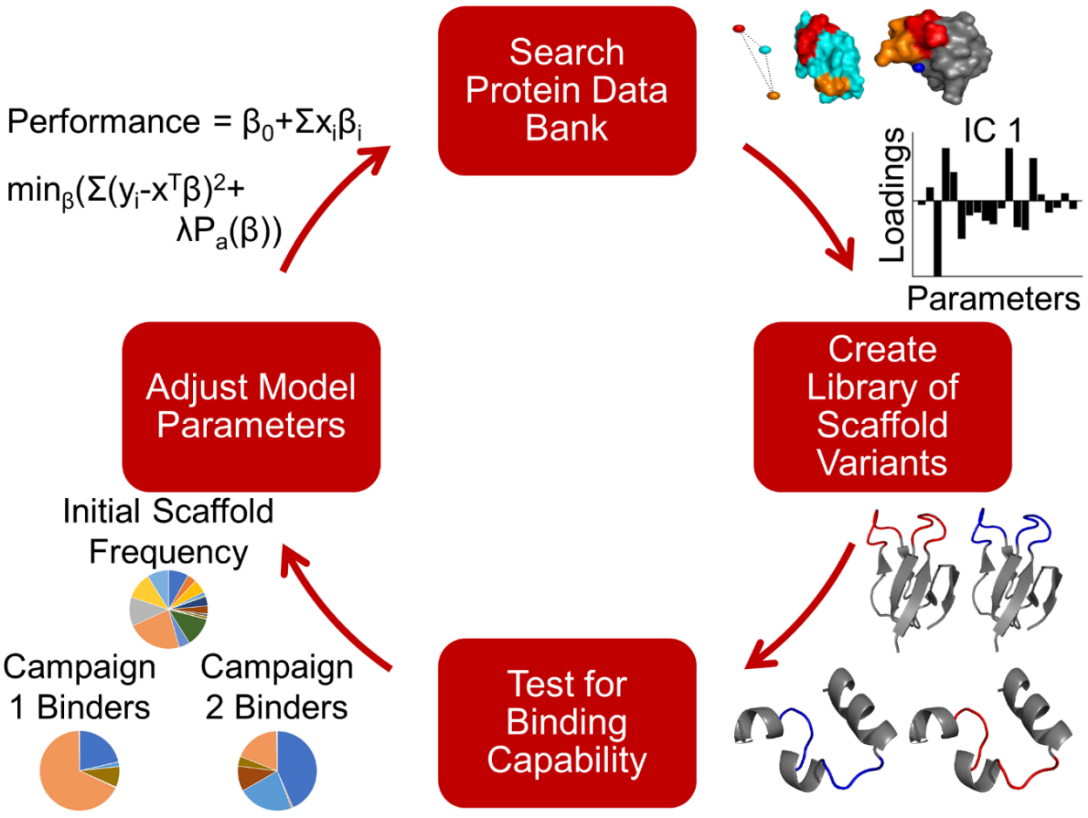

Figure 2.1 - Algorithm for protein scaffold discovery

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.

2.3 Results and Discussion

2.3.1 Computational Scaffold Analysis

We hypothesize that not all proteins possess the characteristics to robustly and efficiently develop 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.[86] Small, single-domain architecture eases fusion and site-specific conjugation for multifunctional constructs. Small size reduces exposed surface area that may lead to undesired non-specific interactions. Moreover, small size heightens the challenge to simultaneously balance evolution of intramolecular stability and intermolecular binding[87,88], 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[79,89], α-helices[42,90], β-strands[91], and mixed topologies[80,81]. Although the impact of entropic cost upon binding[92,93] – 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.[94] 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 (1) filtered for size (30-65 AA pre-trimming) and the presence of two loops with at least four residues. 787 unique protein scaffolds were (2) demarcated into conserved frameworks and diversifiable paratopes and (3) characterized by twenty parameters describing geometrical, chemical, and stability properties (summarized in Table 1 and the following text and described in depth in *Materials and Methods*). *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[95]. *Paratope Connectivity.* Paratope connectivity and flexibility, the latter via normal mode analysis[96], was also calculated as we believed spatially-removed diversifications will be less destabilizing to the remainder of the protein. *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[97] was used to model the protein surface and calculate the chemical nature of the solvent accessible surface area (SASA). *Paratope Size and Topology.* Paratope orientation was parameterized 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. *Computational Stability.* It is proposed that scaffolds must be stable and have mutational stability to maintain structural integrity when obtaining binding function. The FoldX empirical forcefield was used to estimate mutational destabilization and overall stability.[28] The amount of buried non-polar surface area was also estimated as a relationship with stability was recently observed for small proteins.[98] *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. 17 candidate scaffolds (Figure 2A-Q), which provide a range of characteristics (Figure 2R), were chosen for experimental evaluation.

Table 2.1 - Evaluated descriptors of protein scaffolds

Factor	Description	Mean±SD (n=787)
Protein Connectivity		
Contact Degree	Total number of residue contacts within 8Å	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 8Å 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	The 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^o
Paratope SASA	The solvent exposed surface area of an alanine-scanned paratope region	780±360 Å^2
Paratope Separation	The 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±0.4AU
Computational Stability		
Buried NPSA	The amount of buried non-polar 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	The amount of solvent exposed area created when removing unstructured termini	320±260 Å^2
Secondary Structure Percent	The percent of residues in an α-helix or β-sheet	51±12 %
Size	The total number of residues in the scaffold	47±7 AA

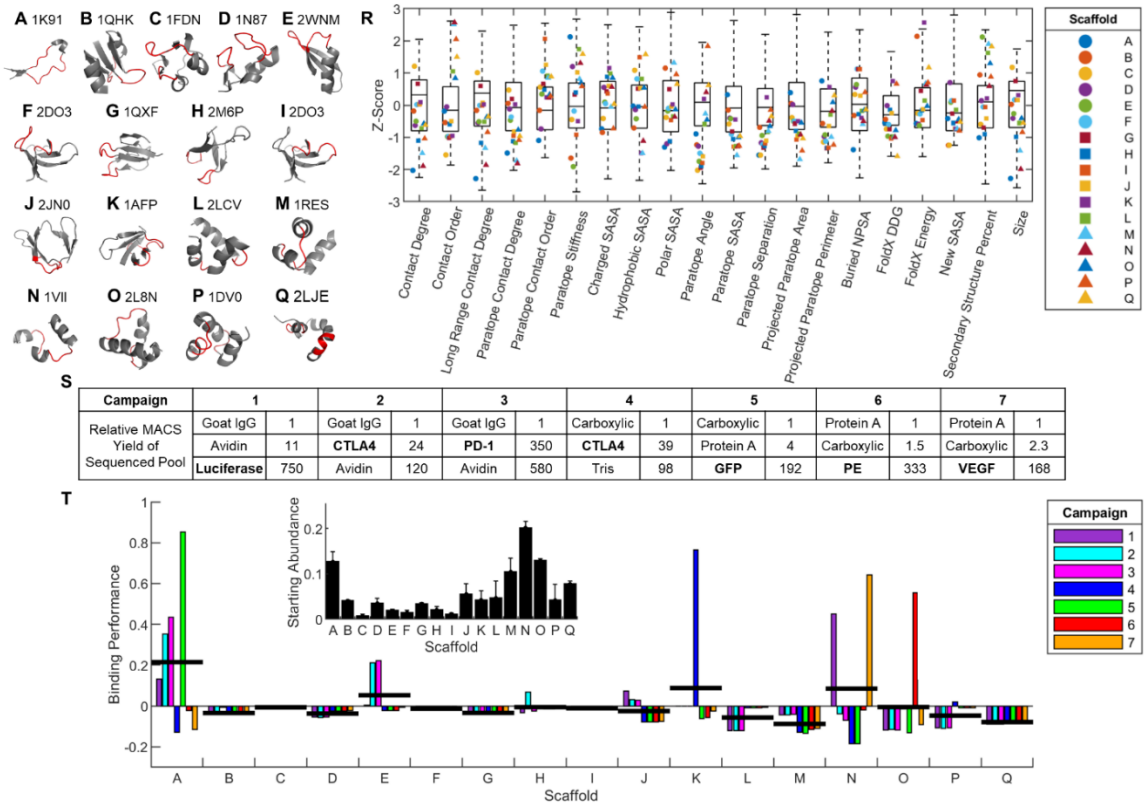

Figure 2.2 - Protein scaffold candidates show varying binding performance

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: 25th – 75th percentile; center bar: median; whiskers: 1.5 x interquartile range). The plotted values for each of the 17 assayed scaffolds indicate a diversity of proteins were assayed. **S.** A pooled sample of 1x1010 variants across 17 scaffolds was enriched for binding variants in seven campaigns. MACS sorting was performed until seven binding populations were identified towards 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).

2.3.2 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 3-9x10^8 variants per scaffold. The 17 scaffold libraries were mixed resulting in a total diversity of 1x10^{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 non-specific 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 3,500-715,000 per campaign. Enriched populations exhibited selective binding (Figure 2S) and were deep sequenced to characterize scaffold variants. 280,000 (range 1,250-115,000 per campaign) full-length reads were obtained yielding 21,000 (range 160-9,000 per campaign) unique binding variants. Individual campaign sorting and sequencing statistics are summarized on Table S1. With oversampled sorting, enrichment is correlated with binding affinity.[99] 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[100]) 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 (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, 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[71], 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.

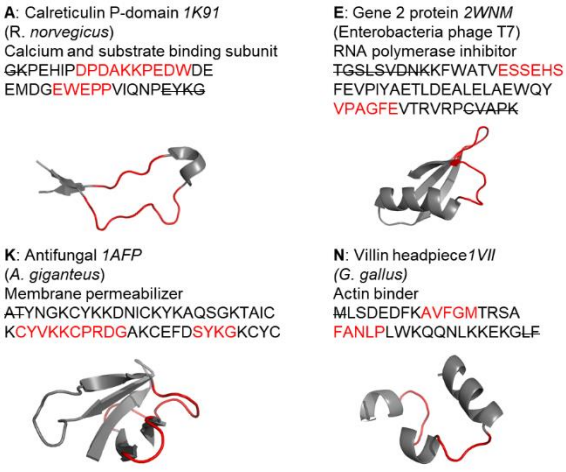

Figure 2.3 - Successful protein scaffolds have diverse topologies

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).

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 over-perform their overall evolvability for multiple reasons. The diversified sites may not be optimal as evolution can be aided by conservation of loop sites[101] and diversification of sites with secondary structure adjacent to paratope[42,101]. Full amino acid

diversity is not optimal for evolution at many sites.[42,101]Yet the 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[102]; 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.

2.3.3 Identifying Functional 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 dataset 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 factors 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 (Figure 4A & 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. 9 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

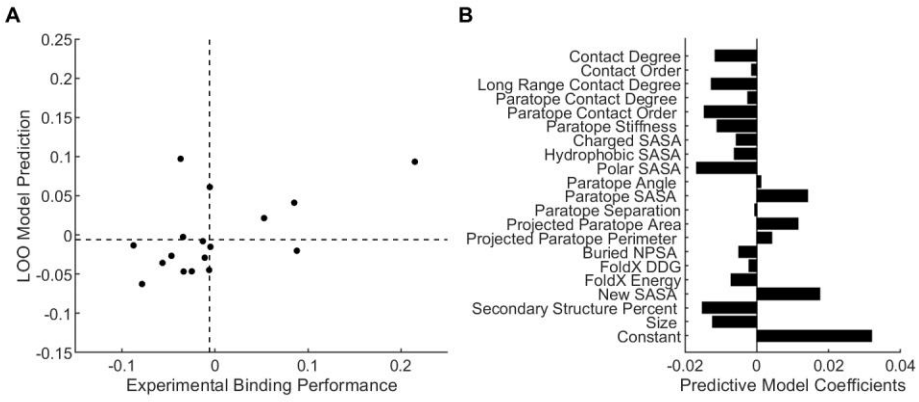

Figure 2.4 - Large disconnected paratopes are associated with increased binding performance

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 twenty calculated parameters and a constant term. The coefficients describe which parameters to modify to improve binding performance of a small protein scaffold.

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. Based upon the linear model term coefficients, the predicted model suggests generally decreasing scaffold connectivity, paratope connectivity, conserved exposed surface area, buried non-polar 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 utilizing 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.

2.3.4 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. Based upon 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 developing 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.

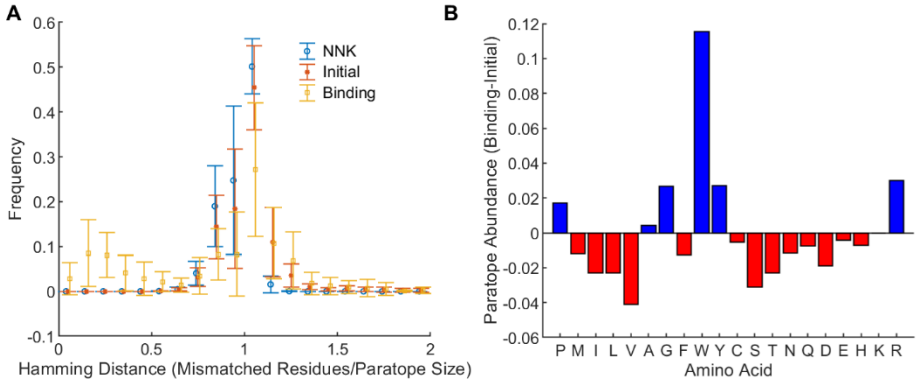

Figure 2.5 - Binding variants describe functional amino acid space

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.

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 due to the ability to partake in different bonds including pi-stacking, hydrogen-bonding, and cation-pi interactions.[103–105] Arginine, which often serves as a hot-spot residue for key interactions but has also been previously associated with non-specific interactions, increased by 3%.[103–105] Glycine increased abundance by 3% perhaps by adding flexibility to the loop regions.[106] Proline increased in abundance by 2%, perhaps by improving scaffold stability by reducing the conformational entropy of the unfolded state.[106] Interestingly, serine has previously shown to be upregulated in binding variants, but was greatly reduced in this study.[103–105] The raw abundance for each residue in the various sequencing populations is depicted in Supplemental Figure 7.

2.3.5 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 *E. 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 producible clones but did not recover any poorly produced proteins.

		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 2.6 - 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 # produced / # attempted.

Based upon 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 development. Additionally, the data recommend that protein scaffolds should be derived from highly developable proteins, rather than engineering developable parameters post-identification of binding functionality. Interestingly, the ability of a parental clone to produce was not indicative of variant producibility (p = 0.3).

2.3.6 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 1,300 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.[98]

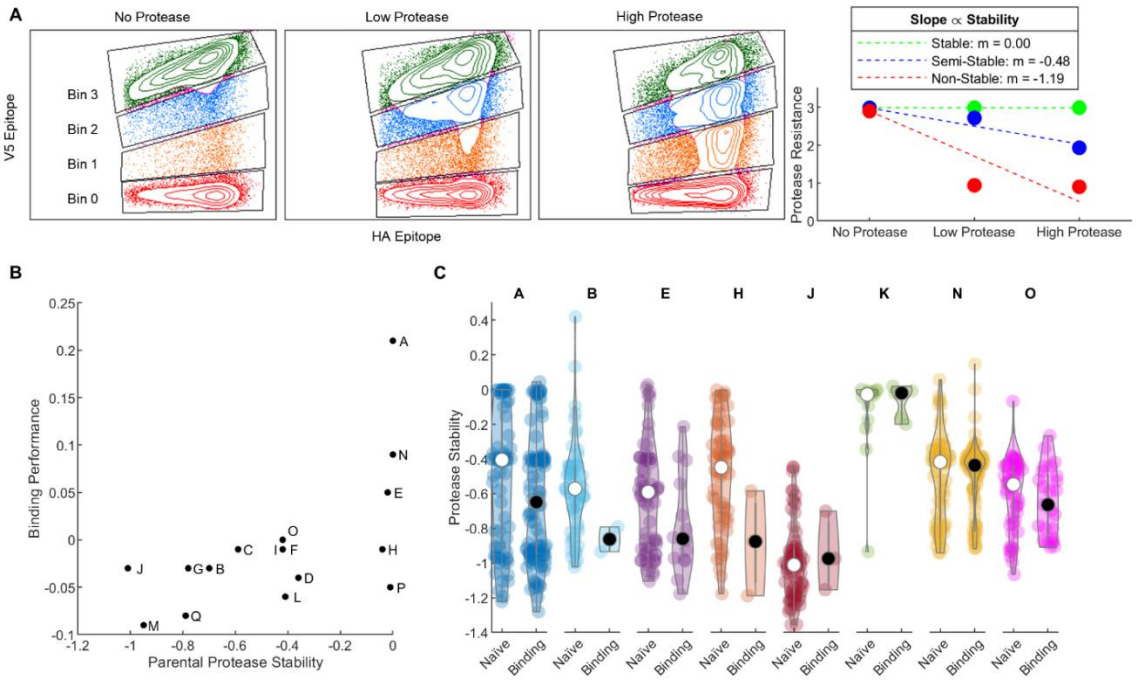

Figure 2.7- Proteolytic stability assay identifies stability requirement for binding

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 ($p = 0.034$).

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 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 non-binding 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.[39,43]

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.

2.4 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 anti-microbial peptide activity. The current limitation in such studies is the collection of a sufficiently rich dataset to build a robust computational model.

2.5 Experimental Procedures

2.5.1 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 outer most 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 ≤ 8 Å are considered contacts, consistent with ranges found in literature.[95] Three parameters are calculated: (1) contact degree: the total number of contacts;

$$Contact\ Degree = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \begin{cases} 1 & \left\| AA_i, AA_j \right\|_2 \leq 8\text{Å} \\ 0 & else \end{cases}$$

(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;

$$Contact\ Order = \frac{\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \begin{cases} j-i & \left\| AA_i, AA_j \right\|_2 \leq 8\text{Å} \\ 0 & else \end{cases} \right)}{N * Contact\ Degree}$$

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.

$$Long\ Range\ Contact\ Degree = \frac{\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \begin{cases} 1 & \left\| AA_i, AA_j \right\|_2 \leq 8\ \text{Å}\ \&\ \text{j} - \text{i} > 12 \\ 0 & else \end{cases} \right)}{N}$$

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[107,108] 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;

$$Paratope\ Contact\ Degree = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \begin{cases} 1 & \left\| AA_i, AA_j \right\|_2 \leq 8\text{Å}\ \&\ \text{AA}_\text{i} \oplus AA_j \in paratope \\ 0 & else \end{cases}$$

(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;

$$Paratope\ Contact\ Order = \frac{\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\begin{cases} j-i & \left\|AA_i, AA_j\right\|_2 \leq 8\text{Å}\ \&\ \text{AA}_\text{i} \oplus AA_j \in paratope \\ 0 & else \end{cases}\right)}{Size\ of\ Paratope * Paratope\ Contact\ Degree}$$

(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[96] - 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 destabilizing due to 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 towards 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; (14) projected paratope perimeter: the number of paratope pixels boarded 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[43] and sought to computationally estimate stability based upon existing correlations. Three parameters were calculated: (15) buried non-polar surface area (buried NPSA)[98]: the sum of solvent exposed non-polar amino acids in Gly-X-Gly[109] minus the sum of solvent exposed non-polar amino acids in the fold protein; (16) FoldX DDG: the mean difference in force field energy between mutant and parental variants; (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[28], 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; (20) size: the number of residues in the scaffold after removal of non-secondary structured termini.

2.5.2 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. 24 scaffolds were chosen to balance breadth of parental proteins and experimentally achievable depth of scaffold variants. 7 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, 1X 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 NdeI, PstI-HF, and BamHI-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 *S. cerevisiae* yeast (EBY100) via homologous recombination. Transformation followed previously described protocols[110], with the addition of 30% v/v PEG 8000 in step 39, which was found to increase transformation efficacy.[111] 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[110], and 10x the transformed diversity of each sub-library 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 hours at 4^{o}C and placed on a magnet for 5 minutes 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, non-binding 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 PBSA (1X phosphate buffed saline with 1g/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.

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 towards luciferase, GFP, PE, and VEGF121, respectively. Campaigns 2, 3, and 4 isolated binders towards CTLA4/Avidin, PD 1/Avidin, and CTLA4/Tris. Though binding was observed towards two molecules, the specificity over a third negative target signifies an enriched population with binding functionality. For avidin-based sorts: 10 µL beads were mixed with 5 or 10 µL of 3.3 µM target in 100 µL PBSA; beads were rotated at room temperature for 1 hour, isolated via magnet, aspirated, and washed with 1 mL PBSA before cells were added to the tube. For carboxylic acid sorts: manufacture's two-

step coating protocol (without NHS) was followed except the following modification: 2 µL of beads were used for each target to match total beads to avidin sorts.

2.5.3 Evaluation of Binder Performance via Deep Sequencing

DNA encoding for scaffolds was isolated from yeast using zymolyase (Zymo Research). Briefly, $1x10^8$ cells are incubated in 200 µL lysis solution (50 mM phosphate buffer, 1M sorbitol, 10 mM β–mercaptoethanol, 75 U/mL zymolyase longlife) for 30 minutes at 37ºC 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[112] 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.[100] 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.

$$Abundance\ (scaffold\ X) = \frac{\sum_{Unique\ Sequence\ i=1}^{Sequences\ for\ scaffold\ X} Reads\ of\ Sequence_i^{\frac{1}{4}}}{\sum_{Unique\ Sequence\ i=1}^{Sequences\ for\ All\ Scaffolds} Reads\ of\ Sequence_i^{\frac{1}{4}}}$$

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 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 (Figure S1 & S2).

2.5.4 Evolutionary Model

With more calculated parameters than experimental datapoints (*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 dataset into mutually independent latent variables.[113] 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.[114] 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.[115] 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 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.

2.5.5 Protein Production

Genes encoding for observed and parental scaffold variants were obtained from Twist BioScience. 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 37ºC for 2 hours 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, 5mM CHAPS, 25 mM imidazole, 2 mM $MgCl_2$, 20 mM NaCl, 7 U/uL benzonase, 50

mg/mL lysozyme, EDTA-free protease inhibitor, and 5% v/v glycerol) and incubated at 37ºC 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 (20mM 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 ºC, washed 4x 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 hour at room temperature, washed 4x, treated with 100 µL of 3,3',5,5'-tetramethylbenzidine (TMB) for 15 minutes, 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 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.

2.5.6 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). $1x10^6$ yeast induced to display protein were incubated in 50 µL PBSA with 0, $4x10^{-6}$, or $22x10^{-6}$ U/µL proteinase K at 37ºC for 10 minutes, 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 anti-chicken IgY (H+L) (Thermo Fisher Scientific) and AlexaFluor647-conjugated goat anti-mouse IgG (H+L) (Thermo Fisher Scientific). Labeling was performed as follows: $1x10^6$ cells were rotated for 30 minutes at room temperature in 50 µL of PBSA with 1 ng/µL primary antibodies, pelleted at 8000*g* for 1 minute, aspirated, washed with 1 mL PBSA, incubated for 20 minutes at 4ºC in 50 µL PBSA with 1 ng/µL secondary antibody; pelleted, washed, and resuspended at $2x10^7$ 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[116] 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).

2.3 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.

2.4 Supplemental Information

Table S2.1 - Sorting and Sequencing Summary

Campaign	1	2	3	4	5	6	7
Maximum Diversity by Sorting Yield	140,000	5,000	3,500	3,900	715,000	220,000	7,500
# of Reads Obtained	64,542	115,519	45,359	1,381	51,367	1,251	6,137
# of Unique Sequences Obtained	4,012	8,969	4,864	214	2,403	162	431
Most Abundant Sequence Reads	34,630	27,611	7,140	327	18,460	775	4,242

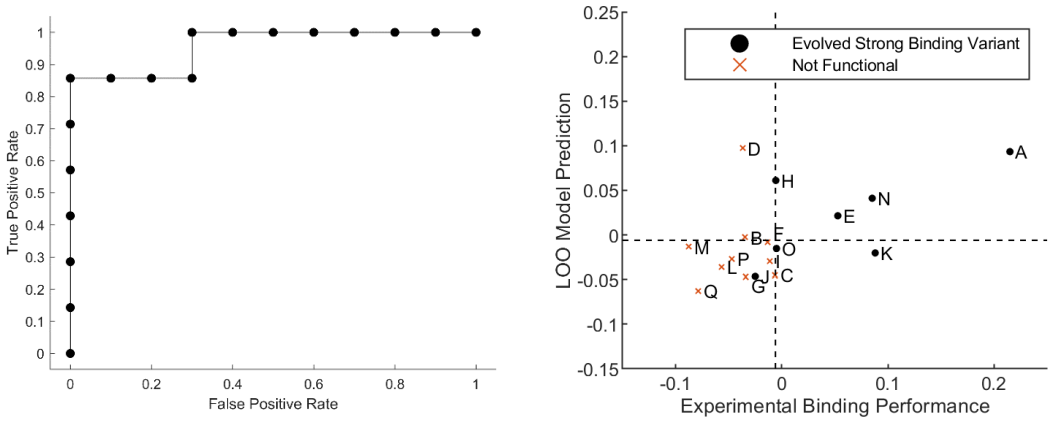

Figure S2.1 - Calibration of Binding Performance

Protein scaffolds were classified by the ability to produce a strong binding variant (campaign abundance >1%, see Supplemental Figure 2). ***Left.*** The ROC curve used to determine a threshold value of binding performance (difference between mean abundance in binding populations and abundance in initial population). A value of -0.006 was chosen to optimize true discoveries and minimize false discoveries. ***Right.*** The ability for the predictive model to correctly classify protein scaffolds also utilized this definition. Scaffold functional classifications are as described above.

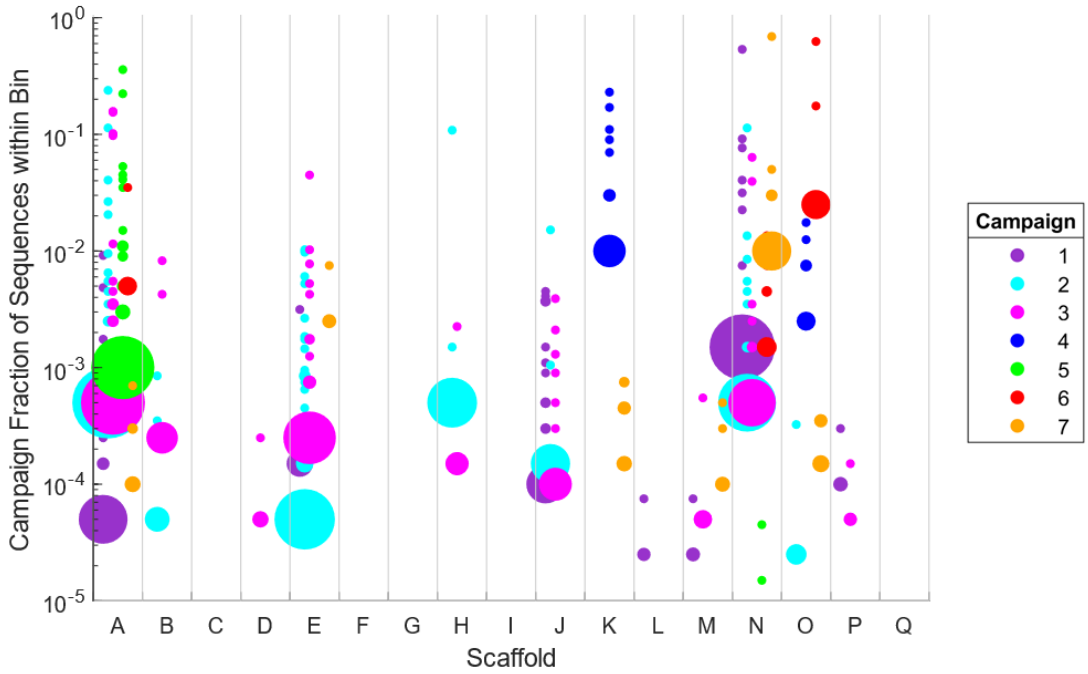

Figure S2.2 - Bubble plot of scaffold performance against each molecular target

Sequences were binned by abundance within campaign, which each bin represented a bubble. The number of unique sequences represented per bin is proportional to the bubble area[2]. Histogram bins were calculated via MATLAB's automated algorithm to best display the distribution of the data.

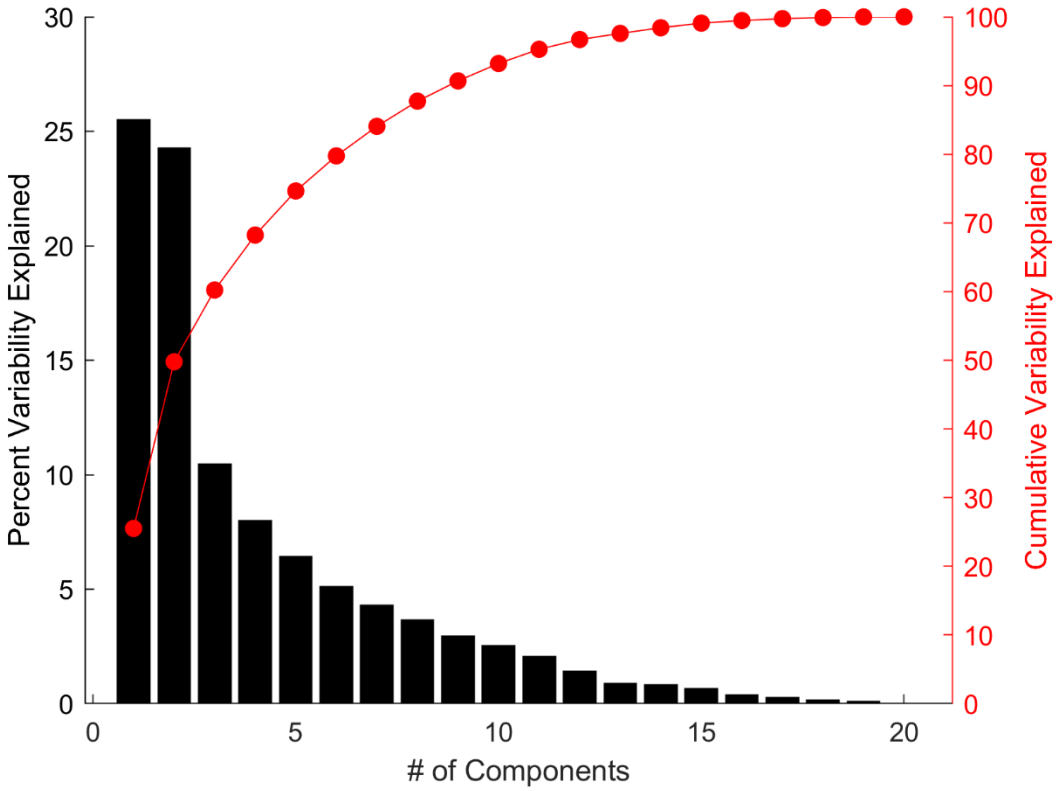

Figure S2.3 - Principal component analysis

The variability of the 20 calculated scaffold factors for 787 proteins were described by principal components. 6 components were included in further analysis as they individually described at least 5% of variability individually with a sum of 80% variability.

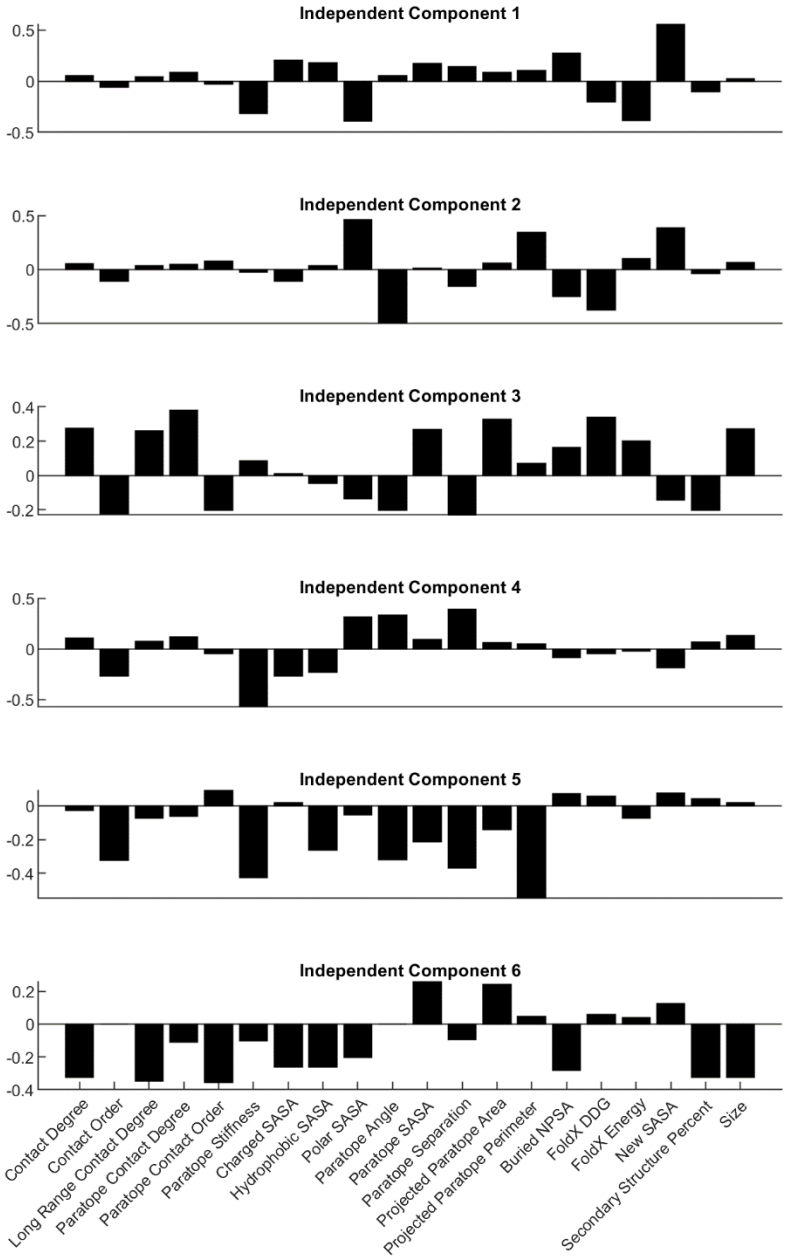

Figure S2.4 - Independent component analysis

Six reconstructed independent components (IC) were obtained via MATLAB's *rica* function following whitening via principal component analysis. After regularization, ICs 1 and 6 were found to be correlated with scaffold binding performance.

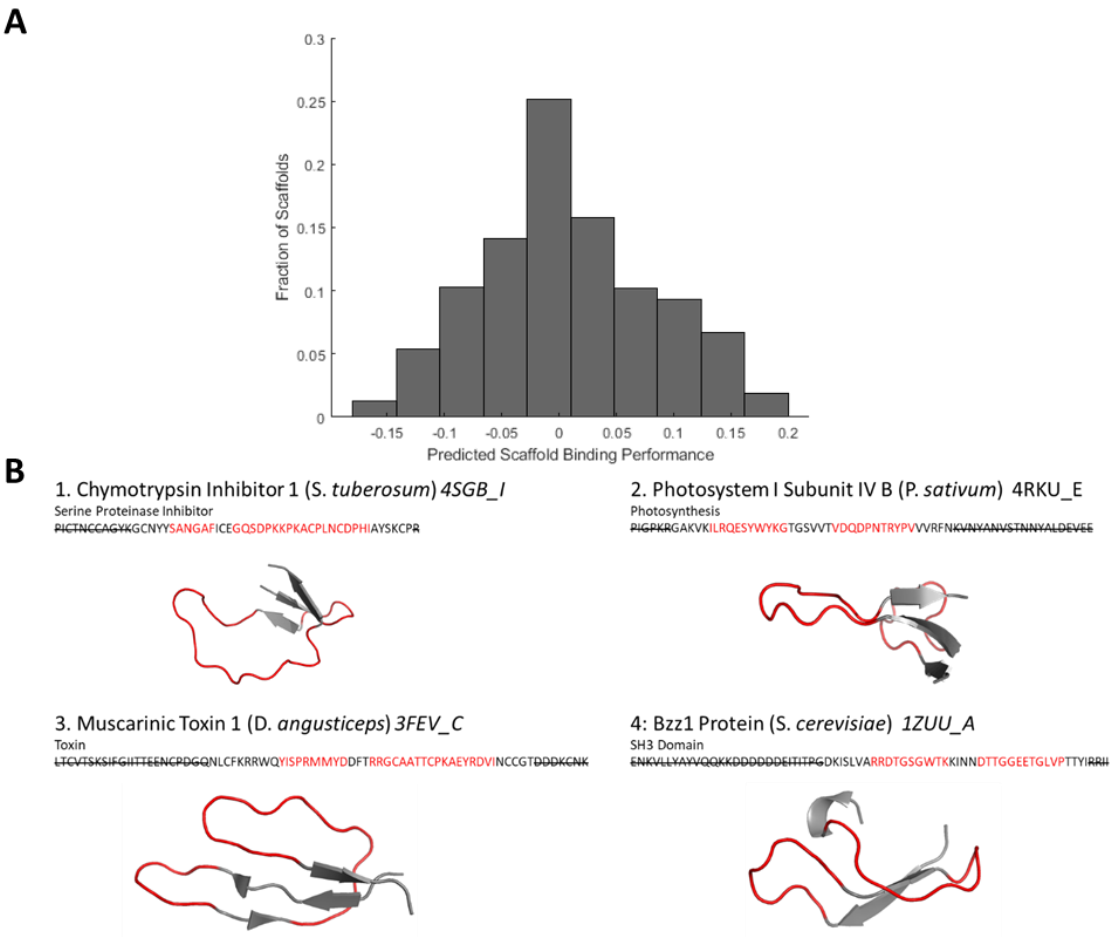

Figure S2.5 - Predicted scaffold performance

A. A histogram of all potential scaffold performance of the 787 calculated scaffolds. Note: as binding performance was a comparative metric, the predicted binding performance simply describes relative scaffold binding performance. ***B.*** The top predicted scaffolds are displayed with large paratope regions. A full list of scaffold parameters can be found on GitHub (HackelLab-UMN).

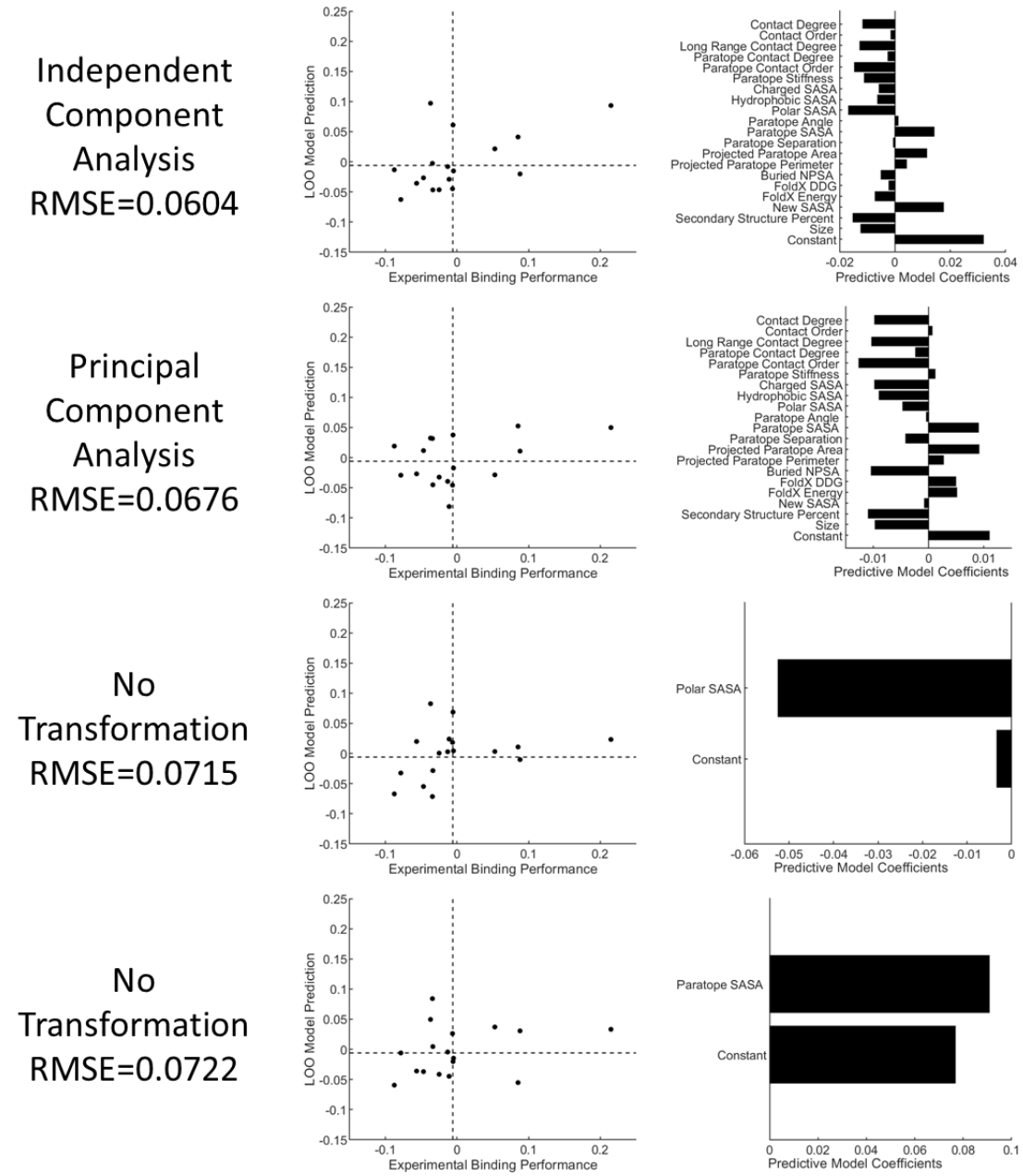

Figure S2.6 - Alternative predictive models

The predictive performance and biophysical parameters are displayed for various approaches. For principal component analysis, the 6 components were fed into elastic net regularization, yielding a single principal component which predicted binding performance. Additionally, each of the individual parameters were used to predict performance. The top two predictive properties included a minimization of conserved polar SASA or a maximization of paratope SASA.

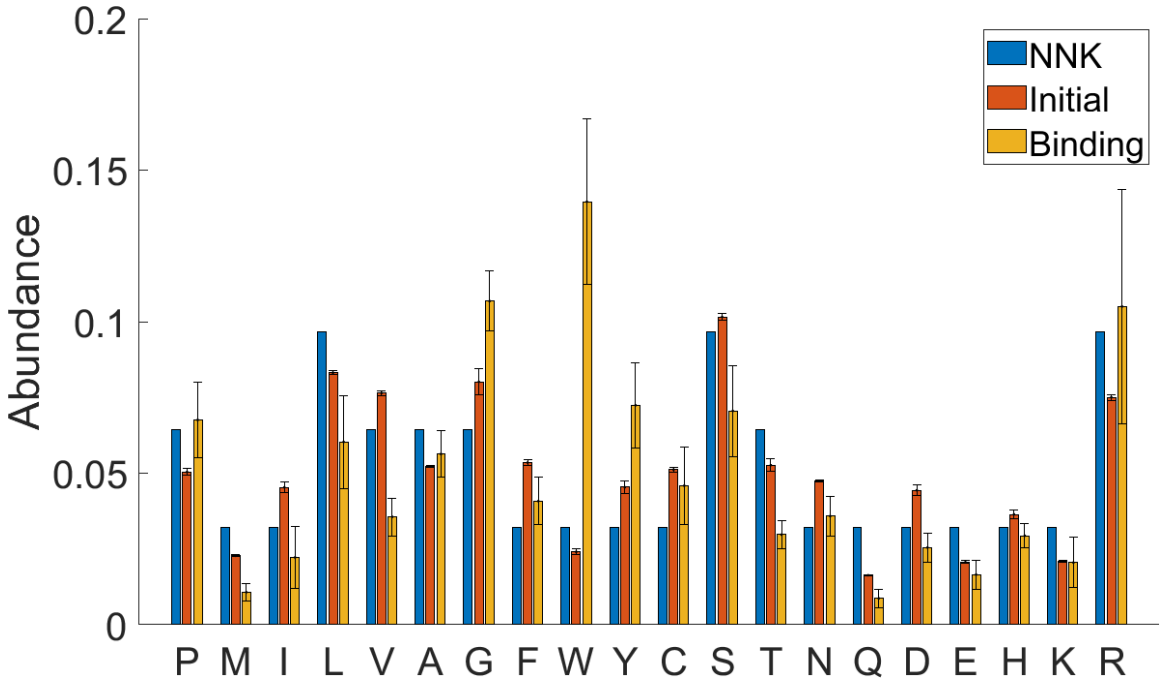

Figure S2.7 - Amino acid abundance across all protein scaffold paratopes

NNK: theoretical NNK codon. *Initial*: unsorted library. *Binding*: sorted binding population. Data are presented as the mean ± standard error of n=3 initial pools and n=7 binding pools.

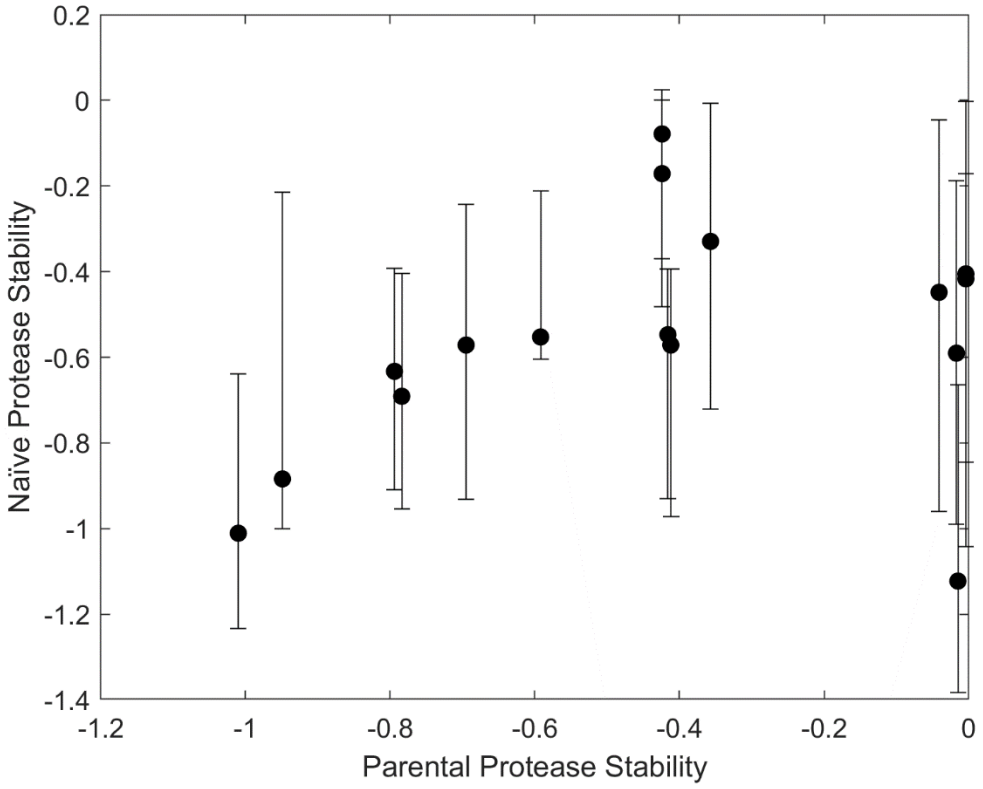

Figure S2.8 - Proteolytic stability comparison

The proteolytic stability of parental proteins and an average of 60 naïve variants per protein scaffold were measured on the surface of yeast. A range of stabilities were observed for each scaffold. There is no significant correlation between parental stability and naïve variant stability. Marker represents median with error bars drawn to the 10^{th} and 90^{th} percentile of naïve proteolytic stability.

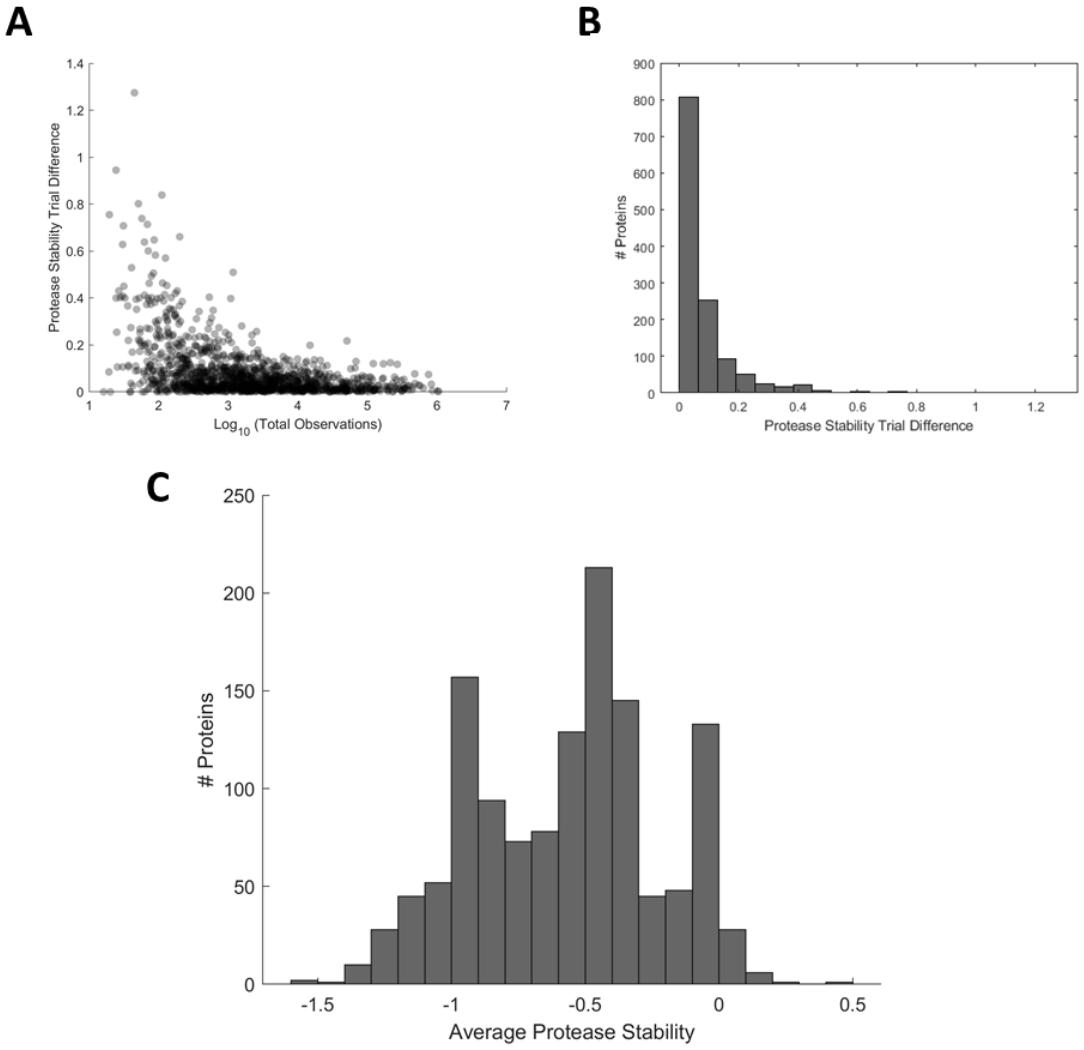

Figure S2.9 - Proteolytic stability of yeast-displayed proteins

Protease stability is determined by the amount of protein cleavage with increasing protease concentration. Two technical replicates of flow cytometry were completed, and the protein stability is reported as the average from each trial. **A**. The error between trials decreases for more observed proteins. **B**. A majority of the scaffolds showed less than 0.05 error in protease stability between trials. **C**. The total distribution protease stability. Values above zero indicated less cleavage with increasing protease and are likely due to error.

Chapter 3 - High-Throughput Developability Assays Enable Library-Scale Identification of Producible Protein Scaffold Variants

Adapted from "Alexander W. Golinski, Katelynn M. Mischler, Sidharth Laxminarayan, Nicole L. Neurock, Matthew Fossing, Hannah Pichman, Stefano Martiniani, and Benjamin J. Hackel. 'High-Throughput Developability Assays Enable Library-Scale Identification of Producible Protein Scaffold Variants.' PNAS, June 8, 2021, 118 (23) e2026658118."

3.1 Abstract

Proteins require high developability - quantified by expression, solubility, and stability - for robust utility as therapeutics, diagnostics, and in other biotechnological applications. Measuring traditional developability metrics is low-throughput in nature, often slowing the developmental pipeline. We evaluated the ability of ten variations of three high-throughput developability assays to predict the bacterial recombinant expression of paratope variants of the protein scaffold Gp2. Enabled by a phenotype/genotype linkage, assay performance for 10^5 variants was calculated via deep sequencing of populations sorted by proxied developability. We identified the most informative assay combination via cross-validation accuracy and correlation feature selection and demonstrated the ability of machine learning models to exploit nonlinear mutual information to increase the assays' predictive utility. We trained a random forest model that predicts expression from assay performance that is 35% closer to the experimental variance and trains 80% more efficiently than a model predicting from sequence information alone. Utilizing the predicted expression, we performed a sitewise analysis and predicted mutations consistent

with enhanced developability. The validated assays offer the ability to identify developable proteins at unprecedented scales, reducing the bottleneck of protein commercialization.

3.2 Significance Statement

Poor protein developability is a critical hindrance to biologic discovery and engineering. Experimental capacity limits variant analysis. We demonstrate the ability of an on-yeast protease assay, a split GFP assay, and a split β-lactamase assay to predict recombinant protein production yields in bacteria. The assays presented increase the ability to measure protein developability by more than 100-fold over traditional approaches. Compared to models trained using sequence information alone, the assays are 35% more accurate and require 80% less data to achieve the same prediction accuracy as sequence-based models. The assays were evaluated via randomized protein variants within a protein scaffold topology and offer a method to remove the limitation of variant developability quantification.

3.3 Introduction

A common constraint across diagnostic, therapeutic, and industrial proteins is the ability to manufacture, store, and use intact and active molecules. These protein properties, collectively termed developability, are often associated to quantitative metrics such as recombinant yield, stability (chemical, thermal, and proteolytic), and solubility[48,49,56,117,118]. Despite this universal importance, developability studies are performed late in the commercialization pipeline[117,118] and limited by traditional experimental capacity[47]. This is problematic because: (i) proteins with poor developability limit practical assay capacity for measuring primary function, (ii) optimal developability is often not observed with proteins originally found in alternative formats (such as display or two-hybrid

technologies[119]), and (iii) engineering efforts are limited by the large gap between observation size (~10^2) and theoretical mutational diversity (~10^{20}). Thus, efficient methods to measure developability would alleviate a significant bottleneck in the lead selection process and accelerate protein discovery and engineering.

Prior advances to determine developability have focused on calculating hypothesized proxy metrics from existing sequence and structural data or developing material- and time-efficient experiments. Computational sequence-developability models based on experimental antibody data have predicted post-translational modifications[120,121], solubility[122,123], viscosity[124], and overall developability[49]. Structural approaches have informed stability[125] and solubility[55,122]. However, many in silico models require an experimentally solved structure or suffer from computational structure prediction inaccuracies[126]. Additionally, limited developability information allows for limited predictive model accuracy[127]. In vitro methods have identified several experimental protocols to mimic practical developability requirements (*e.g.*, AC-SINS[51] and chemical precipitation[128] as metrics for solubility). However, traditional developability quantification requires significant amounts of purified protein. Noted in both fronts are numerous in silico and/or in vitro metrics to fully quantify developability[48,56].

We sought a protein variant library that would benefit from isolation of proteins with increased developability and demonstrate the broad applicability of the process. Antibodies and other binding scaffolds, comprising a conserved framework and diversified paratope residues, are effective molecular targeting agents[33,36,46,77,129]. While significant progress has been achieved with regards to identifying paratopes for optimal binding strength and specificity[42,130], isolating highly developable variants remains plagued. One

particular protein scaffold, Gp2, has been evolved into specific binding variants toward multiple targets[26,71,131]. Continued study improved charge distribution[132], hydrophobicity[133], and stability[26]. While these studies have suggested improvements for future framework and paratope residues (including a disulfide stabilized loop), a poor developability distribution is still observed[134] (Figure 1a,b). Assuming the randomized paratope library will lack similar primary functionality, the Gp2 library will simulate the universal applicability of the proposed high-throughput (HT) developability assays.

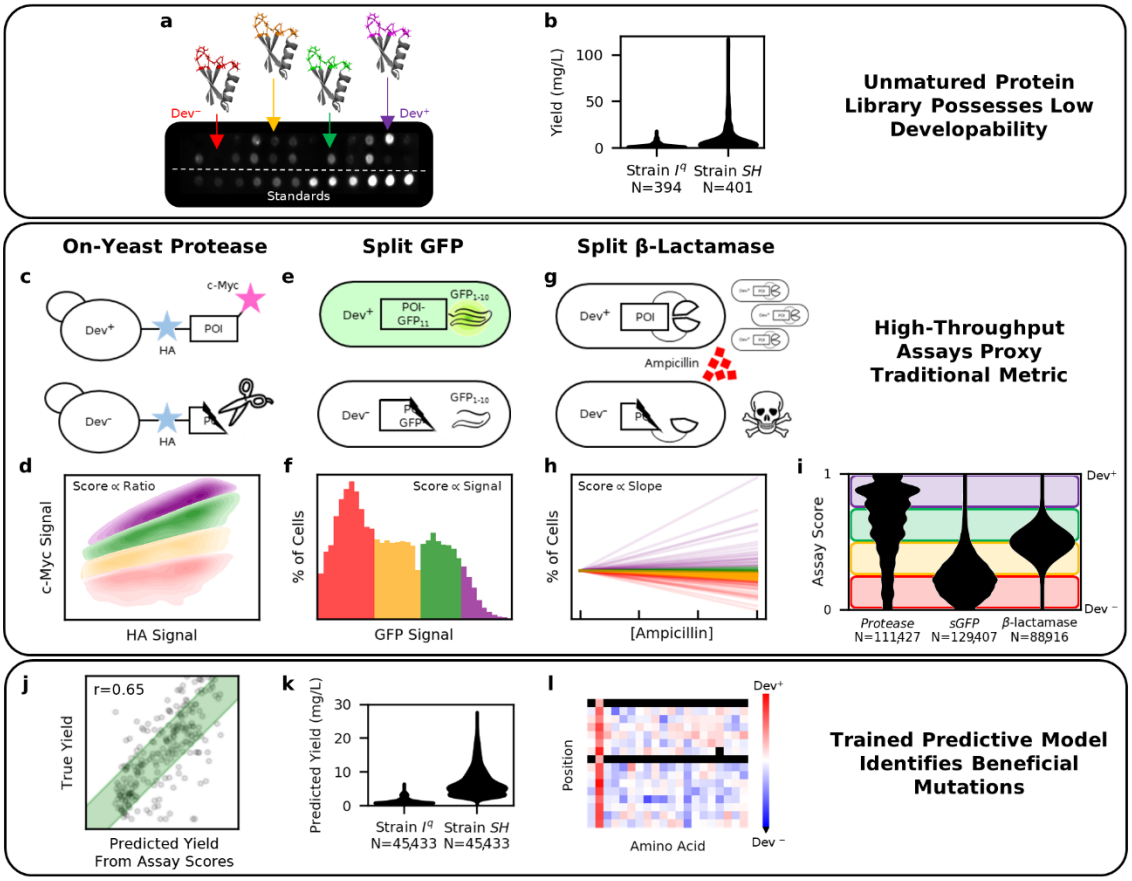

Figure 3.1 - High-throughput (HT) assays were evaluated for the ability to identify protein scaffold variants with increased developability.

a,b) Gp2 variant expression, commonly measured via low-throughput techniques such as the dot blot shown, highlights the rarity of ideal developability. **c,d)** The HT on-yeast protease assay measures the stability of the protein of interest (POI) by proteolytic extent. **e,f)** The HT split-GFP assay measures POI expression via recombination of a genetically fused GFP fragment. **g,h)** The HT split β-lactamase assay measures the POI stability by observing the change in cell growth rates when grown at various antibiotic concentrations. **i,j)** Assay scores, assigned to each unique sequence via deep-sequencing, were evaluated by predicting

expression (see Figure 3). **k,l)** HT assay capacity enables large-scale developability evaluation and can be used to identify beneficial mutations (see Figure 4).

We sought HT assays that allow protein developability differentiation via cellular properties to improve throughput. Variations of three primary assays were examined: 1) On-yeast stability (Figure 1c,d) - previously validated to improve the stability of de novo proteins[98], antimicrobial lysins[135], and immune proteins[136] - measures proteolytic cleavage of the protein of interest (POI) on the yeast cell surface via fluorescence activated cell sorting (FACS). We extend the assay by performing the proteolysis at various denaturing combinations to determine if different stability attributes (thermal, chemical, protease specificity) can be resolved. 2) Split green fluorescent protein (GFP, Figure 1e,f) - previously used to determine soluble protein concentrations[137] - measures the assembled GFP fluorescence emerging from a 16-amino acid fragment (GFP_{11}) fused to the POI after recombining with the separably expressed GFP_{1-10}. We extend the assay by utilizing FACS to separate cells with differential POI expression to increase throughput over the plate-based assay. 3) Split β-lactamase (Figure 1g,h) - previously used to improve thermodynamic stability[138] and solubility[139] - measures cell growth inhibition via ampicillin to determine functional lactamase activity achieved from reconstitution of two enzyme fragments flanking the POI. We expand assay capacity by deep sequencing populations grown at various antibiotic concentrations to relate change in cell frequency to functional enzyme concentration.

In this paper, we determined the HT assays' abilities to predict Gp2 variant developability. We deep-sequenced the stratified populations and calculated assay scores (correlating to hypothesized developability) for ~10^5 Gp2 variants (Figure 1i). We then converted the assay scores into a traditional developability metric by building a model that

predicts recombinant yield (Figure 1j). The assays' capacity enabled yield evaluations for >100-fold traditional assay capacity (Figure 1k, compared to Figure 1b) and provide an introductory analysis of factors driving protein developability by observing beneficial mutations via predicted developable proteins (Figure 1l).

3.4 Results

3.4.1 Gp2 Paratope Library Quantification

We first evaluated the assays' ability to separate sequence classes with a hypothesized difference in developability. 204,174 observed Gp2 variants belonged to one of four classes (Figure 2a): *GaR*: a thermostable variant[71], *Stop*: 13,690 nonfunctional truncated variants; *CC+*: 128,854 variants with a hypothesized[26] stabilizing cysteine pair at sites 7 and 12; *CC−*: 61,629 variants without conserving sites 7 and 12. *CC+*, *CC−*, and *Stop* classes utilize a previously optimized conserved framework[26] and two paratope loops, each with 6-8 'NNK' degenerate codons encoding all 20 amino acids (Figure 2b). The library was widely diversified, averaging 13.1 differences between observed sequence pairs (Figure 2c).

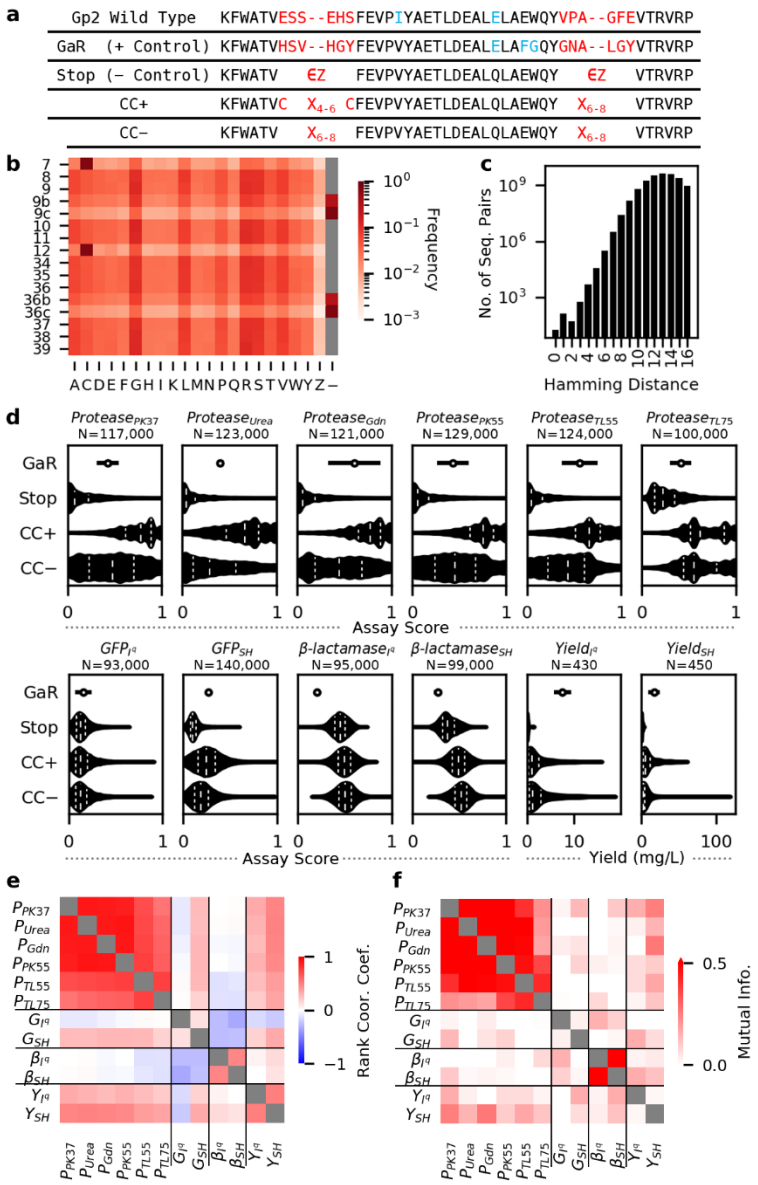

Figure 3.2 - Developability Characterization of Loop-Diversified Gp2 Library

a) Sequence alignment of assayed sequence classes: *GaR* (single variant control), *Stop*: (sequences with stop codon, Z), *CC+*: (hypothesized to be more developable), *CC−*: (hypothesized to be less developable) **b**) Diversified paratope frequency heatmap. **c**) Histogram depicting the pairwise distances between 190,483 full length and unique variants. **d)** Assay performance distributions divided by class. *Top Row*: various on-yeast protease assay reaction conditions. *Bottom Row*: bacterial assays performed in strain I^q and strain SH. GaR error bars represent the standard deviation (N=3 trials). Total unique variants for Stop, CC+, and CC− range 93,178-140,229 for HT assays, and 431-447 for yield (See SI Appendix, Figure S2) **e)** The Spearman's rank correlation coefficient between and **f)** mutual information between HT assays and yield.

3.4.2 Recombinant Yield as a Traditional Developability Metric

We sought a traditional developability metric that was translationally relevant and scalable to train and validate predictive models. A key step in developing and using a protein involves recombinant production. Bacterial cells are often chosen due to affordability, ease, and speed[140]. However, with limited production machinery, expressed proteins must rely on inherent developability parameters to achieve high soluble concentrations. Also, considering alternative assays require high purified protein quantities, we selected bacterial recombinant yield as the metric of interest. The Gp2 titer in the soluble lysate fraction was measured using a chemiluminescent quantitative dot blot protocol[141] via a C-terminal His_6 tag (Figure 1a,b).

Different bacterial strains have been evolved containing additional machinery to obtain increased yield. We chose to include two *E. coli* strains (T7 Express lysY/I^q (I^q) and SHuffle® T7 Express lysY (SH), New England Biolabs) for improved developability resolution. SH was chosen to stabilize disulfide formation and increase cysteine-free variant yields[142]. This was confirmed by *GaR* having a significantly higher yield in SH despite not having cysteines ($p<0.05$ in one-way Student's t-test using trial-averaged yield, n=8 plates per strain).

The recombinant yield of unique Gp2 sequences in each class was measured in triplicate (Figure 2d): *GaR* (both strains), *Stop* (I^q: 37 Gp2 variants, SH: 46), *CC-* (I^q: 98, SH: 117), and *CC+* (I^q: 296, SH: 284). *GaR* had a significantly higher yield than most *Stop* sequences (I^q: 100%, SH: 63% of unique *Stop* sequences, $p<0.05$ in one-way Student's t-test using plate-averaged *GaR* standard deviation, n=3 trials), validating the dot blot controls while suggesting slight noise with SH. *CC+* did not have significantly different yields than *CC-* ($p = 0.40$ in two-way Mann-Whitney U test) in I^q, while the populations

were significantly different in SH ($p<0.05$, one-way Mann-Whitney U test). This implies SH is forming a disulfide bond, thus increasing *CC+* sequence developability.

3.4.3 HT Developability Assays

The Gp2 variants were sorted into populations of varying developability and were assigned an HT assay score as the mean over three independent trials (SI Appendix, Figure S1). Below we motivate score calculation, followed by assay score distribution analysis (Figure 2d, SI Appendix, Figure S2).

3.4.3.1 On-yeast Stability

The on-yeast stability assay evaluates protein stability by measuring proteolytic cleavage (Figure 1c). Using yeast surface display technology[67], the POI is expressed between two tags (N-terminal HA, C-terminal cMyc). The protein-displaying yeast are exposed to a protease at a concentration that produces a distribution of cleavage (as determined by cMyc:HA ratio) across protein variants. The Gp2 library was sorted into four populations (Figure 1d). Sequencing scored every collected variant on a cell-weighted average: 1 (intact), 2/3, 1/3, 0 (fully cleaved).

We performed the proteolysis using various conditions to determine if additional stability metrics could be obtained (SI Appendix, Figure S1). From our baseline condition (P_{PK37}), we studied chemical stability by adding 1.5 M urea (P_{Urea}) or 0.5 M guanidinium chloride (P_{Gdn}). We explored protease specificity by using proteinase K (P_{PK55}) and thermolysin (P_{TL55}). Finally, we examined thermostability for each enzyme at an additional temperature (P_{PK37} vs. P_{PK55} and P_{TL55} vs. P_{TL75}).

Assay scores were calculated for $>10^5$ unique Gp2 variants in each of the 6 reaction conditions. The assay score distributions per class (Figure 2d) matched hypothesized

developability in all conditions except P_{TL75}. Standard deviations were small (0.17 – 0.20, except P_{TL75}: 0.29). *Stop* variants scored low (0.04 – 0.08, except P_{TL75}: 0.23). *GaR* scored higher than most *Stop* variants (67 – 81%, except P_{TL75}: 35%). One potential hypothesis for P_{TL75} is the increased temperature may lead to non-specific binding of surface-aggregated proteins. Nevertheless, all reaction conditions, displayed a significantly higher distribution of assay scores for *CC+* vs. *CC-* (one-way Mann-Whitney U test, $p<0.001$), validating each condition's utility.

3.4.3.2 Split GFP

The split GFP assay measures POI concentration with a C-terminus fused 11th strand of GFP (Figure 1e). Upon recombination with GFP strands 1-10, which was separately induced following POI production and a one-hour gap, the POI fusion remaining soluble in the cytosol will produce a fluorescent signal detectable by FACS (Figure 1f). The library was sorted into four populations based on GFP signal and assigned an assay score as a cell-weighted average: 1 (highest signal), 2/3, 1/3, 0 (background signal).

The assay score distributions (Figure 2d) are consistent with expectations in SH (G_{SH}) with limited resolution in I^q (G_{Iq}). While both distributions display a low assay score skew, *GaR* had a significantly higher score than 76% of *Stop* in G_{SH}, compared to 8% in G_{Iq}. Additionally, G_{SH} produced a significantly higher assay score distribution for *CC+* compared to *CC-* (one-way Mann-Whitney U test. $p<0.001$) whereas G_{Iq} scores were only nominally higher ($p=0.15$). Thus, G_{SH} is a compelling candidate for HT developability analysis.

3.4.3.3 Split β-lactamase

In the split β-lactamase assay, the POI is inserted in a loop distal to the active site (final construct: $\beta\text{-lac}_{1\text{-}194}\text{-}(G_4S)_2\text{-AS-POI-GS-}(G_4S)_2\text{-}\beta\text{-lac}_{197\text{-}287}$, location previously observed to retain 40% activity[143]). Functional enzyme, hypothesized to be paired with POI solubility and folding robustness[144], provides ampicillin resistance allowing cell reproduction (Figure 1g). The change in growth rates was measured as the change in POI amplicon abundance in cultures grown to saturation with varying antibiotic concentrations (Figure 1h). For comparison to other assays and improved modeling efficiency, slopes were normalized and scaled (see Methods).

The split β-lactamase assay produced assay scores that were contradictory towards hypothesized developability yet were able to differentiate classes, suggesting potential utility despite an unsolved mechanism. We obtained assay scores for $1x10^5$ variants in both I^q (β_{Iq}) and SH (β_{SH}). Independent *GaR* cultures (capable of growing at all concentrations) and *Stop* (unable to grow in non-zero ampicillin concentrations) performed as expected (SI Appendix, Figure S3). Yet in multi-POI culture, *GaR* had a significantly lower assay score than *Stop* (β_{Iq}: 99%, β_{SH}: 70%, one-way Student's t-test, $p<0.05$), and the *CC+* population had a significantly lower assay score distribution than *CC-* (both strains, one-way Mann-Whitney U test, $p<0.001$). See Figure 5 and Discussion for further explanation.

3.4.4 Determination of Most Predictive HT Assay Conditions

While the HT assays broadly differentiated hypothesized class developability, the ability to transform the assay scores to a traditional metric is a superior utility assessment. Despite the limited sensitivity in the split GFP assay and the counterintuitive split β-lactamase distributions with minimal rank correlation to yield (Figure 2e), the assays have

nonzero mutual information (MI) with yield, suggesting utility as long as the predictive model is capable of exploiting the nonlinear relationships captured by MI (Figure 2f). In this section, we determine the optimal HT assay set (assay type, reaction conditions, and/or bacterial strain) by the ability to predict recombinant yield with the lowest mean-squared error (MSE) loss.

With a potential complex relationship between developability and assay scores, we designed our model to maximize the ability to detect assay utility. Correlation of yields in both strains was observed (ρ *CC*+: 0.65, *CC*-: 0.61; SI Appendix, Figure S4); thus, a multitask model (Figure 3a) was utilized to include both strains' yield measurements via a one-hot (OH) encoded vector. We included relevant comparisons for model inputs: a null strain-only model (predicts the mean yield per strain) and a OH sequence model (encoded and flattened paratope sequence). To capture possible linear and nonlinear relationships between assay scores, sequences, strains, and yield, four model architectures (ridge, random forest, support vector machine, and a feedforward neural network) were employed.

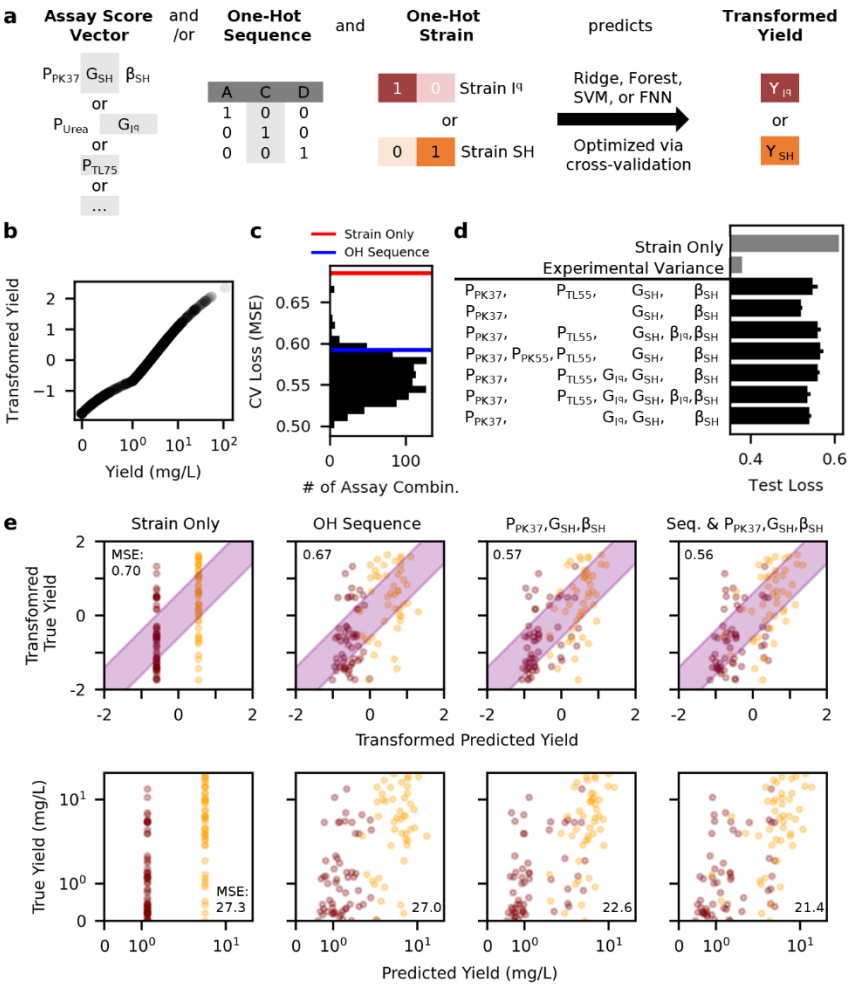

Figure 3.3 - Determination of Predictive HT Developability Assays

a) Model visualization utilizing HT assay scores, a one-hot paratope sequence, and a one-hot strain identifier to predict the recombinant yield in both cell types. **b**) Power-transformation and standardization of yields to remove correlation between yield and error (See SI Appendix, Figure S5). **c**) Predictive model loss (mean squared error (MSE) between predicted and actual yields) distribution for 1,023 HT assay combinations. **d)** The top combinations from CV (listed top down) were tested for generalizability by the predictive loss against independent set of 44 sequences. **e)** Representative scatter plots of predicted versus measured yield (I^q: purple, SH: orange; *top:* power-transformed and normalized, *bottom:* non-transformed) during final evaluation on set of 97 sequences. Purple shaded area represents true yield ± square root of sequence-averaged experimental variance.

Cross-validation (CV) and hyper-parameter optimization were trained by 195 unique sequences observed in all HT assays and for which yield was measured in at least one strain. A Yeo-Johnson[145] power transform and normalization was applied to remove correlation between error and yield ($\lambda = -0.324$, SI Appendix, Figure S5). The experimental

variance (measurement accuracy) was calculated as the sequence-averaged trial-to-trial (n=3) variance after applying the transformation to trial yields.

Despite potential limitations, all 1023 assay combinations of the 10 HT conditions predicted yield with a lower CV loss than the strain-only control, and 92% of the combinations outperformed the OH sequence model (Figure 3b) suggesting all conditions possess utility. There were seven assay combinations (using seven of the ten assays) that performed optimally and equally (SI Appendix, Figure S6, one-way Student's t-test against top model, $p>0.05$). To determine the most generalizable collection, the yield for an independent set of 44 sequences (not utilized during CV but observed in top seven HT assays) was predicted revealing the most informative set: P_{PK37}, G_{SH}, β_{SH} (Figure 3c, one-way Student's t-test against top model, $p<0.05$).

The top three HT assays can provide substantial predictive power for variant developability over sequence or strain information alone. The yield for a second set of 97 sequences (not utilized during CV but observed in top three HT assays) was predicted (Figure 3d, SI Appendix, Figure S6). The assay model (MSE: 0.565) was able to significantly (one-way Student's t-test, $p<0.05$) outperform the one-hot sequence model (MSE: 0.667) and strain-only model (MSE: 0.697). A model utilizing both sequence and assay information (MSE: 0.562) did not have significantly different ($p>0.05$) performance from the assay model alone, suggesting little aid of sequence knowledge as currently implemented. The model utilizing sequence and assay information, while predicting better than alternatives, required a nonlinear random forest architecture with 325 trees for optimal predictive performance that still trails the experimental variance (0.364), suggesting room for future improvement. As performed, the assays reduce the gap between prediction and

experimental error of developability evaluation by 35% compared to sequence information alone.

A practical application of the HT developability assays is the ability to isolate sequences with increased developability from those without. To this effect, we calculated a receiver-operator curve (ROC) and precision-recall curve via pretrained models to classify the independent test sequences in the top 50^{th} percentile of each strain (SI Appendix, Figure S7). When utilizing the HT assay scores, the area under the ROC was improved from 0.59 to 0.71 (Strain I^q) and from 0.55 to 0.69 (Strain SH) over the OH sequence model. The average precision, a metric more focused on correctly identifying the positive class, was improved from 0.56 to 0.71 (Strain I^q) and 0.55 to 0.70 (Strain SH) demonstrating the HT assays are also capable of isolating developable sequences.

3.4.5 Optimal Paratope Sequence Identification

With a predictive model to translate the assay scores to recombinant expression, we aimed to understand the sequence-developability relationship. The predictive model utilizing $P_{PK37,}$ G_{SH}, β_{SH} assay scores and OH sequence was used to predict the yield for 45,433 unique sequences in both strains (Figures 1k & 4a). After observing the predicted yield distribution, 6,394 sequences with a predicted I^q yield > 2.5 mg/L (transformed yield > 0.0) and SH yield > 6.4 mg/L (transformed yield > 0.75) were isolated as Dev^+. The pairwise Hamming distance distribution for the Dev^+ sequences (median 12.3) is shifted to significantly lower values than the initial distribution (median 13.0, χ^2, $p < 0.05$), suggesting that developable sequences exist in a partially constrained subset of sequence space.

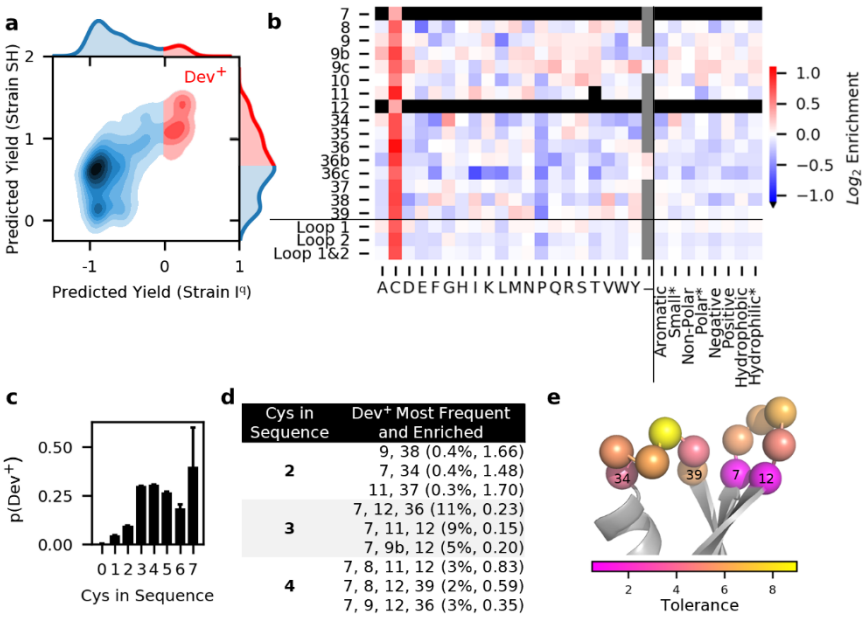

Figure 3.4 - HT Assays Enable Prediction of Gp2 Variants with High Developability

a) Kernel density plot of the predicted yield of 45,433 unique sequences in each bacterial strain. 6,394 sequences with high predicted yield in both strains were isolated as Dev$^+$ (red). **b)** Sitewise enrichment heatmap (Dev$^+$ versus all predicted sequences) for each amino acid and averaged groups with similar chemical properties: aromatic (F, W, Y), small* (A, G, S), non-polar aliphatic (A, G, I, L, M, P, V), polar uncharged* (N, Q, S, T), negative charged (D, E), positive charged (H, K, R), hydrophobic (A, F, G, I , L, M, P, V, W, Y), and hydrophilic* (D, E, H, K, N, Q, R, S, T). *Note: cysteine was removed to identify any further enrichment of the groups. Loop 1: positions 8-11. Loop 2: positions 34-39. **c)** The proportion of sequences predicted identified as Dev$^+$ as a function of the number of cysteines in the sequence. Error bars: 1 / number of predicted sequences. **d)** The most frequent (percent of Dev$^+$) and enriched (log_2 of Dev$^+$ versus all predicted) positions for combinations of cysteines that result in high developability proteins. **e)** Wild type paratope positions of Gp2 (PDB: 2WMN) colored by the mutational tolerance calculated as the inverse of the average magnitude of amino acid enrichment.

To identify beneficial, tolerable, and detrimental mutations to developability, the log_2 difference in amino acid frequency at each position between Dev$^+$ and all predicted sequences was calculated (Figure 4b). Cysteine was the only positively enriched amino acid at positions 7 and 12 (confirming CC+ stability) but was also the most enriched at every position. The high cysteine enrichment was also observed when analyzing predictions of an assay score model without sequence information (SI Appendix, Figure S8). Regarding epistasis, we analyzed the probability of Dev$^+$ as conditioned by number of

cysteines in the sequence, finding 3 or 4 cysteines most optimal (Figure 4c). There also appears to be a benefit of 7 cysteines, however the limited number of sequences (n=5) limits the confidence in the benefit. To determine the best cysteine locations to improve developability, the Dev^+ frequency and $\log_2$ enrichment were calculated (Figure 4d). It should be noted the 7 and 12 pair had a negative enrichment likely due to the artificially increased initial frequency. As additional cysteines may be disfavored for downstream processing flexibility, the enrichment of sequences only containing cysteines at positions 7 and 12 was calculated (SI Appendix, Figure S9). Enabled by the assay throughput, less-extreme enrichment values observed for cysteine-rich sequences (compared to sequences with fewer cysteines) suggests the cysteines are buffering stability and permitting a wider sequence set. The preference of cysteines in Dev^+ sequences could be partially impacted by disulfide-driven protease resistance in the on-yeast stability assay, e.g. with a free cysteine located near the active site of proteinase K^{146}. However, both the OH model and a model utilizing only assays G_{SH} and β_{SH} also indicate a stabilizing effect of additional cysteines (SI Appendix, Figure S10c-f). Moreover, recombinant yield increased at higher cysteine frequencies of synthesized variants (ρ: I^q = 0.28, SH = 0.48, SI Appendix, Figure S12a,b).

Additional analysis enabled by the HT assays were used to hypothesize properties that drive Gp2 stability. The enrichment of small residues (alanine, glycine, and serine) at position 34, the proline depletion in the second loop, and gap enrichment at positions 36b,c (enriching sequences of wild type length) suggests the second loop may be geometrically constrained. We assessed positional mutational tolerance (ability to mutate without modifying developability) by calculating the inverse of the average enrichment score

magnitude (Figure 4e). Positions 7 and 12 were the most constrained (tolerances: 0.5), signifying the need to be cysteines. While position 37 was the least constrained position (8.8), as a whole loop 2 (5.5) was less tolerant than loop 1 (5.9, excluding 7 and 12). We hypothesize either i) the second loop is a poor paratope in terms of allowing broad diversity with favorable developability or ii) the stabilizing disulfide bond offsets unfavorable mutations within the first loop.

3.4.6 β_{SH} Assay Predictive Performance Explained by Mutual Information

Like amino acid preference, we sought a first-order understanding of optimal assay scores by looking at the Dev^{+} distribution compared to all observed unique sequences (Figure 5a). Matching the sequence class distributions (see Figure 2), P_{PK37} and G_{SH} assay scores of Dev^{+} sequences were significantly higher, and β_{SH} assay scores was significantly lower than the initial distribution (Figure 5a, one-way Mann-Whitney U test, $p<0.05$). However, the rank correlation between β_{SH} and yield is slightly positive (Iq: 0.00, SH: 0.11), suggesting the model is exploiting a nonlinear relationship.

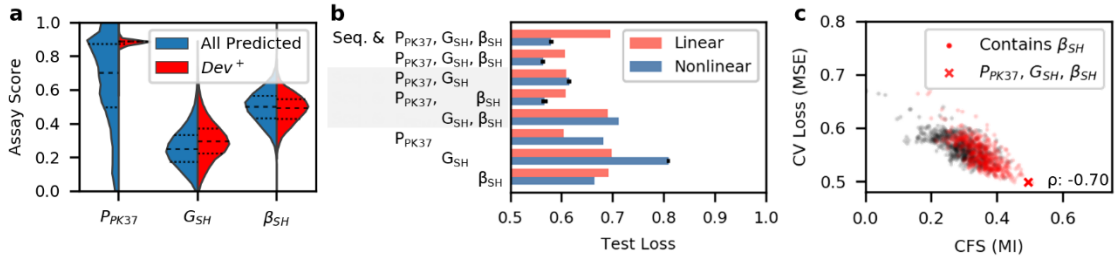

Figure 3.5 - Nonlinear models can extract nonlinear developability mutual information (MI) from the split β-lactamase assay

a) Comparison of assay score distributions between 45,433 unique sequences with observed $P_{PK37,}$ G_{SH}, β_{SH} assay scores (blue) versus 6,394 of the sequences with high predicted developability (Dev^{+}, red). **b)** The predictive performance of model input combinations in both a linear architecture (ridge regression) and nonlinear architectures (reported top performance of random forest, support vector machine, and a feed-forward Neural Network). Error bars in nonlinear models represent standard deviation in MSE from n=10 stochastically trained models. **c)** The correlation-based feature selection (CFS) as calculated by MI for 1023 assay combinations versus the CV loss utilizing the best of linear and nonlinear model architectures. The

Spearman's rank correlation coefficient (ρ) between CFS and loss confirms the ability of the models to extract nonlinear MI.

We hypothesize that the counterintuitive relationships between β_{SH} and yield resulted from several competing interactions relating the change in sequence frequency to the concentration of functional enzyme-POI. We tested this by comparing nonlinear versus linear model performances for several model input combinations (Figure 5b). While the P_{PK37} and G_{SH} assays, alone and together, performed better with a linear model, 4 of 5 models using the β_{SH} assay performed best with a nonlinear model.

The correlation-based feature selection[147] (CFS) explains how the nonzero MI between β_{SH} and yield (I^q: 0.16, SH: 0.13) resulted in increased predictive power by supplying non-redundant information with respect to other HT assays. The CFS calculated by MI was significantly higher and CV loss was significantly lower of HT assay combinations containing β_{SH} than assay combinations without (Figure 5c, one-way Mann-Whitney U test, $p < 0.05$). CFS calculated with MI was highly correlated with loss when utilizing nonlinear models ($\rho = -0.70$) remarking its effectiveness as a feature selection tool. We also found CFS calculated by rank correlation was correlated to linear model performance ($\rho = -0.56$) but less so to overall performance ($\rho = -0.30$) as linear models cannot exploit nonlinear relationships (SI Appendix, Figure S11). As a result, the top CFS combination via rank correlation (P_{PK37}, P_{Urea}, P_{PK55}, G_{Iq}, G_{SH}. Ridge MSE: 0.564) increased the prediction error relative to experimental variance by 46% compared to the top model identified by CFS via MI (P_{PK37}, P_{TL55}, G_{SH}, β_{SH}. Forest MSE: 0.497). While the current selection of HT assays were chosen by hypothesized utility, based upon the results of CFS, future HT assays, such as systems for assessing protein foldability[148,149], should be considered if it is believed the assay will provide a non-redundant metric of developability.

3.4.7 Training Sample Size Evaluation

Next, we asked how the predictive performance scales versus the number of training sequences. We first analyzed how many sequences it takes for a model to learn training set developability, as determined by outperforming the strain only model during CV (Figure 6a). With only 10 sequences (5% of data), the P_{PK37}, G_{SH}, β_{SH} model achieves this goal (one-way paired t-test, $p<0.05$). However, models with sequence information required at least 39 sequences (20% of data) to achieve the same accomplishment, suggesting the increased input dimensionality limits the model's ability to learn. When evaluating the models for generalizability against a test set (Figure 6b), the models using assays required only 59 (P_{PK37}, G_{SH}, β_{SH}, 30% of data, $p<0.05$) or 78 (Sequence and P_{PK37}, G_{SH}, β_{SH}, 40% of data, $p<0.05$) training sequences to outperform the strain only model, while the sequence only model required all 195 sequences. The generalizability results suggest the HT assays reduce the training data requirements by 60-70% over sequence information alone.

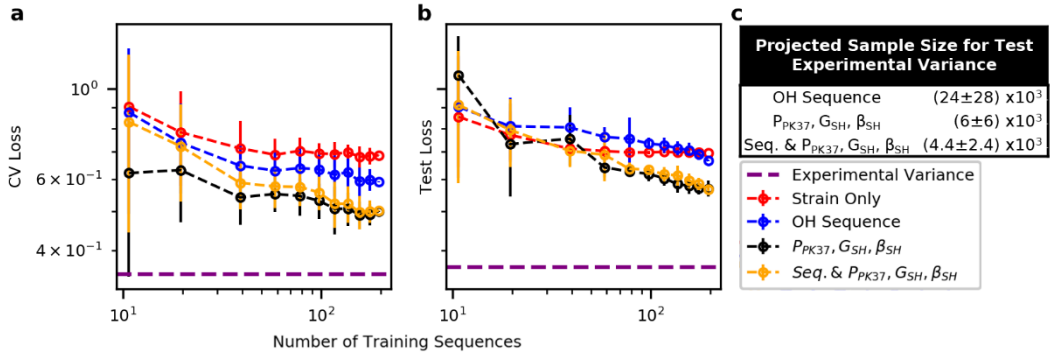

Figure 3.6 - HT developability assays reduce training size requirement

Ten bootstrapped samples for each sample size (ranging 5-100% of available data) were individually trained by CV and evaluated on 97 independent test sequences. Error bars represent standard deviation across models. **a**) The performance during CV describing the model's ability to predict developability. **b)** The predictive performance against the independent sequence set describing the models' ability to generalize beyond the training data. **c)** The generalizable performance was extrapolated to estimate the required number of sequences for the model to perform optimally. Log-log regression was trained with points weighted by the

inverse of test loss variance. The error shown represents the propagated error from the standard errors of the parameter estimates.

We also extrapolate how many additional training sequences would be required to achieve performance within the measurement accuracy (experimental variance). For each model, we extrapolated a best-fit line between the $\log_{10}$ test loss and the $\log_{10}$ number of training sequences weighted by the inverse variance for each sample size (Figure 6c). We predict that utilizing the HT assay scores, the number of unique sequences required to obtain optimal performance is 80±40% (P_{PK37}, G_{SH}, β_{SH}) and 81±24% (Sequence and P_{PK37}, G_{SH}, β_{SH}) lower than what would be required when considering sequence information alone, which demonstrates the efficiency of the HT assays to enable developability engineering.

3.4.8 Error Analysis

While the trio of assays provide valuable developability assessment, we sought to identify factors that limit performance. Due to the sampling strategy (see Methods), the observation frequency of variants analyzed via dot-blot was higher than the distribution of all variants observed in the HT assays (SI Appendix, Figure S12a). However, we observed non-significant correlation (ρ: I^q = 0.02, SH = 0.07, SI Appendix, Figure S12b) between the accuracy of our model and the predictive loss in either strain, suggesting that the predicted yields are not influenced by observation frequencies.

We next assessed if the number of collected populations per assay influenced the ability to predict recombinant yield. The assay scores were recalculated with only two merged populations from the HT assays at various levels of stringency (SI Appendix, Figure S13). We found that, if sequence and assay information is utilized, there is little benefit of utilizing four populations over two provided that the most stringent gate is used.

Interestingly, when only using assay scores to predict yield there was a decrease in predictive accuracy, especially when the highest stringency was not isolated. This suggests that future iterations of assay development may benefit from increasing resolution among the most developable variants.

Finally, we assessed the effect of trial-to-trial assay score variance for the top performing HT assays (SI Appendix, Figure S14). We found that the ability of the HT assays to predict yield increased when averaging assay scores over multiple trials. Thus, while trial-to-trial reproducibility was not limited (ρ: P_{PK37} = 0.66-0.71, G_{SH} = 0.26-0.29, β_{SH} = 0.39-0.48), the increased resolution of multiple trials may improve overall utility.

Combining the analysis of potential sources of error, we believe future studies will benefit most from increased technical replicates, with more moderate gains from increased stringency in isolating populations and minimal benefit from increased resolution via increased observation frequency. Yet, the relatively small impacts of high-throughput error identified in this section paired with moderate MI between assays and yield (SI Appendix, Figure 11d) suggest a more likely limitation is the difference in mechanisms driving success in each assay. For example: 1) the protease assays utilize a eukaryotic cell with more complex cellular machinery than the prokaryotic *E. coli*; 2) the split GFP assay measures intracellular protein concentration rather than the amount of extractable soluble protein during cell lysis; and 3) the split β-lactamase assay ties transport to the periplasm and enzymatic activity on top of the producibility measured via dot blot. Thus, pursuit of additional assays with non-redundant metrics of developability and closer mechanisms to the traditional metric should be sought to augment the significant predictive power already achieved with the current assays.

3.5 Discussion

Traditional protein developability measurements are restricted in practical throughput, reducing the number of protein variants that can be reasonably characterized. We evaluated HT assays that genetically encode the POI in a context where the cell's phenotype is related to the POI's developability. The on-yeast protease, split GFP, and split β-lactamase assays exhibited their ability to proxy protein developability via prediction of recombinant yield for Gp2 scaffold variants. HT assays increased the scale of protein developability differentiation by 100-fold (in this study: 400 yield measurements versus predicted yield via 40k HT assay measurements) and potentially enable analysis of developable sequences beyond those presented in this manuscript. Ligation efficiency for bacterial transformations and the sequencing depth per cost are current capacity limitations. However, future studies utilizing the narrowed set of optimal assay conditions determined in this work could potentially screen millions of unique variants with minimal modifications.

The most useful conditions were determined by comparing the predictive model performance of a traditional developability metric. Only one of six protease assay conditions were utilized in the top model, indicating that other conditions (chemical denaturants, elevated temperature, and alternative protease) were not needed to increase the predictive accuracy of recombinant soluble yield. This may be because the assays were unable to capture alternative stability metrics, or that a single stability metric is sufficient to predict developability. Additional assays may be useful for predicting other traditional developability metrics, such as thermostability. For example, P_{TL55} was found in 5 of 7 top CV models and may aid thermal predictions. The split GFP and split β-lactamase assays were most beneficial when utilizing SH assay scores despite predicting both strain's yield.

We hypothesize SH was able to increase developability resolution over I^q in our library by promoting stabilizing disulfide bonds and chaperoning the production of even weakly developable variants.

A nonlinear model was required to convert the split β-lactamase HT assay scores to a traditional developability metric. The reference assay evaluated enzymatic activity via minimum inhibitory concentration (MIC) of ampicillin by clonal colony growth on an agar plate[138]. While the exact differences between our measured assay score and the traditional MIC remains unclear, one possible explanation is a decrease in growth rate with increased protein production[150], lowering the frequency of highly produced variants. Library plating on agar plates could reduce this mechanism but with throughput limitations to achieve sufficient physical spacing to avoid bystander ampicillin reduction. In any event, despite the discrepancy, we have shown nonlinear models can extract useful developability information to predict recombinant yield. One assay limitation is the inability to perform direct selection, which is possible for the on-yeast protease and split GFP, based upon the linear model performance. A potential solution to streamline the discovery would be serial direct selections via on-yeast protease and split GFP, followed by a sequenced stratification via the split β-lactamase to increase accuracy.

The Gp2 library (~10^{20}) is well beyond the capacity of traditional developability assays that often fail to produce predictive sequence-based models. Utilizing the HT assays, we predicted yields 35% closer to experimental accuracy than a one-hot encoded sequence-based model trained on the same sequence set, proving their utility over naïve computational approaches in the vast protein domain. We studied the sitewise amino acid biases based upon predicted yield of 40k unique paratopes, which can be used to design

more effective libraries[41,42,91,151,152]. However, the analysis utility is limited by multi-site interactions (observed with cysteine) and model accuracy. We believe the increased knowledge will enable more advanced sequence-based models, capable of extrapolating developability to unobserved variants. The efficiency and accuracy of measuring developability proxies via HT assays empowers such models.

We estimate the HT assays will reduce the number of sequences required to produce an optimal predictive model by 80% compared to sequence information alone. Advances in experimental protocol (beyond those evaluated in this study) and alternative model architectures may provide other routes for increased utility. The assays presented in this work have shown the ability to evaluate the developability for a substantially higher number of unique sequences compared to traditional methods. These assays are essentially independent of protein primary function (assuming naïve Gp2 variants tested have no known primary function). Future work will validate the utility of integrating developability assays with discovery and evolution of primary function. Continued improvements of HT assay development may revolutionize the candidate selection process by presorting proteins for ideal developability before the primary function is evaluated, removing a discovery and engineering bottleneck.

3.6 Materials and Methods

The following section contains a summary of relevant information to perform the HT assays and predictive analyses. Additional methods can be found in the SI Appendix.

3.6.1 Subsampling Gp2 Library

We chose to subsample the transformed population to increase assay resolution by sampling multiple cells per sequence and performing assays in triplicate. We projected 10

reads per sequence for on-yeast protease and split GFP, and 10 reads per sequence per antibiotic concentration for the split β-lactamase assay, summing to 160 reads per sequence per trial across all ten assays. We found the limiting factors to be the capacity of high-throughput sequencing and bacterial ligation efficiency. Given that an Illumina NovaSeq SP flowcell can achieve 400 x 10^6 reads per lane for about $3,000, we decided on utilizing two lanes to analyze the 10^6 sequences to balance information and experimental cost. The realized difference in obtained sequence information is likely due to stochastic sampling leading to a bias in sequence frequencies.

3.6.2 On-Yeast Protease Assay

Dilutions of proteases and yeast were separately prepared on ice. Proteinase K (P8107S, New England Biolabs) was diluted to twice the reaction concentration in PBSA (P_{Urea} was diluted using 3 M urea in PBSA, P_{Gdn} was diluted using 1 M guanidium chloride in PBSA). Thermolysin (V4001, Promega) was reconstituted to 1 mg/mL in 50 mM Tris at pH 8 with 0.5 mM calcium chloride and diluted with PBSA on the day of experiment. Exposure time with protease at reaction temperature was held constant while the concentrations of protease were modified to obtain a roughly equal distribution of FACS gates' occupancy (SI Appendix, Figure S1).

Ten million yeast cells expressing the subsampled library were centrifuged at 5000g for 1 min, aspirated, resuspended in 1 mL cold PBSA, centrifuged, resuspended in 50 µL of PBSA, and transferred to a 0.2 mL PCR tube on ice. 50 µL of the diluted enzyme was added to the cells and mixed via pipetting on ice. The enzyme-yeast mixture was placed in a pre-chilled 4 °C PCR block where a preset program heated the mixture to the reaction temperature for 10 min and returned the mixture to 4 °C. Both heating and cooling rates

were set to the maximum ramp speed on the Eppendorf Mastercycler Nexus GX2. The enzyme-yeast mixture was then added to 1 mL of cold PBSA and the epitopes were labeled following the protocol used during library subsampling.

The cells were separated via FACS into four populations based upon the cMyc to HA ratio. The undigested gate (highest cMyc:HA ratio) was determined by the location of the library in a no-enzyme control. The fully digested gate (lowest cMyc:HA ratio) was determined by the location of the no-enzyme control where the primary mouse-anti-cMyc antibody was omitted. The other two gates were drawn to divide the remaining space in half. Collected cells were centrifuged and stored at -80 °C without allowing propagation.

3.6.3 Split GFP Assay

Frozen aliquots of cells were thawed and grown in 5 mL LB+Amp+Kan overnight. Part of the overnight culture was added to 5 mL fresh LB+Amp+Kan at an OD_{600} of 0.1 and grown for 90 min. Gp2-GFP_{11} production was induced by the addition of 0.5 mM IPTG. For the remainder of split-GFP protocol, both I^q and SH strains were grown at 37 °C. Production continued for 2 h, followed by a centrifugation (3000g for 3 min). Cells were then resuspended in 5 mL fresh LB+Amp+Kan and incubated for 1 h to end Gp2-GFP_{11} expression. GFP_{1-10} expression was then induced by adding 2 mg/mL arabinose and production continued for 2 h. Finally, the culture was centrifuged, resuspended in 1 mL cold PBSA, and stored on wet ice.

FACS was used to separate bacterial cells based upon the GFP signal. Background fluorescence was determined by cells containing the stop-GFP_{11} plasmid. The remainder of cells were divided into three equally (log scale) spaced gates. The collected populations

were centrifuged (3000g for 10 min) and frozen at -80 °C to inhibit growth. The cells were then thawed and miniprepped to obtain the Gp2-encoding plasmids.

3.6.4 Split β-lactamase Assay

Frozen aliquots of cells were thawed and grown in 5 mL LB+Kan overnight. Part of the overnight culture was added to 5 mL fresh LB+Kan at an OD_{600} of 0.01 and grown for 90 min. The split β-lactamase production was induced by the addition of 0.5 mM IPTG. Production was continued for 2 h at 37 °C (strain I^q) or 4 h at 30 °C (strain SH). The culture was then divided into 6 x 300 µL wells per concentration of ampicillin in a 96 well plate. 30 µL per well of diluted ampicillin was spiked in to achieve the desired final concentrations. The cultures were then monitored in a Synergy H1 microplate reader (BioTek) with continuous double-orbital shaking and the 600 nm absorbance obtained every five minutes. All wells for a given concentration of ampicillin when the average unnormalized absorbance reached 0.35 were removed from the plate, centrifuged (12,000 g for 3 min), and frozen at -80 °C to stop growth. The cells were then thawed and miniprepped to obtain the Gp2-encoding plasmids.

3.6.5 High-Throughput Assay Score Calculations

3.6.5.1 On-yeast protease and Split GFP Assay Score Calculation

The four collection gates in the FACS based assays were drawn to bin cells via hypothesized developability. Thus, we defined an assay score which correlates to the relative position of a sequence. To increase resolution, we collected an average of 6.7X (on-yeast protease) and 7.9X (split-GFP) the hypothesized diversity of cells per trial and assigned a score correlating to the average cell location.

For each population, the read frequency of every sequence was converted to the number of cells collected via FACS (Equation 1).

$$cells\ of\ sequence\ i\ in\ gate\ j = \frac{reads\ of\ sequence\ i}{total\ filtered\ reads\ in\ gate\ j} * number\ of\ cells\ collected\ in\ gate\ j \quad \text{(Eq 1)}$$

The assay score for a sequence was calculated by assigning each gate a score [0, 1/3, 2/3, 1] and determining the cell-averaged score (Equation M2). For on-yeast protease, 1 was given to full length sequences and 0 was given to fully digested sequences. For split GFP, 0 was given to no detected GFP signal and 1 was given to the highest amount of GFP signal.

$$score\ of\ sequence\ i = \frac{\sum_{j\ gates} cells\ of\ sequence\ i\ in\ gate\ j * score\ of\ gate\ j}{\sum_{j\ gates} cells\ of\ sequence\ i\ in\ gate\ j} \quad \text{(Eq 2)}$$

The final assay score was determined by the average score for a sequence in each trial. Sequences without reads in at least one gate per trial were removed from the dataset.

3.6.5.2 Split β-lactamase Assay Score Calculation

We aimed to assign an assay score that would correlate to the total activity of β-lactamase enzyme in each cell. We assumed that cells with active enzyme grown in ampicillin will retain the ability to grow and divide (and thus increase DNA frequency), whereas cells with inactive enzyme grown in ampicillin will stop growth (and thus prevent any increase in DNA frequency). To increase resolution, we chose ampicillin concentrations that produced approximately 10%, 30%, and 60% of uninhibited growth for each cell strain. Briefly, we estimated the max growth rate and determined the extra number

of doublings required to reach a given concentration. Assuming all cells are growing with no ampicillin, the relative number of dividing cells can be determined by the initial number of cells. The assay score for each sequence was determined by the relative change in read frequency with increasing ampicillin concentrations. For simplicity, the ampicillin concentrations were assigned to [0, 1, 2, 3] where 0 represented the no-ampicillin control and 3 represented the highest ampicillin concentration.

The final assay score was determined by the average score for a sequence in each trial. Sequences without a read in the no-ampicillin population in each trial were removed from the dataset. To scale the assay scores within the range [0,1], scores for *CC+* and *CC-* sequences (not including the independent Test sequences to prevent data leaking) were normalized via scikit-learn's quantile transformer with a normal output distribution followed by a minmax scaler.

3.6.6 Dot Blots to Quantify Expression

3.6.6.1 Production of Gp2 Library for Dot Blot

Frozen cells from deep well 96-well plates were scraped and seeded into 500 µL/well fresh LB+Kan and grown overnight (I^q was grown at 37 °C and SH was grown at 30 °C for the entire production). The following day, 25 µL/well of overnight culture was added to 1 mL/well of fresh LB+Kan and grown for 90 min. The protein production was induced by the addition of 0.5 mM IPTG (diluted in LB+Kan to add 100 µL/well). Production was continued for 2 h (I^q) or 4 h (SH) followed by centrifugation (3,000g for 5 minutes) and freezing of the cell pellet at -80 °C overnight. The pellet was thawed by the addition of 100 µL/well lysis buffer (only change is 0.1 mg/mL lysozyme) and shaken at 37 °C for 1 hour. The plates were centrifuged (3,000g for 5 min) and 25 µL/well of the soluble fraction was added to 25 µL/well of denaturing buffer Protein lysates from SH were

diluted an additional 5X in denaturing buffer to ensure signals were within the range of standards. The plates were incubated at 70 °C for 5 min to ensure denaturation and full accessibility of the His_6 tag.

3.6.6.2 Dot-Blot Protocol

A section of 0.2 µm pore polyvinylidene fluoride (PVDF, 1620177, BioRad) was cut to size and placed in a box (15.2 cm × 10.2 cm × 3.2 cm, Z742094, Sigma Aldrich). The membrane was soaked in 50 mL methanol for 30 s, followed by 50 mL dH_2O for 2 min. Finally, the membrane was equilibrated in 50 mL TBST (0.05% v/v Tween 20 in tris-buffered saline (TBS)) for 5 min. The membrane was then placed on a TBST soaked filter paper and padded dry with a Kimwipes™. Using a multichannel pipet, 2 µL/well of protein samples were added to the membrane and allowed to fully absorb. The membrane was then transferred to a dry filter paper and placed in a fume hood for 30 min until dry. The membrane was then placed back in the box with 50 mL blocking solution (5% (w/v) nonfat dry milk in TBST) and rocked overnight at 4 °C. The membrane was then labeled with 50 mL of 0.2 µg/mL anti His_6-HRP (ab1187, Abcam) in blocking solution for 30 minutes at room temperature. Excess antibody was washed via 3 washes of 50 mL TBST for 10 min at room temperature. The membrane was then soaked in 25 mL of SuperSignal™ West Pico PLUS Chemiluminescent Substrate (ThermoFisher) for 5 min. Then membrane was then placed inside a transparency and exposed 10-30 s on a ChemiDoc MP Imaging System (BioRad).

3.6.7 Identification of HT Assay Predictiveness

3.6.7.1 Code Availability

Python scripts used for deep-sequencing and model evaluation, as well as datasets to train, evaluate, and plot predictive performance are available at https://github.com/HackelLab-UMN/DevRep.

3.6.7.2 Cross-Validation Performance

A set of 195 unique Gp2 variants contained measured HT assay scores in all 10 assay conditions, and a yield in at least one of the strains. We performed 10 x 10 repeated K-fold cross-validation to determine which of the 1,023 combinations of HT assay conditions predicted the "left-out" set of sequences' yield with the least error. Each HT assay combination was evaluated for predictive performance on four different model architectures summarized in Table 1. We utilized the Hyperopt[153] library to determine the optimal hyperparameters for each architecture. We allowed 50 trials (or a maximum of 24 hours of computational time for FNN) and recorded the trial with the lowest predictive error.

Table 3.1 - Description of model architectures utilized when evaluating HT assay predictive performance

"Uniform" and "quniform" refers to stochastic search spaces defined in the Python *hyperopt* library[154].

Architecture	Description	Hyperparameter Space
Ridge	sklearn.linear_model.Ridge	10^{α}: uniform[-5,5]
Forest	sklearn.ensemble.RandomForestRegressor	n_estimators: quniform[1,500], max_depth: quniform[1,100], max_features: uniform[0,1]
SVM	sklearn.svm.SVR	10^{γ}: uniform[-3,3], 10^{C}: uniform[-3,3]
FNN	tf.keras.layers.Dense	10^epochs: uniform[0,2], batch size: quniform[10,200], hidden layers: quniform[0,4], nodes/hidden layer: quniform[1,100]

3.6.7.3 Test Performance

When evaluating performance on the independent test sequences, the best model architecture and hyperparameters were chosen by cross-validation, but the weights for the model were refit utilizing the entire cross-validation training set. The independent test set was not used in training data transformations or models.

3.6.7.4 Correlation Feature Selection (CFS)

CFS identifies the optimal feature set by maximizing the relationship between features (x, HT assays) and target (y, yield) while minimizing the inter-feature relationships[147]. We calculated the CFS for every set (S_x) of 1023 HT assay combinations. We defined the relationship (r) as the absolute value of Spearman's rank correlation coefficient (ρ) or the mutual information (MI) to capture linear and nonlinear relationships (Equation 3).

$$CFS\ (S_x) = \frac{\sum_{x=1}^{k}(r_{y_{Iq},f_x} + r_{y_{SH},f_x})}{\sqrt{k + 2\ \sum_{x=1}^{k-1}\left(\sum_{z=x+1}^{k}\left(r_{f_x,f_z}\right)\right)}} \qquad \text{(Eq 3)}$$

3.6.7.5 Subsampling Training Data

When evaluating the predictive performance of assays with varying number of training datapoints, we bootstrapped the dataset for cross-validation ten times. Each random dataset had separately optimized architectures and hyperparameters determined by cross-validation. Due to the computational constraints, FNN architecture was not evaluated when subsampling the training dataset.

3.6.7.6 Propagation of Uncertainty

Calculations involving propagation of uncertainty for predicted sample size were performed using the uncertainties[155] Python package.

3.7 Acknowledgements

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 thank Daniel Woldring for useful feedback on the manuscript. We appreciate support from the University of Minnesota Flow Cytometry Core, University of Minnesota Genomics Center, and the Minnesota Supercomputing Institute (MSI) at the University of Minnesota.

3.8 Author contributions

A.W.G., K.M.M, and B.J.H. designed experiments. A.W.G., K.M.M, S.L., N.L.N., M.F., and H.P. performed experimental methods. A.W.G., S.M., and B.J.H. analyzed, interpreted, and wrote the manuscript.

3.9 Supplemental Materials and Methods

3.9.1 Library Generation and Selection

3.9.1.1 Gp2 Insert Preparation

The Gp2 libraries for high-throughput (HT) assays were created via polymerase chain reaction (PCR) overlap extension on oligonucleotides purchased from Integrated DNA Technologies. PCR conditions: 0.02 U/µL Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs), 1X Q5 reaction buffer, 200 µM dNTPs, 0.5 µM primers (SI Appendix, DNA Table A: 1 & 17), 0.05 µM internal single stranded template (SI Appendix, DNA Table A: 2, 3, 4, 11, 15, 16), 0.008 µM degenerate strands for loop 1 (SI Appendix, DNA Table A: 5-10), 0.017 µM degenerate strands for loop 2 (SI Appendix, DNA Table A: 12-14) diluted to 50 µL with deionized water (dH_2O). Thermocycling routine: 98 °C for 30 s, (98 °C for 10 s, 59 °C for 30 s, 72 °C for 20 s) x 30 cycles, 72 °C for 120 s. The 300 bp resulting DNA was then gel extracted from a 2% agarose gel and

purified via silicon spin column (Epoch Life Sciences) eluting with 30 µL dH_2O per manufacturer's instructions.

The constructed library was then amplified for electroporation. PCR conditions: 0.02 U/µL Phusion High-Fidelity DNA Polymerase (New England Biolabs), 1X Phusion HF buffer, 200 µM dNTPs, 0.5 µM primers (SI Appendix, DNA Table A: 1 & 17), 15 µL purified template, diluted to 400 µL total volume with dH_2O. Thermocycling routine: 98 °C for 30 s, (98 °C for 10 s, 66 °C for 30 s, 72 °C for 20 s) x 35 cycles, 72 °C for 120 s. The DNA was concentrated and purified via ethanol precipitation: 40 µL 3 M sodium acetate at pH 5.2 and 1200 µL ethanol was added to the post-PCR product and the mixture was then incubated at 4 °C for 10 min. The insoluble DNA was pelleted via centrifugation at 15,000g for 20 min at 4 °C. The DNA was then washed with 1 mL 70% ethanol in dH_2O, centrifuged, washed with 1 mL ethanol, centrifuged, aspirated, and dried overnight to R.T. air. The reaction was then resuspended in 30 µL of Buffer E' (0.5 M sorbitol and 0.5 mM calcium chloride in dH_2O).

3.9.1.2 Yeast Surface Display Plasmid Preparation

A yeast plasmid display vector pCT from Kruziki et al[26] was modified to contain a stop codon before the cMyc epitope tag to serve as a negative control (final construct: Aga2-HA-Stop-cMyc, Plasmid Sequence 1). A plasmid expressing the parental Gp2 programmed death-ligand 1 (PD-L1) binding clone E4 was restriction enzyme digested (2 µg plasmid, 20 U BamHI-HF, 20 U PstI-HF, 5U Quick CIP, 1X CutSmart Buffer, diluted to 50 µL with dH_2O and incubated at 37 °C for 1 hr), extracted from a 2% agarose gel, and purified via silica column eluted with 30 µL of dH_2O. The motif was inserted via NEBuilder® HiFi DNA Assembly (New England Biolabs): 35 ng of the digested vector,

1.8 ng template DNA (SI Appendix, DNA Table B), 5 µL HiFi Master Mix, diluted to 10 µL with dH_2O was incubated at 50 °C for 20 mins. 2 µL of the reaction mixture was added to 25 µL NEB 5-alpha Competent *E. coli* and transformed according to the manufacturer's protocol. The transformed cells were plated onto 100 µg/mL ampicillin LB agar plates. A colony was plucked, grown in LB (10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride) with 100 µg/mL ampicillin, and miniprepped to obtain 50 µg of plasmid. The plasmid was restriction enzyme digested (50 µg plasmid, 200 U BamHI-HF, 200 U PstI-HF, 50 U Quick CIP, 1X CutSmart Buffer, diluted to 500 µL with dH_2O, incubated at 37 °C for 2 h), and ethanol precipitated. The digested plasmid was reconstituted in 50 µL Buffer E'. The resulting concentration of vector was measured via absorbance with a NanoDrop (Thermo Fisher).

3.9.1.3 Yeast Transformation

The Gp2 gene library was inserted into the yeast display vector (final construct: Aga2–HA–GS linker–Gp2–cMyc) via homologous recombination into *S. cerevisiae* yeast (EBY100) (steps 36-48 in Chao *et al.*[110]) with the following modifications: step 37: inoculated 100 mL of culture, steps 38/41: all (50 µL) of Gp2 insert and 6 µg of digested plasmid was used per 100 mL culture, step 39: included 30% (v/v) PEG 8000 during incubation. Plating a dilution of the transformed cells on selective media estimated the creation of 3.6×10^8 variants of Gp2. Surface display was induced (Step 2[110]) by the introduction of galactose containing media followed by growth at 30 °C overnight.

3.9.1.4 Epitope Labeling for Yeast Flow Cytometry

Labeling of the HA and cMyc epitope tags were used to enrich the frequency of full length Gp2 variants while subsampling from the initial population via fluorescence-

activated flow cytometry (FACS). Ten million yeast cells were centrifuged at 5000g for 1 min. The induction media was removed, and the cell pellet was resuspended with 1 mL cold PBSA (1 g/L bovine serum albumin in 1X phosphate-buffered saline (PBS)). The cells were pelleted, aspirated, resuspended in 500 µL PBSA containing 0.5 µg of a chicken-anti-HA antibody (ab9111, Abcam) and 1 µg of a mouse-anti-cMyc antibody (9E10, Biolegend), and rotated for 30 min at R.T. The cells were then centrifuged, washed with 1 mL cold PBSA, and resuspended in 500 µL PBSA containing 0.5 µg of a goat-anti-chicken AlexaFluor488 antibody (A-11039, Invitrogen) and 1 µg of a goat-anti-mouse AlexaFluor647 antibody (A-21235, Invitrogen) and incubated at 4 °C for 20 min while protected from light. Finally, the cells were centrifuged, washed with 1 mL cold PBSA, and stored as a pellet after centrifugation until sorting. FACS was performed at the University of Minnesota Flow Cytometry Resource facilities. Cells were resuspended at 2 x 10^7 cells/mL in PBSA and 1 x 10^6 cells displaying positive 488 (HA) and 647 (cMyc) signals were collected. Perhaps resulting from multi-vector transformants[156], 6.7% of the sequences contained a stop codon in the paratope. Additional propagations were performed before completing high-throughput assays to mitigate this issue. Yeast expressing the GaR clone, obtained from Kruziki[71], were added to the subsampled population at an intended ratio of 100 GaR : 1 random variant from library (obtained 172:1 via sequencing).

3.9.2 On-Yeast Protease Assay

3.9.2.1 Yeast DNA Extraction

Frozen populations were thawed, and the DNA was obtained via Zymoclean Gel DNA Recovery Kit (Zymo Research) following the manufacturer's protocol. Following the elution into 30 µL of dH_2O, half of the DNA was mixed with 2 µL ExoI (M0293S, New England Biolabs), 1 µL of Lambda Exonuclease (M0262S, New England Biolabs)

and 2 µL of 10 X Lambda Exonuclease Buffer, incubated at 30 °C for 90 min to remove genomic DNA, and 80 °C for 20 min to inactivate the enzymes. The DNA was then purified via silica column purification and eluted with 30 µL dH_2O.

3.9.2.2 Preparation of DNA for Deep Sequencing

The DNA was prepared for Illumina sequencing and genetically barcoded for population identification by two successive PCR reactions. The first PCR specifically amplified the region of DNA encoding for Gp2: PCR conditions: 0.02 U/µL Q5 High-Fidelity DNA Polymerase (New England Biolabs), 1X Q5 reaction buffer, 200 µM dNTPs, 0.1 µM of 5-forward and 5-reverse primers to add length diversity for sequencing (SI Appendix, DNA Table C), 15 µL (half) of the DNA extracted from yeast, diluted to 50 µL total volume with dH_2O. Thermocycling routine: 98 °C for 30 s, (98 °C for 10 s, 60 °C for 30 s, 72 °C for 20 s) x 16 cycles, 72 °C for 120 s. Unreacted primers were then removed by the addition of 4 U ExoI (37 °C for 30 min, inactivated at 80 °C for 20 min). The second PCR added trial-specific I5 barcode and a gate-specific I7 barcode. PCR conditions: 0.02 U/µL Q5 High-Fidelity DNA Polymerase, 1X Q5 reaction buffer, 200 µM dNTPs, 0.5 µM of forward primer (SI Appendix, DNA Table D) and reverse primer (SI Appendix, DNA Table E), 1 µL of the DNA from the first PCR, diluted to 50 µL total volume with dH_2O. Thermocycling routine: 98 °C for 30 s, (98 °C for 10 s, 67 °C for 30 s, 72 °C for 20 s) x 16 cycles, 72 °C for 120 s. The DNA was purified via agarose gel extraction and quantified via absorbance on a NanoDrop. DNA within the same assay was mixed at the ratio of cells collected during FACS. DNA across assays were evenly mixed for each trial.

3.9.3 Split GFP Assay

3.9.3.1 Creation of GFP1-10 Bacterial Production Plasmid

Plasmid pcDNA3.1-GFP(1-10) was a gift from Bo Huang[157] (Addgene plasmid 70219). The fragment encoding for GFP_{1-10} was isolated via PCR. PCR conditions: 0.02 U/µL Q5 High-Fidelity DNA Polymerase, 1X Q5 reaction buffer, 200 µM dNTPs, 0.5 µM forward and reverse primers (SI Appendix, DNA Table F), 1 ng pcDNA3.1-GFP(1-10), diluted to 50 µL with dH_2O. Thermocycling routine: 98 °C for 30 s, (98 °C for 10 s, 72 °C for 50 s) x 30 cycles, 72 °C for 120 s. The DNA was then purified via silica column.

Plasmid pBAD-His-6-Sumo-TEV-LIC cloning vector (8S) was a gift from Scott Gradia (Addgene plasmid 37507). The plasmid was modified via restriction enzyme digestion (final construct: His_6-GFP_{1-10}, Plasmid Sequence 2). Digestion conditions: 2 µg plasmid, 20 U NheI-HF, 20 U BamHI-HF, 5U Quick CIP, 1X CutSmart Buffer, diluted to 50 µL with dH_2O and incubated at 37 °C for 1 h. The plasmid was isolated via agarose gel extraction and silica column purification.

GFP_{1-10} was inserted into the pBAD plasmid via NEBuilder® HiFi DNA Assembly (New England Biolabs): 25 ng of the digested vector, 2 ng of GFP_{1-10} encoding DNA, 5 µL HiFi Master Mix, diluted to 10 µL with dH_2O and was incubated at 50 °C for 20 min. The assembled plasmid was transformed into NEB 5-alpha Competent *E. coli* as per the manufacturer's protocol using the ampicillin selection marker.

3.9.3.2 Creation of GFP11 Production Plasmid

A pET production plasmid was obtained[134] and modified to serve as a non-fluorescent control with a stop codon before the C-terminal GFP_{11} (final construct: MAS–Stop–GSGGGGS–GFP_{11}, Plasmid Sequence 3). Two rounds of restriction enzyme digestion and HiFi assembly processes were used to complete construction. Digestion 1: 2

µg plasmid, 20 U NheI-HF, 20 U StyI, 5U Quick CIP, 1X CutSmart Buffer, diluted to 50 µL with dH_2O and incubated at 37 °C for 1 h The plasmid was isolated via agarose gel extraction and silica column purification. HiFi assembly 1: 30 ng plasmid, 2 ng pET-GFP11 gBlock (SI Appendix, DNA Table G), 5 µL HiFi Master Mix, diluted to 10 µL with dH_2O, incubated at 50 °C for 20 min, and transformed using the kanamycin selection marker. Digestion 2: 2 µg plasmid, 20 U NheI-HF, 20 U BamHI-HF, 5U Quick CIP, 1X CutSmart Buffer, diluted to 50 µL with dH_2O and incubated at 37 °C for 1 h. The plasmid was isolated via agarose gel extraction and silica column purification. HiFi assembly 1: 30 ng plasmid, 2 ng GFP11-stop insert (SI Appendix, DNA Table G), 5 µL HiFi Master Mix, filled to 10 µL with dH_2O, incubated at 50 °C for 20 min, and transformed using kanamycin selection marker.

3.9.3.3 Ligation of Gp2 Library into GFP_{11} Production Plasmid

The Illumina prepared DNA resulting from the on-yeast protease assay (equal mixture of 6 reaction conditions of trial 1) was used as the source of the Gp2 library for the split GFP assay. The DNA was prepared for ligation via restriction enzyme digest. Digestion conditions: 1.25 µg DNA, 25 U NheI-HF, 25 U BamHI-HF, 1X CutSmart Buffer, filled to 62.5 µL with dH_2O and incubated at 37 °C for 1 h. The digested DNA was isolated via agarose gel extraction and silica column purification. All pre-ligation gel extractions used Zymoclean Gel DNA Recovery Kits (Zymo Research). The GFP_{11} plasmid was prepared for ligation in a similar process. Digestion conditions: 10 µg DNA, 200 U NheI-HF, 200 U BamHI-HF, 50 U CIP, 1X CutSmart buffer, diluted to 500 µL with dH_2O and incubated at 37 °C for 1 hour.

Five ligations were required to obtain more than 10^6 transformed colonies, providing 63% likelihood of sampling any clone of the subsampled library. Ligation conditions: 370 ng vector, 30 ng insert, 10,000 U T4 DNA Ligase (New England Biolabs), 1X T4 Buffer, 1 mM ATP (New England Biolabs), filled to 100 µL with dH_2O prepared on ice. The reaction was mixed via gentle pipetting and incubated at 22 °C for 15 min, followed by ligase deactivation via incubation at 60 °C for 10 min. The ligated DNA was purified and concentrated into 10 µL dH_2O via MinElute PCR Purification Kit (Qiagen). The plasmids were transformed into NEB 5-alpha Electrocompetent *E. coli* following the manufacturer's protocol using 2.5 µL of DNA per 25 µL of cells. An average of 2 x 10^5 transformed cells was obtained per 100 µL ligation plated on LB agar plates containing 50 mg/L kanamycin. Colonies were scraped from plates and miniprepped to transfer the DNA to production cell lines.

3.9.3.4 Transformation of Split-GFP Production Cells

The GFP_{1-10} plasmid was transformed into T7 Express lysY/I^q Competent *E. coli* (I^q, c3013, New England Biolabs) and SHuffle T7 Express lysY Competent *E. coli* (SH, c3030, New England Biolabs) following the manufacturer's heat-shock protocol and using the ampicillin selection marker. A single colony from each bacterial strain was plucked and prepared for electroporation: the colony was grown in 100 mL SOB + Amp (2% tryptone, 0.5% yeast extract, 10 mM sodium chloride, 2.5 mM potassium chloride, 10 mM magnesium chloride, 10 mM magnesium sulfate, and 100 mg/L ampicillin in dH_2O) to an optical density at 600 nm (OD_{600}) of 0.5. Unless otherwise stated, strain I^q was grown at 37 °C and strain SH was grown at 30 °C. The culture was then placed on wet ice for 15 min and centrifuged (5000g for 10 min). The cells were then resuspended and centrifuged twice

with 200 mL of 10% (v/v) glycerol in water. Finally, the cells were resuspended in the residual glycerol before flash freezing with liquid nitrogen and storage at -80 °C.

The Gp2-GFP_{11} plasmids were then electroporated into the prepared competent cells. Frozen cells were thawed on wet ice for 10 min. 20 ng of the plasmid was added to 25 µL of cells and transferred to a cold 1 mm electroporation cuvette. The cells were shocked (2.0 kV, 200 Ω, 25 µF), resuspended in 975 µL SOC (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose in dH_2O), and incubated for 1 h. The cells were plated on selective LB agar (containing 100 mg/L ampicillin, and 50 mg/L kanamycin in dH_2O). Transformations were repeated until $>10^7$ colonies were obtained. The colonies were scraped from plates and grown in 100 mL LB with 100 mg/L ampicillin and 100 mg/L kanamycin (LB+Amp+Kan) for two h. Aliquots were created by mixing 1 mL culture with 500 µL glycerol and storing at -80 °C.

3.9.3.5 Preparation of DNA for Deep Sequencing

DNA was prepared for Illumina as above for the on-yeast protease assay: except for unique primers (SI Appendix, DNA Table H) and a 62 ºC annealing temperature in the first PCR.

3.9.4 Split β-lactamase Assay

3.9.4.1 Creation of Production Plasmid

A pET production plasmid was obtained in house and modified via two rounds of restriction digest and HiFi assembly to create a non-functional lactamase control with a stop codon before the second half of the enzyme (final construct: β-$lactamase_{1-196}$-$(G_4S)_2$AS-Stop-GS$(G_4S)_2$-β-$lactamase_{197-286}$, Plasmid Sequence 4). Digestion 1 conditions: 2 µg plasmid, 20 U NdeI-HF, 20 U StyI-HF, 5 U Quick CIP, 1X CutSmart buffer, diluted

to 50 µL with dH_2O and incubated at 37 °C for 1 h), gel extracted, and purified via silica column. HiFi assembly 1: 25 ng plasmid, 2 ng of each insert (Beta-Lac A+B, SI Appendix, DNA Table I), 10 µL HiFi Master Mix, filled to 20 µL with dH_2O, reacted at 50 °C for 15 min, and transformed using kanamycin selection marker. Digestion 2 conditions: 2 µg plasmid, 20 U NheI-HF, 20 U BamHI-HF, 5 U Quick CIP, 1X CutSmart Buffer, filled to 50 µL with dH_2O and incubated at 37 °C for 1 h), gel extracted, and purified via silica column. HiFi assembly 1: 25 ng plasmid, 2 ng insert (β stop insert, SI Appendix, DNA Table I), 10 µL HiFi Master Mix, diluted to 20 µL with dH_2O, reacted at 50 °C for 15 min, and transformed using kanamycin selection marker. Plasmid for ligation was obtained via miniprep.

3.9.4.2 Split β-lactamase Library Creation and Transformation into Production Cells

The Illumina prepared DNA resulting from the split GFP assay (equal mixture of 2 cell strains from trial 1) was used as the source of the Gp2 library. The DNA was prepared for ligation via restriction enzyme digest. Digestion conditions: 2 µg DNA, 20 U NheI-HF, 20 U BamHI-HF, 1X CutSmart Buffer, diluted to 50 µL with dH_2O and incubated at 37 °C for 1 h. The digested DNA was isolated via agarose gel extraction and silica column purification. All pre-ligation gel extractions utilized Zymoclean Gel DNA Recovery Kits (Zymo Research). The β-lactamase plasmid was prepared for ligation in a similar process. Digestion conditions: 10 µg DNA, 200 U NheI-HF, 200 U BamHI-HF, 50 U CIP, 1X CutSmart Buffer, diluted to 500 µL with dH_2O and incubated at 37 °C for 1 h.

Ligations were repeated, using the same conditions as for the split GFP pool above, to obtain more than 10^6 transformed colonies. Transformed cells were plated on LB agar plates containing 50 mg/L kanamycin. Colonies were scraped from plates and miniprepped

to transfer the DNA to production cell lines. The library was transformed into I^q and SH following the manufacturer's heat-shock protocol and using the kanamycin selection marker. Transformations were replicated until more than 10^7 colonies were obtained. The colonies were scraped from plates and grown in 100 mL LB with 100 mg/L kanamycin for two h. Aliquots were created by mixing 1 mL culture with 500 µL glycerol and storing at -80 °C.

3.9.4.3 Preparation of DNA for Deep Sequencing

DNA was prepared for Illumina as above for the protease assay except for unique primers (SI Appendix, DNA Table J) and a 59 ºC annealing temperature in the first PCR.

3.9.5 High-Throughput Assay Score Calculations

3.9.5.1 Illumina Sequencing and Read Filtering

The prepared DNA from each assay was sequenced via two SP lanes of an NovaSeq 6000 (Illumina) with the help of the University of Minnesota Genomics Center. The first trial for all assays was sequenced in the first lane, and the second and third trials were equally mixed in a second lane after confirming the preliminary success of the first trial.

Sequence analysis was performed using the computational resources of the Minnesota Supercomputing Institute utilizing USearch[116] to merge, align, filter, denoise, and dereplicate the sequences. Merged reads were clipped to the region between NheI and BamHI prior to quality filtering, where we accepted sequences with less than one expected error based upon the reported quality scores of each nucleotide. A total of 832 x 10^6 sequences passing filter were obtained (Trial 1: 434 x 10^6, Trial 2: 211 x 10^6, Trial 3: 188 x 10^6).

A contamination of DNA encoding for Gp2 with a PD-L1 binding paratope and framework mutations was experimentally confirmed in the on-yeast protease assay

sequencing primer stocks. Thus, these sequences were removed from the on-yeast assays but were included in the split GFP and split β-lactamase assays as the sequences were obtained from the on-yeast DNA.

Beyond the physical contamination, an average of 56 x 10^6 unique sequences was obtained per trial, well beyond the expected max diversity of 1 x 10^6. We hypothesize the "true" sequences to have high observation frequency (most number of reads) and contain highly different sequences compared to "false" sequences (due to subsampling of theoretical library, it is unlikely to see two very similar sequences). We denoised each trial independently utilizing the UNOISE[158] algorithm with observation minimums chosen for computational efficiency (Trial 1: 100 total reads, Trial 2 and 3: 50 total reads). A total of 294,644 unique sequences observed in all three trials were obtained from denoising. We then mapped the filtered false sequences to the true sequence via 97% genetic similarity. Finally, 204,173 unique sequences for *CC+*¸*CC-*, and *Stop* were isolated via requiring 100% genetic match of the conserved portion of Gp2. *CC+* was identified via two cysteines located at each end of loop 1 (positions 7 & 12), whereas other sequences were classified *CC-*. *Stop* sequences contained at least one stop codon located inside either loop, but otherwise matched library design. No sequences containing synonymous codons were observed during the measurement of recombinant yield.

3.9.6 Dot Blots to Quantify Expression

3.9.6.1 Creation of Production Plasmid

A pET production plasmid was obtained in-house and modified via restriction digest and HiFi assembly to create a His_6-less negative control with a stop codon placed between the restriction sites. (final construct: V5-AS-Stop-GS-His_6, Plasmid Sequence 5). Digestion conditions: 2 µg plasmid, 20 U NheI-HF, 20 U BamHI-HF, 5 U Quick CIP, 1X

CutSmart Buffer, diluted to 50 µL with dH_2O and incubated at 37 °C for 1 h), gel extracted, and purified via silica column. HiFi assembly 1: 25 ng plasmid, 2 ng insert (SI Appendix, DNA Table K), 10 µL HiFi Master Mix, diluted to 20 µL with dH_2O, incubated at 50 °C for 15 min, and transformed utilizing kanamycin selection marker. Plasmid for ligation was obtained via miniprep.

3.9.6.2 Dot Blot Library Creation and Transformation into Production Cells

DNA encoding for Gp2 variants for the dot blots were obtained from two sources: i) the Illumina prepared DNA resulting from the split β-lactamase assay (equal mixture of the no-ampicillin population from both cell strains in trial 1). ii) an oligopool (Oligopool.fasta, Twist Bioscience) encoding for the 2,000 most frequently observed Gp2 variants found in all 10 assays of trial 1. The oligopool was amplified to create double stranded DNA: PCR conditions: 0.02 U/µL Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs), 1X Q5 reaction buffer, 200 µM dNTPs, 0.5 µM primers (SI Appendix, DNA Table L), 10 ng of the oligopool resuspended at 1 ng/µL in EB (Epoch Life Science), diluted to 50 µL total volume with dH_2O. Thermocycling routine: 98 °C for 30 s, (98 °C for 10 s, 61 °C for 30 s, 72 °C for 20 s) x 12 cycles, 72 °C for 120 s. The amplified oligopool was silica column purified eluting with 50 µL dH_2O.

The DNA from each source was separately prepared for ligation via restriction enzyme digest. Digestion conditions: 2 µg DNA, 20 U NheI-HF, 20 U BamHI-HF, 1X CutSmart Buffer, filled to 50 µL with dH_2O and incubated at 37 °C for 1 h. The digested DNA was isolated via agarose gel extraction and silica column purification. All pre-ligation gel extractions used Zymoclean Gel DNA Recovery Kits (Zymo Research). The His_6 production plasmid was prepared for ligation in a similar process. Digestion conditions: 2

µg DNA, 20 U NheI-HF, 20 U BamHI-HF, 1X CutSmart Buffer, diluted to 50 µL with dH_2O and incubated at 37 °C for 1 hr.

Ligations were repeated to obtain more than 10^3 transformed colonies, to obtain a feasible testable diversity of sequences. Ligation conditions: 37 ng vector, 3 ng insert, 1,000 U T4 DNA Ligase (New England Biolabs), 1X T4 Buffer, 1 mM ATP (New England Biolabs), filled to 10 µL with dH_2O prepared on ice. The reaction was mixed via gentle pipetting and incubated at 22 °C for 15 min, followed by ligase deactivation via incubation at 60 ° C for 10 min. The plasmids were heat-transformed into NEB 5-alpha Competent *E. coli* following the manufacturer's protocol and using 2.5 µL of DNA per 25 µL of cells. Transformed cells were plated on LB agar plates containing 50 mg/L kanamycin. Colonies were scraped from plates and miniprepped to transfer the DNA to production cell lines. The library was transformed into I^q and SH following the manufacturer's heat-shock protocol and using the kanamycin selection marker. Transformations were replicated until more than 10^3 colonies were obtained. Single colonies were grown in 1 mL LB+Kan for two hours in separate wells of a deep 96 well plate. Aliquots were created by mixing 1 mL culture with 500 µL glycerol and storing at -80 °C.

3.9.6.3 Identifying Plate Location of Gp2 Variants

We appended genetic barcodes representing the plate, row, and column via PCR while simultaneously preparing the sequences for Illumina sequencing. PCR 1 conditions: 0.02 U/µL Q5 High-Fidelity DNA Polymerase (New England Biolabs), 1X Q5 reaction buffer, 200 µM dNTPs, 0.5 µM of a row-specific forward primer (SI Appendix, DNA Table M FApETV5N50X), 0.5 µM of a column-specific forward primer (SI Appendix, DNA Table M RApETN7XX), 1 µL bacterial culture, filled to 20 µL total volume with dH_2O.

Thermocycling routine: 98 °C for 5 min, (98 °C for 10 s, 61 °C for 30 s, 72 °C for 20 s) x 16 cycles, 72 °C for 120 s. The DNA for each plate was pooled at equal volume (2 µL) and the unreacted primers were then removed by the addition of 8 U ExoI (37 °C for 30 minutes, inactivated 80 °C for 20 min). The second PCR added a plate specific barcode. PCR conditions: 0.02 U/µL Q5 High-Fidelity DNA Polymerase, 1X Q5 reaction buffer, 200 µM dNTPs, 0.5 µM of forward primer (SI Appendix, DNA Table D) and reverse primer (SI Appendix, DNA Table E), 1 µL of the DNA from the first PCR, filled to 50 µL total volume with dH_2O. Thermocycling routine: 98 °C for 30s, (98 °C for 10 s, 67 °C for 30 s, 72 °C for 20 s) x 16 cycles, 72 °C for 120 s. The DNA was isolated via agarose gel extraction and quantified via NanoDrop. DNA across plates and cell strain were equally mixed and sequenced via Illumina iSeq, aiming to obtain ~1,000 reads per well.

Sequence analysis was performed using the computational resources of the Minnesota Supercomputing Institute utilizing USearch[116] to merge, align, filter, denoise, and dereplicate the sequences. Merged reads were clipped to the region between NheI and BamHI prior to quality filtering, where we accepted sequences with less than one expected error based upon the reported quality scores of each nucleotide. To identify single-variant wells, the most abundant sequence had to have >100 reads, the next most sequence had to have <100 reads, and the top sequence had to occupy > 80% (DNA from β-lactamase) or > 40% (DNA from oligopool, changed based upon resequencing previously identified variants) of the total reads for a well. Sequences obtained from dot blot sequencing were genetically paired (requiring 100% matching) with sequences from HT assays.

3.9.6.4 Preparation of Protein Standard

GaR was separately ligated into the production vector (final construct V5-*GaR*-His_6) and transformed following the same protocol used for the library of Gp2 variants, with the difference of transforming the post-ligation product directly into I^q via heat-shock transformation. A scrape from the frozen stock of *GaR* cells was grown in 5 mL LB+Kan overnight. Part of the overnight culture was added to 200 mL fresh LB+Kan at an OD_{600} of 0.1 and grown for 90 min. The protein production was induced by the addition of 0.5 mM IPTG. Production was continued for 2 h at 37 °C followed by centrifugation (3,000 g for 15 min) and freezing of the cell pellet at -80 °C overnight. The pellet was thawed by the addition of 2 mL lysis buffer: 1 mg/mL lysozyme (L6876, Millipore Sigma), 10 U/mL benzonase nuclease (E1014, Millipore Sigma), protease inhibitor pellet (A32953, Thermo Fisher Scientific), 20 mM sodium chloride, 2 mM magnesium chloride, 25 mM imidazole, 5 mM 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate hydrate (CHAPS), 5% (v/v) glycerol, 50 mM (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) (HEPES) in dH_2O at pH 8.0. The lysate and lysis buffer were shaken at 37 °C for 1 h to promote enzymatic activity. The soluble lysate was isolated via centrifugation (15,000g for 10 min) and filtered through a 0.22 µm membrane. *GaR* was isolated via immobilized metal affinity chromatography utilizing HisPur Cobalt Resin (89964, ThermoFisher Scientific); wash buffer: 500 mM sodium chloride, 20 mM HEPES, 20 mM imidazole, pH 7.4; elution buffer: as in wash but 500 mM imidazole. The protein was desalted on a PD-10 column (Fisher Scientific, eluted into 0.5 M sodium chloride, 20 mM HEPES, pH 7.4). The identity and purity were confirmed via matrix-assisted laser desorption/ionization (MALDI) and polyacrylamide gel electrophoresis (PAGE), and concentration was determined via 280 nm absorbance on a NanoDrop. The protein was diluted into 4 standard concentrations (103

ng/µL, 52 ng/µL, 26 ng/µL, and 13 ng/µL) and flash frozen in aliquots via liquid nitrogen and stored at -80 °C. On the day of use, the aliquot was thawed and 25 µL of protein was mixed with 25 µL of denaturing buffer (1 g/L SDS, 500 mM sodium chloride, 20 mM imidazole, 20 mM HEPES, pH 7.4) and incubated at 70 °C for 5 min.

3.9.6.5 Quantification of Chemiluminescent Intensities

Intensity measurements were quantified utilizing Fiji[159,160]. The average intensity of a constant diameter of a circular region of interest for each lysate impression was recorded. Ten randomly chosen background locations were also measured and subtracted from the intensity measurements. A row of standards (4 concentrations, each concentration in triplicate) was used to generate a linear standard curve from average intensity to yield (mg/L). To correct for non-specific binding of *E. coli* lysate proteins (not present in standard curve), the 75th percentile value of yield for wells containing *Stop* sequences in each trial was subtracted. Sequences with negative corrected yields were set to 0 mg/L. The final yield for model evaluation was reported as the average of three yield measurements, grown from separate starter cultures on different days. *GaR* was tested on each plate for both cell types to obtain an estimate of variance. It was observed that day-to-day coefficient of variation (I^q: 43%, SH: 77%) was higher than plate-to-plate variation (I^q: 20%, SH: 25%).

3.9.7 Identification of HT Assay Predictiveness

3.9.7.1 Sequence Encoding

To create models with the amino acid sequence, we only considered the amino acids in the modified paratope loops. To conserve possible interactions with the first/last position and the conserved residues of the protein, gap characters were placed in the middle of loops

during sequence alignment. The gap character was treated as a 21st amino acid during the one-hot encoding of the sequence.

3.10 Supplemental Figures

	Reaction Conditions	FACS Gate	Barcode for sequencing	Cells Collected			Filtered Reads			Reads/Cell			Score of Barcode		
				Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
On-Yeast Protease	**P_PK37**	P4	RPI 1	720,695	2,410,750	2,844,812	1,672,486	4,519,546	4,036,169	2.3	1.9	1.4	1.00	1.00	1.00
	2.7 U/L Proteinase K	P5	RPI 2	2,368,334	1,007,885	1,253,317	266,178	2,644,637	2,313,456	0.1	2.6	1.8	0.00	0.00	0.00
	1g/L PBSA	P6	RPI 3	2,463,131	997,983	1,168,086	6,940,582	1,178,588	2,284,648	2.8	1.2	2.0	0.67	0.33	0.67
	37°C	P7	RPI 4	1,931,957	1,822,412	574,143	4,607,090	3,110,116	1,311,519	2.4	1.7	2.3	0.33	0.67	0.33
	DF: 'Sort1'	Trial_Total		7,484,117	6,239,030	5,840,358	13,486,336	11,452,887	9,945,792	1.8	1.8	1.7			
	P_Urea	P4	RPI 5	685,873	999,145	2,351,724	1,681,542	2,677,268	3,745,159	2.5	2.7	1.6	1.00	1.00	1.00
	8 U/L Proteinase K	P5	RPI 6	2,285,755	1,991,598	2,069,329	8,201,646	4,851,346	3,974,228	3.6	2.4	1.9	0.00	0.00	0.00
	1g/L PBSA + 1.5M Urea	P6	RPI 7	2,484,865	1,703,321	1,350,940	9,701,969	3,548,318	1,533,413	3.9	2.1	1.1	0.67	0.33	0.67
	37°C	P7	RPI 8	2,001,220	1,578,136	890,976	6,435,191	2,119,653	643,801	3.2	1.3	0.7	0.33	0.67	0.33
	DF: 'Sort2'	Trial_Total		7,457,713	6,272,200	6,662,969	26,020,348	13,196,585	9,896,601	3.5	2.1	1.5			
	P_Gdn	P4	RPI 9	965,069	2,172,462	3,354,329	2,073,925	5,056,018	5,346,284	2.1	2.3	1.6	1.00	1.00	1.00
	8 U/L Proteinase K	P5	RPI 10	2,232,193	854,928	1,503,997	6,012,148	2,463,237	3,034,009	2.7	2.9	2.0	0.00	0.00	0.00
	1g/L PBSA + 0.5M Guan. Cl	P6	RPI 11	2,677,672	953,291	1,443,133	8,646,657	2,644,156	2,322,642	3.2	2.8	1.6	0.67	0.33	0.67
	37°C	P7	RPI 12	1,983,694	1,696,519	779,361	5,401,756	4,767,879	1,197,442	2.7	2.8	1.5	0.33	0.67	0.33
	DF: 'Sort3'	Trial_Total		7,858,628	5,677,200	7,080,820	22,134,486	14,931,290	11,900,377	2.8	2.6	1.7			
On-Yeast Protease	**P_PK55**	P4	RPI 13	1,038,298	975,667	3,818,691	4,074,116	2,970,629	5,816,002	3.9	3.0	1.5	1.00	1.00	1.00
	2.7 U/L Proteinase K	P5	RPI 14	1,630,037	1,502,970	1,199,931	6,983,125	4,999,760	1,768,760	4.3	3.3	1.5	0.00	0.00	0.00
	1g/L PBSA	P6	RPI 15	2,811,130	1,500,134	1,583,723	11,636,056	3,663,161	2,760,792	4.1	2.4	1.7	0.67	0.33	0.67
	55°C	P7	RPI 16	1,891,089	1,722,690	778,051	6,224,207	3,513,008	1,154,484	3.3	2.0	1.5	0.33	0.67	0.33
	DF: 'Sort4'	Trial_Total		7,370,554	5,701,461	7,380,396	28,917,504	15,146,558	11,500,038	3.9	2.7	1.6			
	P_TL55	P4	RPI 17	673,921	642,686	1,365,763	888,416	1,539,959	938,231	1.3	2.4	0.7	1.00	1.00	1.00
	0.5 mg/L Thermolysin	P5	RPI 18	2,373,381	1,195,457	2,035,251	6,253,572	2,666,639	3,414,867	2.6	2.2	1.7	0.00	0.00	0.00
	1g/L PBSA	P6	RPI 19	2,377,624	1,614,995	2,244,466	10,723,597	5,647,975	4,096,671	4.5	3.5	1.8	0.67	0.33	0.67
	55°C	P7	RPI 20	2,158,807	2,509,968	1,331,840	7,945,702	5,549,800	2,015,879	3.7	2.2	1.5	0.33	0.67	0.33
	DF: 'Sort5'	Trial_Total		7,583,733	5,963,106	6,977,320	25,811,287	15,404,373	10,465,648	3.4	2.6	1.5			
	P_TL75	P4	RPI 21	831,137	1,789,334	3,515,289	2,373,154	5,432,257	5,463,062	2.9	3.0	1.6	1.00	1.00	1.00
	0.67 mg/L Thermolysin	P5	RPI 22	678,048	765,350	439,280	1,824,231	2,117,123	1,073,922	2.7	2.8	2.4	0.00	0.00	0.00
	1g/L PBSA	P6	RPI 23	3,440,571	2,473,794	1,631,852	19,234,215	7,899,114	3,420,569	5.6	3.2	2.1	0.67	0.67	0.67
	75°C	P7	RPI 24	1,766,041	911,971	370,949	7,565,981	2,379,129	759,692	4.3	2.6	2.0	0.33	0.33	0.33
	DF: 'Sort6'	Trial_Total		6,715,797	5,940,449	5,957,370	30,997,581	17,827,623	10,717,245	4.6	3.0	1.8			

	Cell Strain	FACS Gate	Barcode for sequencing	Cells Collected			Filtered Reads			Reads/Cell			Score of Barcode		
				Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
Split GFP	**G_Iq**	P4	RPI 25	7,663,303	4,546,079	5,763,479	25,814,818	8,996,792	7,589,725	3.4	2.0	1.3	0.00	0.00	0.00
	T7 Express LysY/Iq	P5	RPI 26	392,858	493,224	683,276	1,171,571	555,408	748,363	3.0	1.1	1.1	0.33	0.33	0.33
	0.5mM IPTG, 2 g/L Arabinose	P6	RPI 27	326,900	958,124	1,027,122	974,591	1,249,940	691,451	3.0	1.3	0.7	0.67	0.67	0.67
	37°C	P7	RPI 28	103,978	441,186	300,025	563,859	893,081	236,098	5.4	2.0	0.8	1.00	1.00	1.00
	DF: 'Sort7'	Trial_Total		8,487,039	6,438,613	7,773,902	28,524,839	11,695,221	9,265,637	3.4	1.8	1.2			
Split GFP	**G_SH**	P4	RPI 29	5,016,070	4,725,240	6,257,965	13,362,917	9,054,947	7,218,077	2.7	1.9	1.2	0.00	0.00	0.00
	SHuffle T7 LysY	P5	RPI 30	1,776,832	1,061,176	1,058,452	4,281,977	2,140,059	1,048,734	2.4	2.0	1.0	0.33	0.33	0.33
	0.5mM IPTG, 2 g/L Arabinose	P6	RPI 31	1,937,258	1,138,897	762,107	7,697,521	2,646,845	1,109,272	4.0	2.3	1.5	0.67	0.67	0.67
	37°C	P7	RPI 32	622,872	258,253	347,798	993,264	420,294	486,686	1.6	1.6	1.4	1.00	1.00	1.00
	DF: 'Sort8'	Trial_Total		9,353,032	7,183,566	8,426,322	26,335,679	14,262,145	9,862,769	2.8	2.0	1.2			

	Cell Strain	[Ampicllin]	Barcode	Minutes Grown in Amp.			Filtered Reads			Reads/Diversity			Score of Barcode		
		µg/mL	for sequencing	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
Split β-lactamase	**β_Iq**	0	RPI 33	215	245	210	21,901,407	8,629,167	9,787,372	22	9	10	0	0	0
	T7 Express LysY/Iq	100	RPI 34	290	385	330	18,867,893	8,926,249	10,543,774	19	9	11	1	1	1
	0.5mM IPTG, 2 hr pre-induction	333	RPI 35	385	415	380	23,876,165	10,263,891	12,901,984	24	10	13	2	2	2
	37°C	1000	RPI 36	445	455	435	21,015,169	9,523,461	11,570,331	21	10	12	3	3	3
	DF: 'Sort9'	Trial_Total					85,660,634	37,342,768	44,803,461	86	37	45			
Split β-lactamase	**β_SH**	0	RPI 37	330	355	350	33,702,531	14,174,136	15,385,094	34	14	15	0	0	0
	SHuffle T7 LysY	100	RPI 38	465	405	530	21,877,032	15,752,351	13,703,305	22	16	14	1	1	1
	0.5mM IPTG, 4 hr pre-induction	666	RPI 39	740	690	770	20,017,792	14,048,228	13,405,895	20	14	13	2	2	2
	30°C	6666	RPI 40	1155	885	1075	34,248,595	15,550,720	14,403,817	34	16	14	3	3	3
	DF: 'Sort10'	Trial_Total					109,845,950	59,525,435	56,898,111	110	60	57			

Figure S3.1 - High-throughput (HT) Developability Assay Experimental Results

For each HT assay class, the varying conditions are listed along with the short-hand code and the name of the relevant DataFrame (DF) columns which store assay scores. On-yeast protease and split GFP FACS gates represent polygons named during experimentation. The split β-lactamase assay included a pre-induction of the enzyme-POI fusion before introduction of ampicillin. The Minutes Grown in Amp. represents the amount of time until the culture begins to exit logarithmic growth phase based upon preliminary assay testing. Filtered Reads represents the number of merged and aligned paired-end reads matching design and was used as the denominator when calculating read frequency. The scores for each barcode were utilized when calculating the cell-averaged score (on-yeast protease and split GFP) or when calculating the change in frequency as a function of ampicillin concentration (slope uses scores not actual concentration).

Assay	Total Sequences	Median Obs. per Trial	Average Standard Deviation	GaR Score ± SD (Observations)	Stop Median (Sequences)	% Stop < GaR (T-Test)	CC- Median (Sequences)	CC+ Median (sequences)	CC- < CC+ (MW-Test)
					Assay Score				
P_{PK37}	117,143	22	0.20	0.43±0.12 (1115)	0.08 (5418)	67%	0.47 (33593)	0.78 (78131)	p<0.001
P_{Urea}	122,624	24	0.18	0.40±0.03 (2995)	0.04 (6485)	81%	0.29 (37005)	0.67 (79133)	p<0.001
P_{Gdn}	121,359	24	0.19	0.61±0.28 (964)	0.06 (6379)	75%	0.44 (35848)	0.78 (79131)	p<0.001
P_{PK55}	128,639	24	0.19	0.44±0.17 (1096)	0.06 (6587)	71%	0.46 (38248)	0.76 (83803)	p<0.001
P_{TL55}	124,320	20	0.17	0.56±0.19 (1783)	0.05 (6389)	79%	0.44 (38255)	0.66 (79675)	p<0.001
P_{TL75}	100,455	25	0.29	0.41±0.11 (1989)	0.23 (5098)	35%	0.56 (29736)	0.62 (65620)	p<0.001
GFP_{Iq}	93,179	18	0.18	0.16±0.09 (410)	0.12 (7866)	8.4%	0.13 (27872)	0.13 (57440)	p=0.15
GFP_{SH}	140,300	14	0.18	0.27±0.03 (569)	0.10 (10516)	76%	0.19 (42828)	0.25 (86955)	p<0.001
β_{Iq}	94,675	16	0.07	0.21±0.02 (891)	0.45 (8727)	0.03% (99% >)	0.50 (26807)	0.50 (59140)	(>) p<0.001
β_{SH}	98,522	22	0.08	0.27±0.05 (791)	0.37 (9192)	1.2% (70% >)	0.52 (26825)	0.49(62504)	(>) p<0.001
					Yield (mg/L)				
$Yield_{Iq}$	432	N/A	1.32	7.6±1.6	0.00 (37)	100%	1.27 (98)	1.07 (296)	(=) p=0.40
$Yield_{SH}$	448	N/A	6.80	16.8±10	0.07 (46)	63%	2.52 (117)	5.54 (284)	p<0.001
					Transformed Yield (Yeo-Johnson Transformation, λ=-0.324)				
$Yield'_{Iq}$	432	N/A	0.42	0.87±0.15	-1.63 (37)	100%	-0.53 (98)	-0.64 (296)	(=) p=0.40
$Yield'_{SH}$	448	N/A	0.78	1.42±0.49	-1.25 (46)	91%	0.01 (117)	0.64 (284)	p<0.001

Figure S3.2 - Tabulated Gp2 Library Developability Performance

Total Sequences: all Stop, CC-, and CC+ sequences observed in all 3 trials per assay. *Median Observations per Trial* for on-yeast Protease and Split GFP assays measured observations as the number of cells collected in all gates per trial. Split β-Lactamase measured observations as the number of reads in the no-ampicillin control due to anticipated loss of sequences which fail to replicate in antibiotic-containing conditions. *Average Standard Deviation* represents the square root of the mean (by unique sequence) of experimental variance across 3 independent trials. *% Stop < GaR* is the percentage of unique stop codon sequences for which the assay score was significantly lower than the GaR score (one-way student's t-test, $p<0.05$). For β_{Iq} and β_{SH}, % Stop > GaR is also displayed. *CC- < CC+* tests the hypothesis that sequences with cysteines at positions 7 and 12 significantly increase assay score (one-way Mann-Whitney U test). For β_{Iq} and β_{SH}, the opposite hypothesis, CC- > CC+, is displayed. For $Yield_{Iq}$, the two-sided Mann-Whitney tested is displayed.

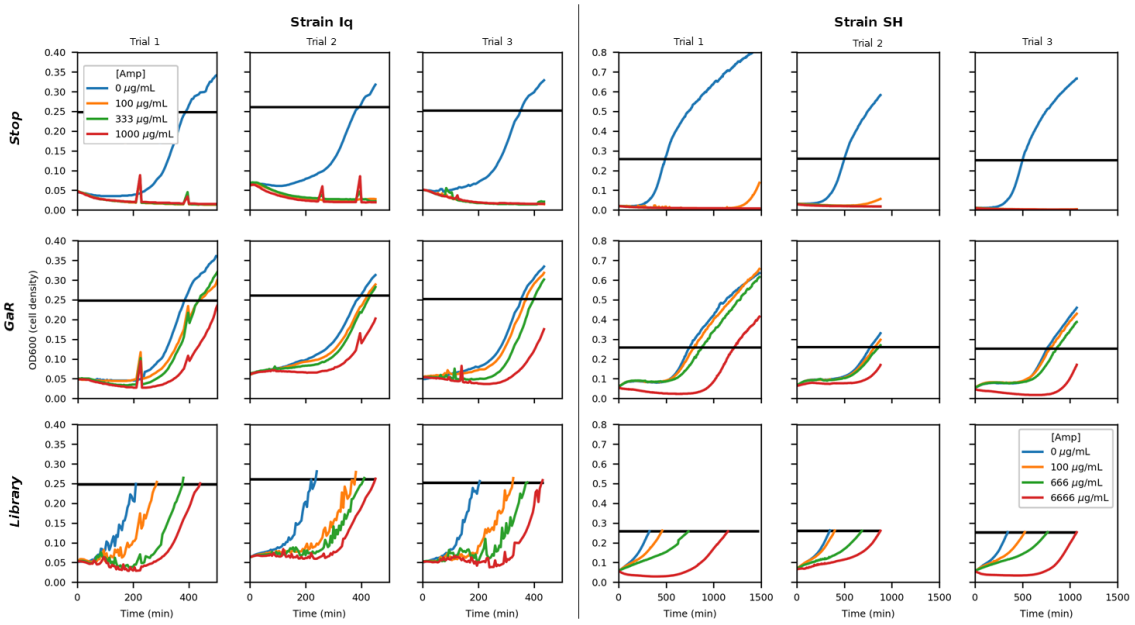

Figure S3.3 - Split β-lactamase Growth Curves of Unmixed Stop and GaR and mixed Library

The growth curves of the control sequences grown in individual wells and the pooled library of cells. The black bar represents the cell density at which the library was collected for sequencing.

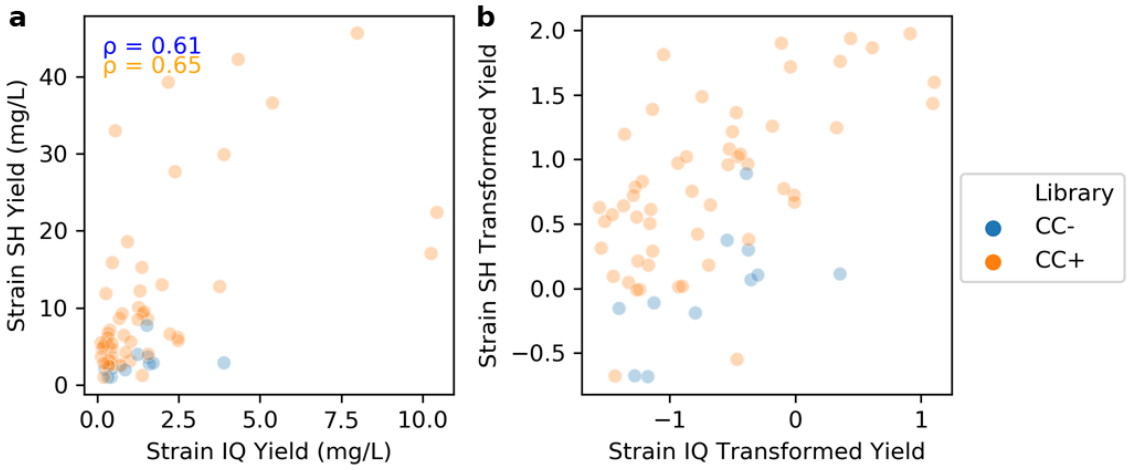

Figure S3.4 - Correlation of yields between cellular strains suggests shared information

a) Scatter plot of recombinant soluble yield of 64 (*CC+*, *orange*: 53, *CC-*, *blue*: 11) unique Gp2 mutants produced in both bacterial strains as measured by chemiluminescent dot blot. The Spearman's rank correlation coefficient (ρ) is shown for each sequence class in the upper left corner. **b)** Scatter plot of the same sequences but utilizing the transformed yield used during HT assay predictions (see Figure S6).

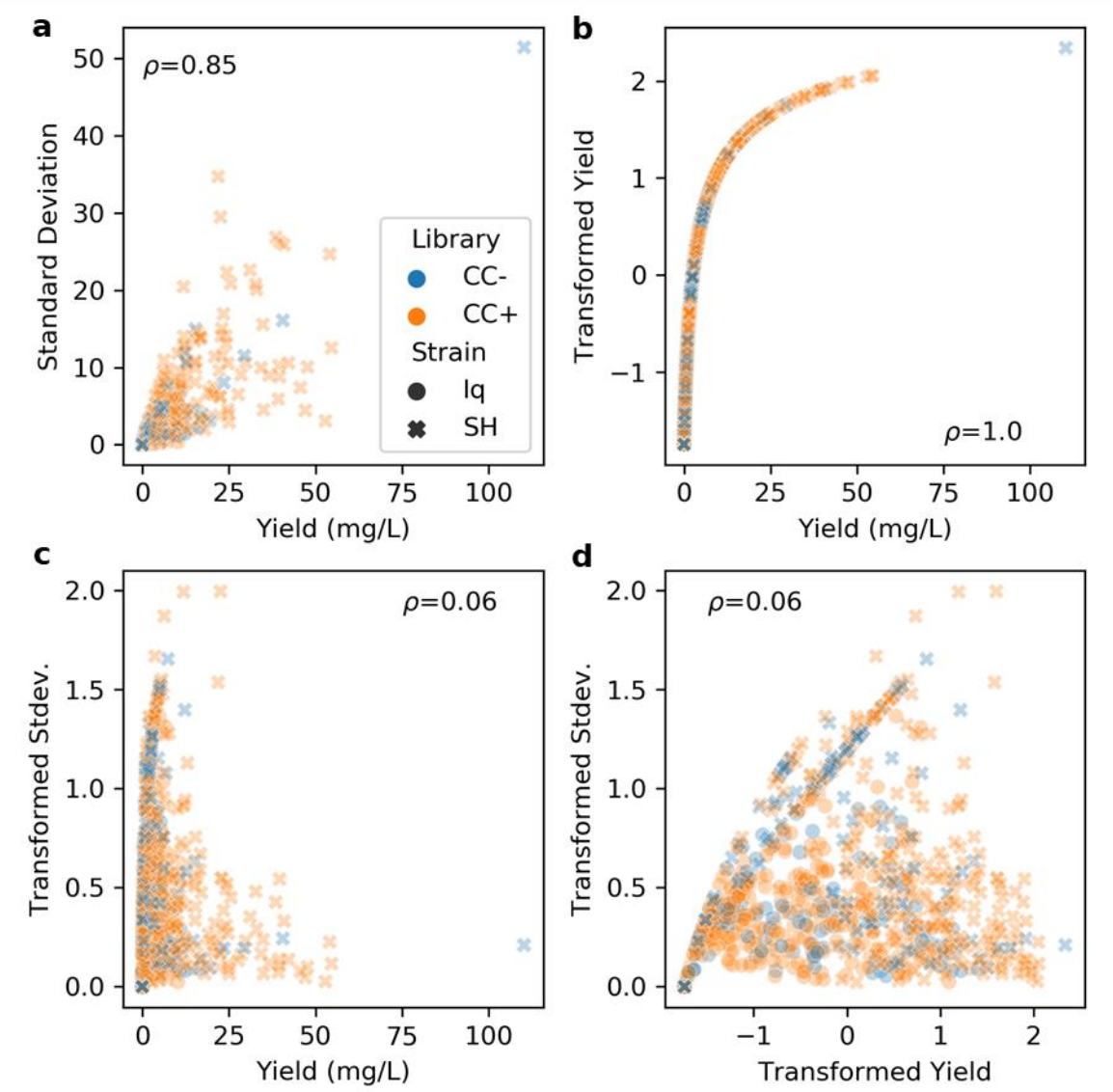

Figure S3.5 - Yeo-Johnson power transformation and normalization of yield

Transformation was performed to remove heteroscedasticity (increasing experimental error with increasing yield). A single transformation parameter (λ = -0.324) was trained using only those sequences utilized in cross-validation for both libraries (CC+: orange, CC-: blue) and production strains (Iq: circle, SH: cross). **a)** Scatter plot of the yield for unique Gp2 sequences versus the trial-to-trial (n=3) standard deviation showing an increasing trend. **b)** Scatter plot of the measured yield via calibration curve of the chemiluminescent dot blot versus the transformed yield used for model evaluation. **c,d)** Scatter plots showing the removal of increasing experimental error with increasing transformed yield.

Model Inputs	CV Loss *N=195*	Test[1] Loss *N=44*	Test[2] Loss *N=97*
Experimental Variance	0.350	0.379	0.364
Strain Only	0.685±0.007	0.612±0.000	0.697±0.000
One-Hot Sequence	0.592±0.011	0.564±0.006	0.667±0.005
Sequence & P_{PK37},G_{SH},β_{SH}	0.500±0.010	**0.470±0.007**	**0.562±0.007**
P_{PK37}, G_{SH}, β_{SH}	0.499±0.008	**0.520±0.003**	**0.565±0.004**
P_{PK37}, P_{TL55}, G_{SH}, β_{SH}	**0.497±0.009**	0.546±0.016	
P_{PK37}, P_{TL55}, G_{SH}, β_{lq}, β_{SH}	0.499±0.010	0.552±0.005	
P_{PK37}, P_{PK55}, P_{TL55}, G_{SH}, β_{SH}	0.503±0.010	0.567±0.008	
P_{PK37}, P_{TL55}, G_{lq}, G_{SH}, β_{SH}	0.505±0.011	0.560±0.008	
P_{PK37}, P_{TL55}, G_{lq}, G_{SH}, β_{lq}, β_{SH}	0.505±0.010	0.541±0.008	
P_{PK37}, G_{lq}, G_{SH}, β_{SH}	0.507±0.016	0.542±0.008	

Figure S3.6 - Tabulated performance of models

Mean squared error ± standard deviation, CV: n=10 repetitions of 10-fold cross-validation. Test: n=10 models of same architecture with varying random state during training). CV predicted the yield of 195 unique sequences (I^q: 73, SH: 122, Both: 39). Test[1] (44 unique, I^q: 26, SH 21, Both: 3) and Test[2] (97 unique, I^q: 57, SH: 46, Both: 6) sequences were independent of model training and were evaluated to determine the model generalization.

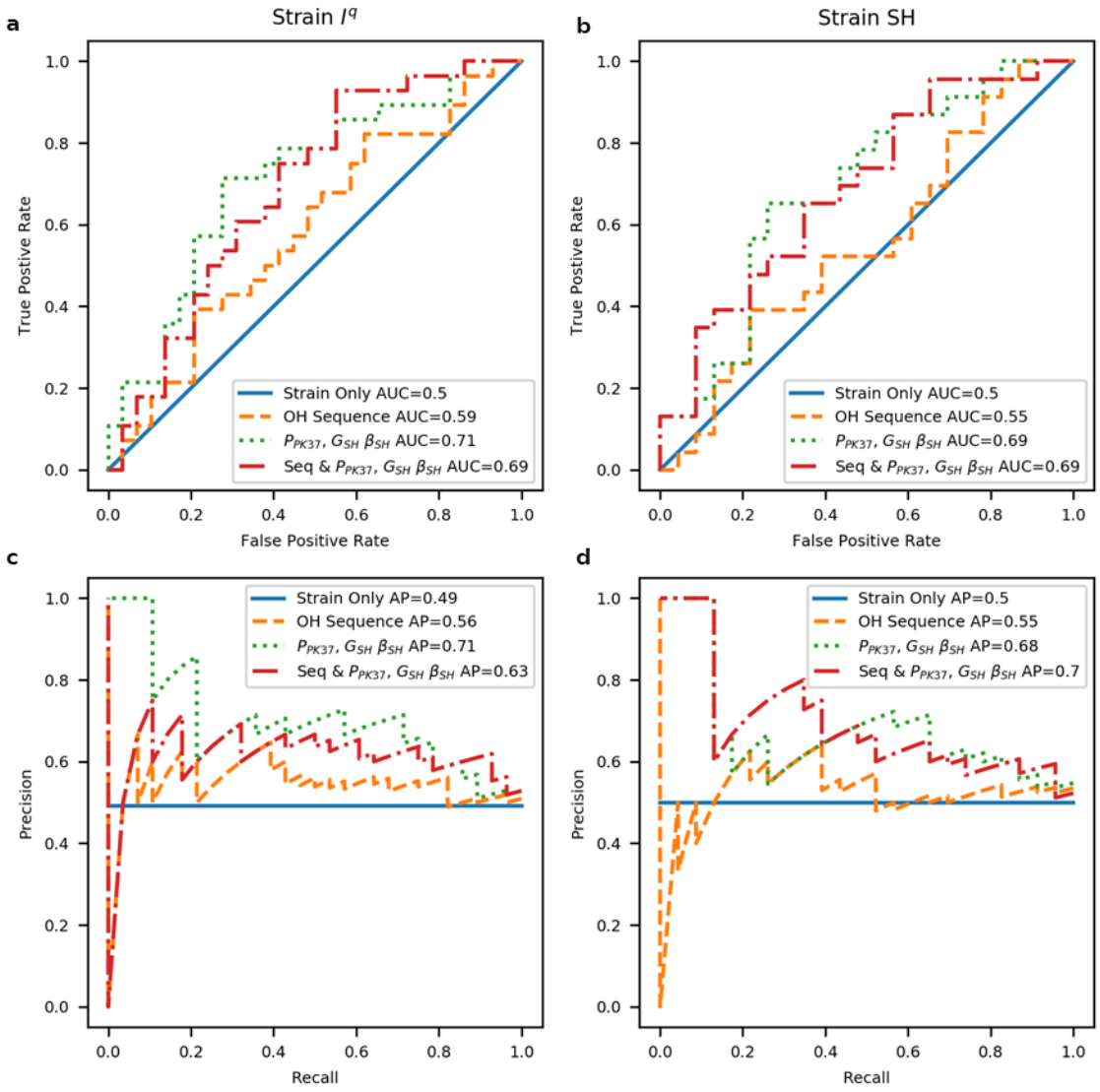

Figure S3.7 - HT assays improve ability to classify sequences with increased developability

The independent test sequences were classified as above or below the median yield per class: Strain I^q transformed yield cutoff = -0.69, n=57 sequences. Strain SH transformed yield cutoff = 0.59, n=46 sequences. The predictive models were assessed by **a,b)** area under receiver operator curve (AUC) or **c,d)** the average precision weighted by the increase in recall (AP).

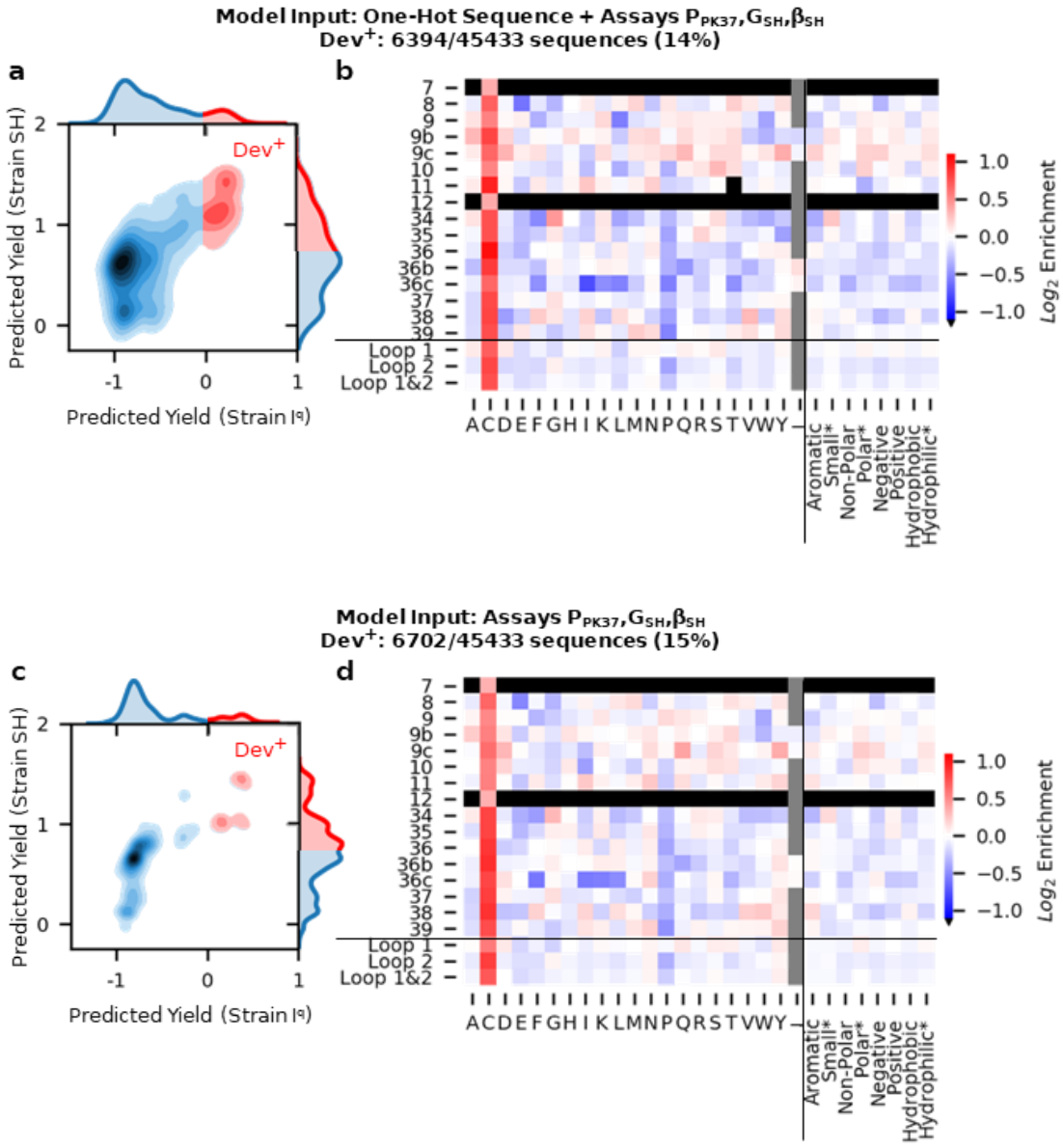

Figure S3.8 - Dev+ sitewise amino acid enrichments displays cysteine preference for models when utilizing HT assay scores

With **(a,b)** and without **(c,d)** sequence information. The predictive performance of both sets of model inputs was not statistically different. This analysis demonstrates the enrichment of cysteine was trained via HT assays, rather than an effect of the inclusion of sequence information in the model.

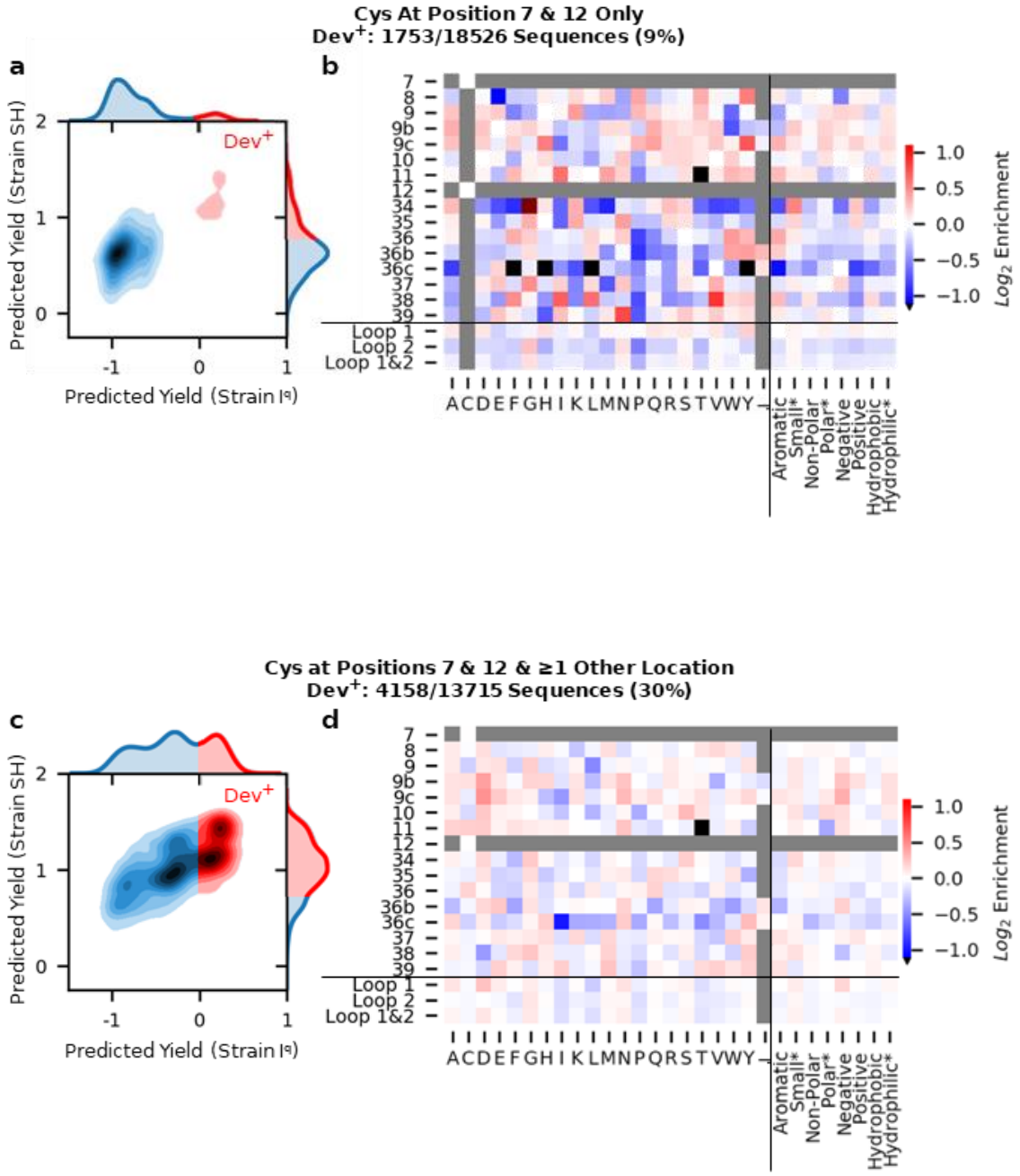

Figure S3.9 - Sitewise enrichment is modified depending on cysteine inclusion outside of positions 7 and 12

Dev$^+$ sequences were predicted to have a transformed yield >0 for strain I^q and >0.75 for Strain SH. Sitewise amino acid enrichment was calculated as the $\log_2$ change in frequency of Dev$^+$ versus all predicted sequences. **a,b)** Predicted recombinant yield and enriched amino acids for sequences containing cysteines only at positions 7 and 12. **c,d)** Predicted recombinant yield and enriched amino acids for sequence containing cysteines at positions 7 and 12 and at least one other position.

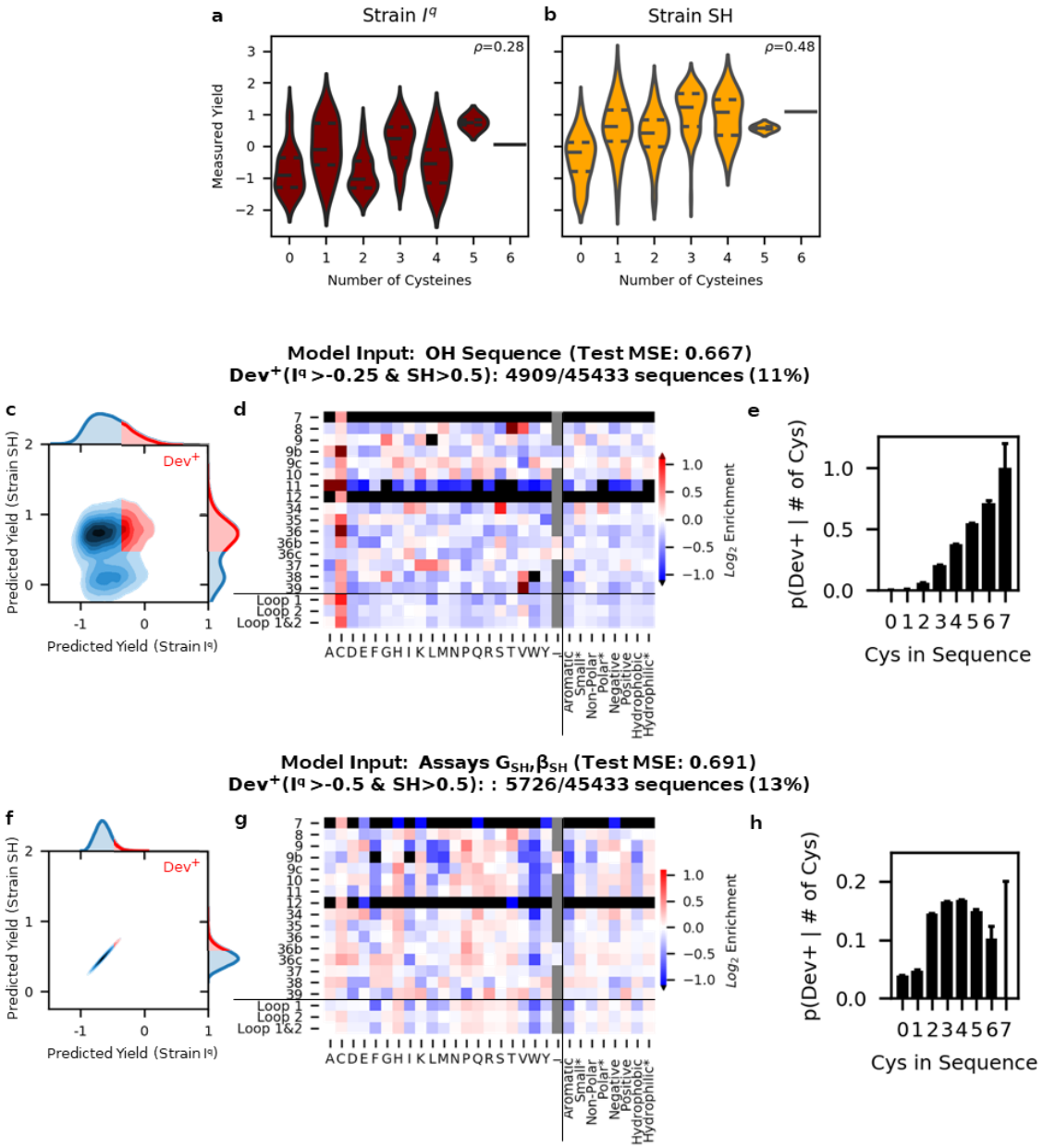

Figure S3.10 - Cysteine Enrichment Independent of On-Yeast Protease Assay

The distributions of experimentally measured yield via dot-blot in **a)** strain I^q and **b)** strain SH broken down by the number of cysteines. The sitewise and overall preference of cysteines in highly developability sequences was also analyzed by **c-e)** an OH sequence-based model and **f-h)** a model utilizing assays G_{SH} and β_{SH}. The Dev^+ predicted yield thresholds were adjusted to isolate approximately 15% of sequences with highest developability. Error bars: 1 / number of predicted sequences.

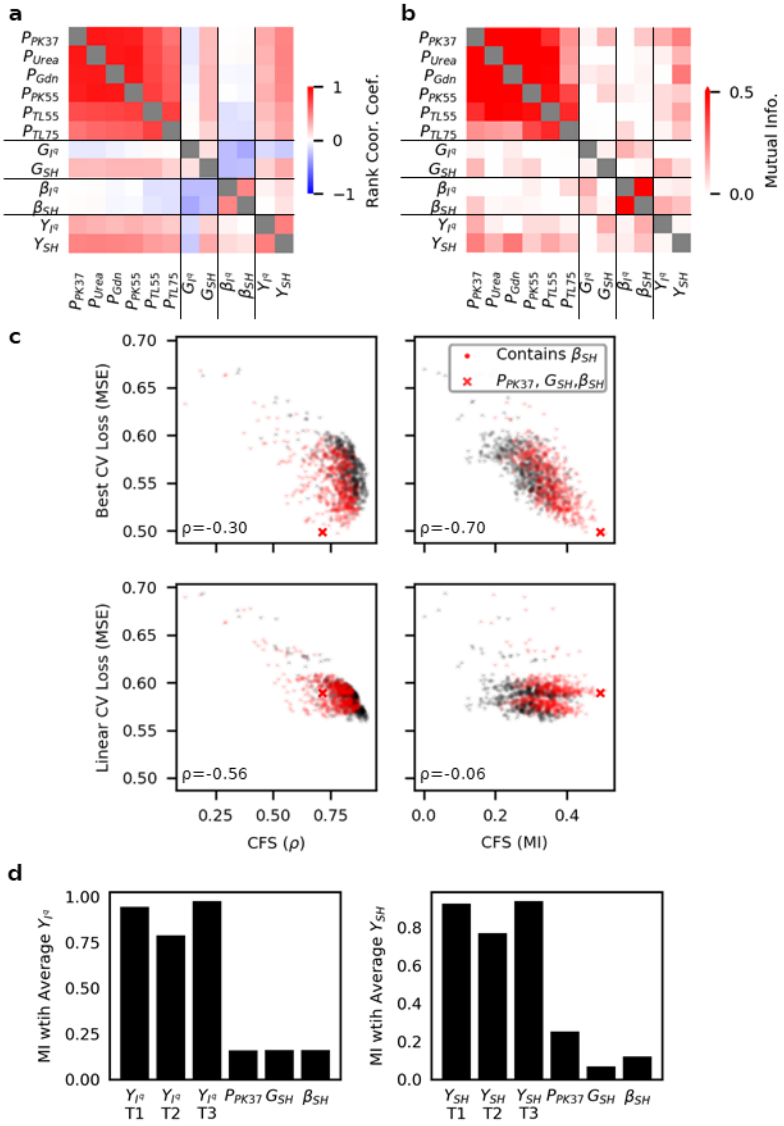

Figure S3.11 - Correlation feature selection (CFS) confirms the selection of most predictive HT assay conditions

a) The Spearman's rank correlation coefficient (ρ) and **b)** the mutual information (MI) between HT assays and yield. **c)** Scatter plot of CFS as calculated by ρ (left) or MI (right) versus predictive loss for the best model architecture (top, best of: Ridge, SVM, Forest, FNN) or a linear model (bottom, Ridge) for the 1023 combinations of HT assays. The ρ between CFS and model loss is presented in the lower-left corner. **d)** Comparison of the trial-to-average mutual information of yield compared to the HT assay-yield mutual information.

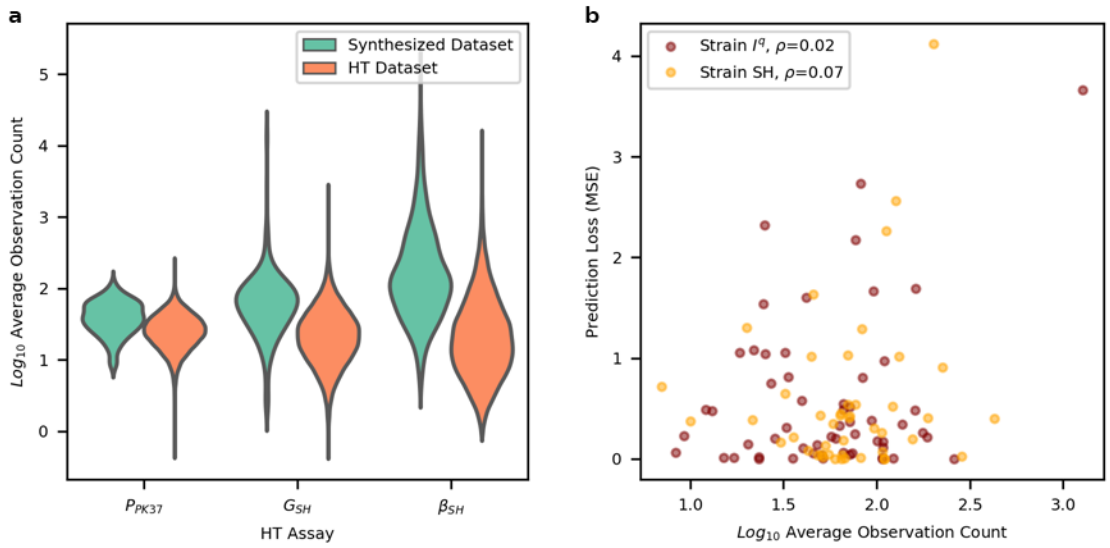

Figure S3.12 - Non-significant correlation between observation frequency and predictive accuracy suggests limited effect of increased sequence observation

a) Distribution of sequence observation frequency (averaged across trials) for sequences synthesized via dot-blot versus all sequences observed during HT assays. **b)** Scatter plot comparing the sequence observation count (averaged across trials and HT assays) and the predictive loss of a model utilizing OH sequence and assays P_{PK37}, G_{SH}, and β_{SH}. The Spearman's rank correlation (ρ) is presented for each strain.

a

4-Gate Stringency	P_{PK37}	G_{SH}	β_{SH}
	Highest cMyc:HA Gate (1) Medium-High Gate (0.67) Medium-Low Gate (0.33) Lowest Gate (0)	Highest GFP Gate (1) Medium-High Gate (0.67) Medium-Low Gate (0.33) Lowest Gate (0)	Slope Utilizing Zero [Amp,] & 3 [Amp.]

2-Gate Stringency	P_{PK37}	G_{SH}	β_{SH}
Low	Highest 3 cMyc:HA Gates (1) Lowest Gate (0)	Highest 3 GFP Gates (1) Lowest Gate (0)	Slope Utilizing Zero [Amp.] & Lowest [Amp.]
Medium	Highest 2 cMyc:HA Gates (1) Lowest 2 Gates (0)	Top 2 GFP Gates (1) Lowest 2 Gates (0)	Slope Utilizing Zero [Amp.] & Middle [Amp.]
High	Highest cMyc:HA Gate (1) Lowest 3 Gates (0)	Highest GFP Gate (1) Lowest 3 Gates (0)	Slope Utilizing Zero [Amp.] & Highest [Amp.]

b

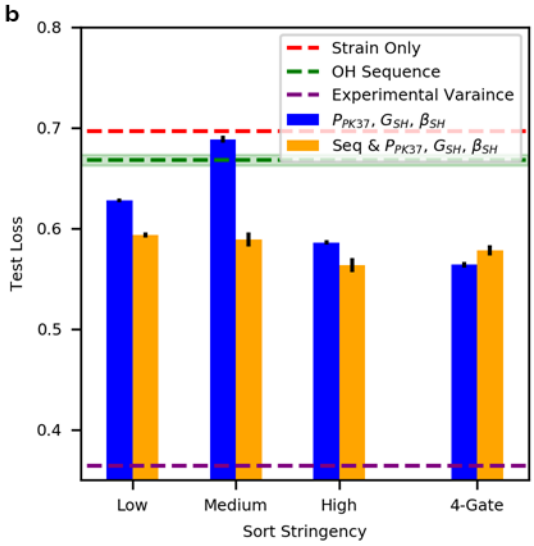

Figure S3.13 - Effect of the number of HT assay populations collected

a) Comparison tables describing the scoring strategy when utilizing 4 populations or 2 populations. The on-yeast protease and split GFP assay's flow cytometry gates were combined by assigning the same score to sequences collected in different gates as described by the value within the parenthesis. **b)** The ability to predict yield when utilizing two populations compared to utilizing all four collected populations. Error bars represent the standard deviation in stochastically trained models.

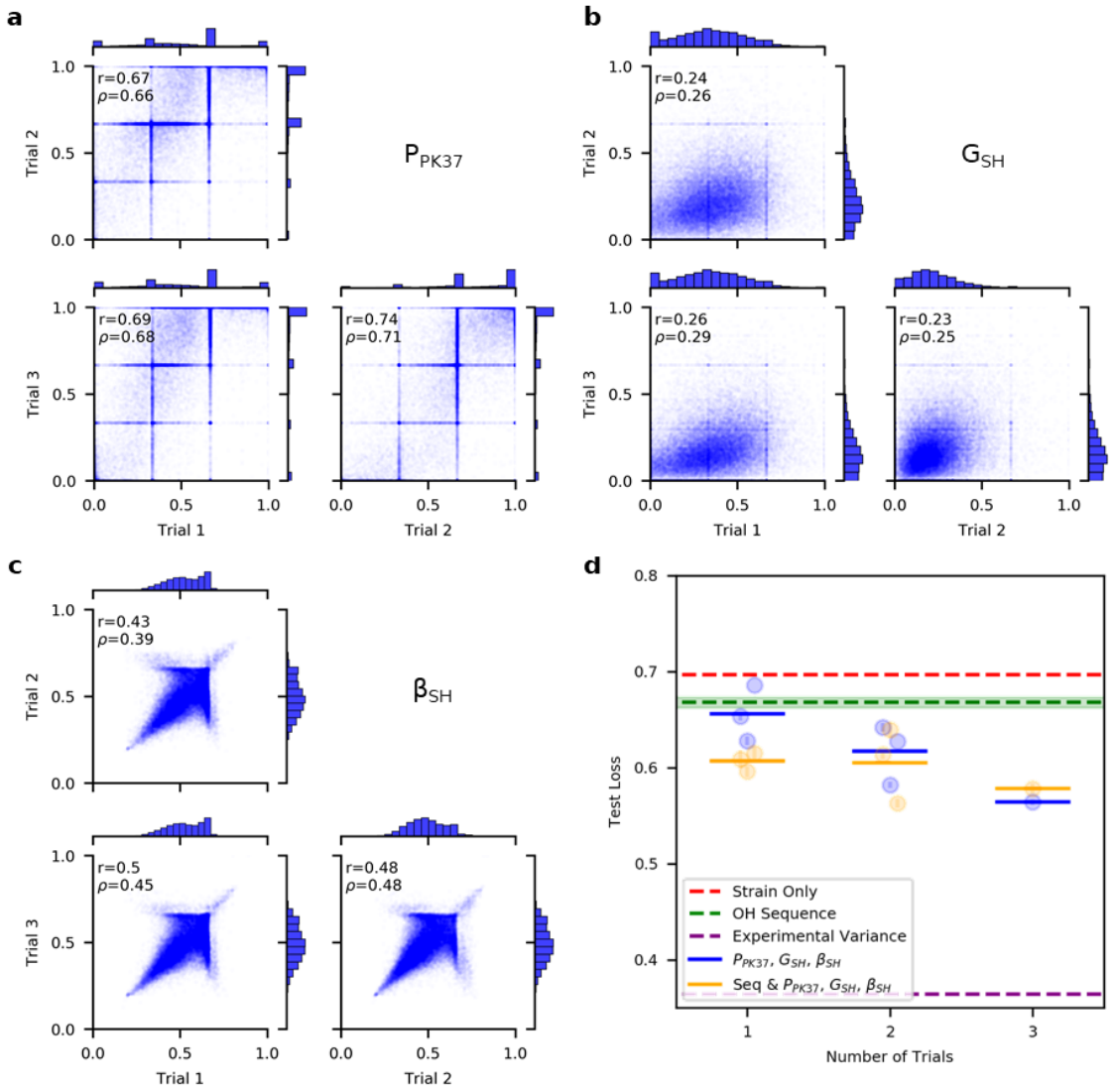

Figure S3.14 - Effect of the number of HT trial replicates

a-c) The trial-to-trial scatter plots with marginal distributions for 45,433 unique sequences. The Pearson's r and Spearman's ρ is printed in the upper corner. **d)** The ability to predict yield when utilizing a subset of the trials. Each point represents a unique combination of trials, with a horizontal bar representing the average performance.

3.11 DNA Tables

DNA Table 3.A - Primers for Gp2 library construction via PCR addition/amplification

1. geneamp5	CGACGATTGAAGGTAGATACCCATACG
2. W5	CGACGATTGAAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAG
3. W5 G4S	TTCCAGACTACGCTCTGCAGGCTAGTAGTGGTGGTGGTTCTGGTGGTGGTGGTTCTGGTG
4. G4S Gp2	GGTTCTGGTGGTGGTGGTTCTGGTGGTGGTGGTTCTGCTAGCAAATTTTGGGCGACTGTA
5. Gp2 1 8	AAATTTTGGGCGACTGTANNKNNKNNKNNKNNKNNKNNKNNKTTTGAGGTGCCGGTGTAT

6. Gp2 1 7	AAATTTTGGGCGACTGTANNKNNKNNKNNKNNKNNKNNKTTTGAGGTGCCGGTGTAT
7. Gp2 1 6	AAATTTTGGGCGACTGTANNKNNKNNKNNKNNKNNKTTTGAGGTGCCGGTGTAT
8. Gp2 1 cys 6	AAATTTTGGGCGACTGTATGTNNKNNKNNKNNKNNKNNKTGTTTTGAGGTGCCGGTGTAT
9. Gp2 1 cys 5	AAATTTTGGGCGACTGTATGTNNKNNKNNKNNKNNKTGTTTTGAGGTGCCGGTGTAT
10. Gp2 1 cys 4	AAATTTTGGGCGACTGTATGTNNKNNKNNKNNKTGTTTTGAGGTGCCGGTGTAT
11. Gp2 mid	TGAGGTGCCGGTGTATGCTGAGACCTTAGACGAAGCTCTTCAGTTAGCTGAATGGCAGTA
12. Gp2 2 8	TTAGCTGAATGGCAGTATNNKNNKNNKNNKNNKNNKNNKNNKGTGACCCGCGTGCGTCCG
13. Gp2 2 7	TTAGCTGAATGGCAGTATNNKNNKNNKNNKNNKNNKNNKGTGACCCGCGTGCGTCCG
14. Gp2 2 6	TTAGCTGAATGGCAGTATNNKNNKNNKNNKNNKNNKGTGACCCGCGTGCGTCCG
15. Gp2 W3	GTGACCCGCGTGCGTCCGGGATCCGAACAAAAGCTTAT
16. W3	GGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGAT
17. geneamp3	ATCTCGAGCTATTACAAGTCCTCTTC

DNA Table 3.B - Primer used to create negative control/baseline vector for on-yeast protease screening

Pct-stop-myc insert	TTCCAGACTACGCTCTGCAGGCTAGCTAATAGATAAGTAGGGGATCCGAACAAAAGCTTATTTCT

DNA Table 3.C - PCR1 primers for Illumina preparation of on-yeast protease screening

FA1GSmyc	TTTCCCTACACGACGCTCTTCCGATCTNNNNTGGTGGTTCTGCTAGC
FA2GSmyc	TTTCCCTACACGACGCTCTTCCGATCTNNNNNTGGTGGTTCTGCTAGC
FA3GSmyc	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNTGGTGGTTCTGCTAGC
FA4GSmyc	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNTGGTGGTTCTGCTAGC
FA5GSmyc	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNTGGTGGTTCTGCTAGC
RA1Gp2-myc	GTTCAGACGTGTGCTCTTCCGATCTNNNNTAAGCTTTTGTTCGGATCC
RA2Gp2-myc	GTTCAGACGTGTGCTCTTCCGATCTNNNNNTAAGCTTTTGTTCGGATCC
RA3Gp2-myc	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNTAAGCTTTTGTTCGGATCC
RA4Gp2-myc	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNTAAGCTTTTGTTCGGATCC
RA5Gp2-myc	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNTAAGCTTTTGTTCGGATCC

DNA Table 3.D - PCR2 Forward primers for Illumina sequencing with trial specific barcodes

Barcode highlighted in red. Note: Trial 1 was run on a separate Novaseq and thus did not have an FB barcode.

Trial 1	Old FB	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
Trial 2	FB-2	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGACGCTCTT
Trial 3	FB-3	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGACGCTCTT

DNA Table 3.E - PCR2 reverse primers for Illumina sequencing with gate specific barcodes

RB: 5'- CAA GCA GAA GAC GGC ATA CGA GAT [Barcode] GTG ACT GGA GTT CAG ACG TGT GCT CTT CC

RPI 1	CGTGAT	RPI 13	TTGACT	RPI 25	ATCAGT	RPI 37	ATTCCG
RPI 2	ACATCG	RPI 14	GGAACT	RPI 26	GCTCAT	RPI 38	AGCTAG
RPI 3	GCCTAA	RPI 15	TGACAT	RPI 27	AGGAAT	RPI 39	GTATAG
RPI 4	TGGTCA	RPI 16	GGACGG	RPI 28	CTTTTG	RPI 40	TCTGAG
RPI 5	CACTGT	RPI 17	CTCTAC	RPI 29	TAGTTG		
RPI 6	ATTGGC	RPI 18	GCGGAC	RPI 30	CCGGTG		
RPI 7	GATCTG	RPI 19	TTTCAC	RPI 31	ATCGTG		
RPI 8	TCAAGT	RPI 20	GGCCAC	RPI 32	TGAGTG		
RPI 9	CTGATC	RPI 21	CGAAAC	RPI 33	CGCCTG		
RPI 10	AAGCTA	RPI 22	CGTACG	RPI 34	GCCATG		
RPI 11	GTAGCC	RPI 23	CCACTC	RPI 35	AAAATG		
RPI 12	TACAAG	RPI 24	GCTACC	RPI 36	TGTTGG		

DNA Table 3.F - Primers used to amplify GFP1-10 from obtained plamid with overlaps to allow for Gibson assembly into pBAD plasmid

GFP1-10 amp top	ATGGTTCTTCTATGGCTAGCATGTCCAAAGGAGAAGAACTGTTTACC
GFP1-10 amp bottom	CCAAAACAGCCAAGGGATCCTTTTTCATTTGGATCTTTGCTCA

DNA Table 3.G - DNA used to add GFP11 to pET and to add the stop codon for the negative control

pET-GFP11 gblock	CTTTAAGAAGGAGATATACATATGGCTAGCGCGTGGGGCGGATCCGGTGGAGGTGGATC GCGTGATCACATGGTATTACATGAATACGTGAACGCTGCTGGGATTACATGATTAACTAA ACGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTG AGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGA
GFP11-stop insert	AGGAGATATACATATGGCTAGCTAATAGATAAGTAGGGGATCCGGTGGAGGTGGAT

DNA Table 3.H - PCR1 primers for illumina preparation of split-GFP assay

FApETN501	TTTCCCTACACGACGCTCTTCCGATCTNNNNTAGATCGCAAGGAGATATACATATGGCTAGC
FApETN502	TTTCCCTACACGACGCTCTTCCGATCTNNNNNCTCTCTATAAGGAGATATACATATGGCTAGC
FApETN503	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNTATCCTCTAAGGAGATATACATATGGCTAGC
FApETN504	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNAGAGTAGAAAGGAGATATACATATGGCTAGC
FApETN505	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGTAAGGAGAAGGAGATATACATATGGCTAGC
RA1 gfp	GTTCAGACGTGTGCTCTTCCGATCTNNNNCCTCCACCGGATCC
RA2 gfp	GTTCAGACGTGTGCTCTTCCGATCTNNNNNCCTCCACCGGATCC
RA3 gfp	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNCCTCCACCGGATCC
RA4 gfp	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNCCTCCACCGGATCC
RA5 gfp	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCCTCCACCGGATCC

DNA Table 3.I - gBlocks for split β-lactamase plasmid

gblock A	AATTTTGTTTAACTTTAAGAAGGAGATATACATATGATGAGTATTCAACATTTCCGTGTCGCCCTTATT CCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCT GAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGA GTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTAT CCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAG TACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCA TAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAAC CGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAG CCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATT AACTGGCGGTGGTGGCGGAAGTGGAGGCGGAGGCAGTGCTAGCAAGTGTTTGAAGTGGGCTGGATCC GGAGGAGGCGGAAGTGGAGGAGGAGGTAGC GAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGAC CACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGG TCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGAC GGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAG CATTGGTAATGATTAACTAAACGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA
gblock B	GGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAA CCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTT
β stop insert	GGAGGCGGAGGCAGTGCTAGCTAATAGATAAGTAGGGGATCCGGAGGAGGCGGAAGT

DNA Table 3.J - PCR1 primers for Illumina preparation of split β-lactamase assay

FA1 blac	TTTCCCTACACGACGCTCTTCCGATCTNNNNGAGGCAGTGCTAGC
FA2 blac	TTTCCCTACACGACGCTCTTCCGATCTNNNNNGAGGCAGTGCTAGC
FA3 blac	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNGAGGCAGTGCTAGC
FA4 blac	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNGAGGCAGTGCTAGC
FA5 blac	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGAGGCAGTGCTAGC
RA1 blac	GTTCAGACGTGTGCTCTTCCGATCTNNNNCTAGAGTAAGTAGTTCGCTAC
RA2 blac	GTTCAGACGTGTGCTCTTCCGATCTNNNNNCTAGAGTAAGTAGTTCGCTAC
RA3 blac	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNCTAGAGTAAGTAGTTCGCTAC
RA4 blac	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNCTAGAGTAAGTAGTTCGCTAC
RA5 blac	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCTAGAGTAAGTAGTTCGCTAC

DNA Table 3.K - Insert to create pET-V5-stop-His6

pET-v5-stop-his6 insert	ACTTTAAGAAGGAGATATACATATGGGCAAACCGATTCCTAATCCGCTTTTAGGTTTGGATAGTAC GGCTAGCTAATAGATAAGTAGGGGATCCCACCATCACCATCATCACT

DNA Table 3.L - PCR primers to amplify the Twist Oligopool

OP_fwd	TTGGATAGTACGGCTAGC
OP_rev	GGTGATGGTGGGATCC

DNA Table 3.M - PCR1 primers with row/column specific barcodes to identify the location of the sequence

FApETV5N501	TTTCCCTACACGACGCTCTTCCGATCTNNNNTAGATCGCTTGGATAGTACGGCTAGC
FApETV5N502	TTTCCCTACACGACGCTCTTCCGATCTNNNNNCTCTCTATTTGGATAGTACGGCTAGC
FApETV5N503	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNTATCCTCTTTGGATAGTACGGCTAGC
FApETV5N504	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNAGAGTAGATTGGATAGTACGGCTAGC
FApETV5N505	TTTCCCTACACGACGCTCTTCCGATCTNNNNGTAAGGAGTTGGATAGTACGGCTAGC
FApETV5N506	TTTCCCTACACGACGCTCTTCCGATCTNNNNNACTGCATATTGGATAGTACGGCTAGC
FApETV5N507	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNAAGGAGTATTGGATAGTACGGCTAGC
FApETV5N508	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNCTAAGCCTTTGGATAGTACGGCTAGC
RApETN701	GTTCAGACGTGTGCTCTTCCGATCTNNNNTCGCCTTAGGTGATGGTGGGATCC
RApETN702	GTTCAGACGTGTGCTCTTCCGATCTNNNNNCTAGTACGGGTGATGGTGGGATCC
RApETN703	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNTTCTGCCTGGTGATGGTGGGATCC
RApETN704	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNGCTCAGGAGGTGATGGTGGGATCC
RApETN705	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNAGGAGTCCGGTGATGGTGGGATCC
RApETN706	GTTCAGACGTGTGCTCTTCCGATCTNNNNCATGCCTAGGTGATGGTGGGATCC
RApETN707	GTTCAGACGTGTGCTCTTCCGATCTNNNNNGTAGAGAGGGTGATGGTGGGATCC
RApETN708	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNCCTCTCTGGGTGATGGTGGGATCC
RApETN709	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNAGCGTAGCGGTGATGGTGGGATCC
RApETN710	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCAGCCTCGGGTGATGGTGGGATCC
RApETN711	GTTCAGACGTGTGCTCTTCCGATCTNNNNNTGCCTCTTGGTGATGGTGGGATCC
RApETN712	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNTCCTCTACGGTGATGGTGGGATCC

3.12 Plasmid Sequences

3.12.1 pCT-HA-stop-Myc

Resistance: Ampicillin

Use: Yeast surface display with myc and no linker unless included in insert. Digest with PstI and BamHI, and electroporate with geneamp3/5 PCR'd insert

Summary:

Aga2--Spacer--FactorXa--HA--PstI--NheI—STOP-BamHI--Myc--2Stop--XhoI--Terminator....

Sequence:

ACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAG
GACGGATCGCTTGCCTGTAACTTACACGCGCCTCGTATCTTTTAATGATGGAATAATTTGGGA
ATTTACTCTGTGTTTATTTATTTTTATGTTTTGTATTTGGATTTTAGAAAGTAAATAAAGAAGGT
AGAAGAGTTACGGAATGAAGAAAAAAAAAATAAACAAAGGTTTAAAAAATTTCAACAAAAAG
CGTACTTTACATATATATTTATTAGACAAGAAAAGCAGATTAAATAGATATACATTCGATTAA
CGATAAGTAAAATGTAAAATCACAGGATTTTCGTGTGTGGTCTTCTACACAGACAAGATGAAA
CAATTCGGCATTAATACCTGAGAGCAGGAAGAGCAAGATAAAAGGTAGTATTTGTTGGCGAT
CCCCCTAGAGTCTTTTACATCTTCGGAAAACAAAAACTATTTTTTCTTTAATTTCTTTTTTTACT

TTCTATTTTTAATTTATATATTTATATTAAAAAATTTAAATTATAATTATTTTTATAGCACGTGA
TGAAAAGGACCCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTT
CTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATA
TTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCA
TTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAG
TTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTT
CGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTA
TCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG
GTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATG
CAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGG
ACCGAAGGAGCTAACCGCTTTTTTTCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTG
GGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAA
TGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAAT
TAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTG
GCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC
TGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGCAGTCAGGCAACTA
TGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGT
CAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGAT
CTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCAC
TGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTA
ATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAG
CTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTT
CTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCT
CTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGAC
TCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACA
GCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCATTGAGAAA
GCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAAC
AGGAGAGCGCACGAGGGAGCTTCCAGGGGGGAACGCCTGGTATCTTTATAGTCCTGTCGGGT
TTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCCGAGCCTATGGA
AAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTT
CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACC
GCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
CAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGT
TTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTACCTCACTCATTAGG
CACCCCAGGCTTTACACTTTATGCTTCCGGCTCCTATGTTGTGTGGAATTGTGAGCGGATAACA
ATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCTCGGAATTAACCCTCACTAAAG
GGAACAAAAGCTGGGTACCCGACAGGTTATCAGCAACAACACAGTCATATCCATTCTCAATTA
GCTCTACCACAGTGTGTGAACCAATGTATCCAGCACCACCTGTAACCAAAACAATTTTAGAAG
TACTTTCACTTTGTAACTGAGCTGTCATTTATATTGAATTTTCAAAAATTCTTACTTTTTTTTTG
GATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACATATCCATAT
ACATATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCT
AAAAAAACCTTCTCTTTGGAACTTTCAGTAATACGCTTAACTGCTCATTGCTATATTGAAGTAC
GGATTAGAAGCCGCCGAGCGGGTGACAGCCCTCCGAAGGAAGACTCTCCTCCGTGCGTCCTC
GTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAA
AGATTCTACAATACTAGCTTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCAC
AAACCTTCAAATGAACGAATCAAATTAACAACCATAGGATGATAATGCGATTAGTTTTTTAGCC
CTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATG
CAAAAACTGCATAACCACTTTAACTAATACTTTCAACATTTTCGGTTTGTATTACTTCTTATTC
AAATGTAATAAAAGTATCAACAAAAAATTGTTAATATACCTCTATACTTTAACGTCAAGGAGA
AAAAACTATAGAATTCTACTTCATACATTTTCAATTAAGATGCAGTTACTTCGCTGTTTTTCAA
TATTTTCTGTTATTGCTTCAGTTTTAGCACAGGAACTGACAACTATATGCGAGCAAATCCCCTC
ACCAACTTTAGAATCGACGCCGTACTCTTTGTCAACGACTACTATTTTGGCCAACGGGAAGGC

AATGCAAGGAGTTTTTGAATATTACAAATCAGTAACGTTTGTCAGTAATTGCGGTTCTCACCC
CTCAACAACTAGCAAAGGCAGCCCCATAAACACACAGTATGTTTTTAAGGACAATAGCTCGA
CGATTGAAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGCTAATAGATAA
GTAGGGGATCC**GAACAAAAGCTTA**TTTCTGAAGAGGACTTGTAATAGCTCGAGATCTGATAA
CAACAGTGTAGATGTAACAAAATCGACTTTGTTCCCACTGTACTTTTAGCTCGTACAAAATAC
AATATACTTTTCATTTCTCCGTAAACAACATGTTTTCCCATGTAATATCCTTTTCTATTTTTCGT
TCCGTTACCAACTTTACACATACTTTATATAGCTATTCACTTCTATACACTAAAAAACTAAGAC
AATTTTAATTTTGCTGCCTGCCATATTTCAATTTGTTATAAATTCCTATAATTTATCCTATTAGT
AGCTAAAAAAAGATGAATGTGAATCGAATCCTAAGAGAATTGAGCTCCAATTCGCCCTATAG
TGAGTCGTATTACAATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGT
TACCCAACTTAATCGCCTTGCAGCACATCCCCCCTTCGCCAGCTGGCGTAATAGCGAAGAGGC
CCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCGACGCGCCCTG
TAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCA
GCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCC
CGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACC
CCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTC
GCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACT
CAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTA
AAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATT
TCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCAGGCAAGTGCACAA
ACAATACTTAAATAAATACTACTCAGTAATAACCTATTTCTTAGCATTTTTGACGAAATTTGCT
ATTTTGTTAGAGTCTTTTACACCATTTGTCTCCACACCTCCGCTTACATCAACACCAATAACGC
CATTTAATCTAAGCGCATCACCAACATTTTCTGGCGTCAGTCCACCAGCTAACATAAAATGTA
AGCTTTCGGGGCTCTCTTGCCTTCCAACCCAGTCAGAAATCGAGTTCCAATCCAAAAGTTCAC
CTGTCCCACCTGCTTCTGAATCAAACAAGGGAATAAACGAATGAGGTTTCTGTGAAGCTGCAC
TGAGTAGTATGTTGCAGTCTTTTGGAAATACGAGTCTTTTAATAACTGGCAAACCGAGGAACT
CTTGGTATTCTTGCCACGACTCATCTCCATGCAGTTGGACGATATCAATGCCGTAATCATTGAC
CAGAGCCAAAACATCCTCCTTAGGTTGATTACGAAACACGCCAACCAAGTATTTCGGAGTGCC
TGAACTATTTTTATATGCTTTTACAAGACTTGAAATTTTCCTTGCAATAACCGGGTCAATTGTT
CTCTTTCTATTGGGCACACATATAATACCCAGCAAGTCAGCATCGGAATCTAGAGCACATTCT
GCGGCCTCTGTGCTCTGCAAGCCGCAAACTTTCACCAATGGACCAGAACTACCTGTGAAATTA
ATAACAGACATACTCCAAGCTGCCTTTGTGTGCTTAATCACGTATACTCACGTGCTCAATAGTC
ACCAATGCCCTCCCTCTTGGCCCTCTCCTTTTCTTTTTTCGACCGAATTAATTCTTAATCGGCAA
AAAAAGAAAAGCTCCGGATCAAGATTGTACGTAAGGTGACAAGCTATTTTTCAATAAAGAAT
ATCTTCCACTACTGCCATCTGGCGTCATAACTGCAAAGTACACATATATTACGATGCTGTCTAT
TAAATGCTTCCTATATTATATATATAGTAATGTCGTTTATGGTGCACTCTCAGTACAATCTGCT
CTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGG
GCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGT
CAGAGGTTTTCACCGTCATCACCGAAACGCGCGA

Protein:

Aga2p – KDNSST – Xa – HA –LQ– AS – STOP (3 frames) – GS – c-myc

MQLLRCFSIFSVIASVLAQELTTICEQIPSPTLESTPYSLSTTTILANGKAMQGVFEYYKSVTFVSNC
GSHPSTTSKGSPINTQYVFKDNSSTIEGRYPYDVPDYALQAS**ISRGSEQKLISEEDL**

3.12.2 pBAD-GFP$_{1-10}$

Resistance: Ampicillin

Use: This construct contains parts 1-10 of the whole GFP which when combined with its final part will fluoresce. Inducible with arabinose

Summary:

--NdeI - 6xHis—-Spacer--NheI—-GFP(1-10)—-BamHI—spacer- stop

Sequence:

AATTCTTGAAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATA
ATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTAT
TTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATA
ACATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCG
GCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGAT
CAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGT
TTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTAT
TATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACT
TGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTA
TGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGA
GGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGT
TGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGC
AATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACA
ATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGC
TGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGC
ACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAA
CTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAA
CCCGGGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAG
GATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTC
CACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGC
GTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAA
GAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTC
CTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTC
GCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTG
GACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCAC
ACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAG
AAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGG
AACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCG
GGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTAT
GGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACAT
GTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGAT
ACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGC
GCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATGGTGCACTCT
CAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGAC
TGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTG
CTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTT
TCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAG
CGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAAT
GTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCT
CCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTC
ACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTA
ATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATA
ATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATT
CATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCG
GTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGA
GCACGATCATGCGCACCCGTGGCCAGGACCCAACGCTGCCCGAGATGCGCCGCGTGCGGCTG

CTGGAGATGGCGGACGCGATGGATATGTTCTGCCAAGGGTTGGTTTGCGCATTCACAGTTCTC
CGCAAGAATTGATTGGCTCCAATTCTTGGAGTGGTGAATCCGTTAGCGAGGTGCCGCCGGCTT
CCATTCAGGTCGAGGTGGCCCGGCTCCATGCACCGCGACGCAACGCGGGGAGGCAGACAAGG
TATAGGGCGGCGCCTACAATCCATGCCAACCCGTTCCATGTGCTCGCCGAGGCGGCATAAATC
GCCGTGACGATCAGCGGTCCAGTGATCGAAGTTAGGCTGGTAAGAGCCGCGAGCGATCCTTG
AAGCTGTCCCTGATGGTCGTCATCTACCTGCCTGGACAGCATGGCCTGCAACGCGGGCATCCC
GATGCCGCCGGAAGCGAGAAGAATCATAATGGGGAAGGCCATCCAGCCTCGCGTCGCGAACG
CCAGCAAGACGTAGCCCAGCGCGTCGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGA
AACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACC
GCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCA
GAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGA
CGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATC
GGTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCG
TTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGC
CACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCC
GATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGA
TGCCGGCCACGATGCGTCCGGCGTAGAGGATCGCGGCCGCCATAATGTGCCTGTCAAATGGA
CGAAGCAGGGATTCTGCAAACCCTATGCTACTCCCTCGAGCCGTCAATTGTCTGATTCGTTAC
CAATTATGACAACTTGACGGCTACATCATTCACTTTTTCTTCACAACCGGCACGGAACTCGCTC
GGGCTGGCCCCGGTGCATTTTTTAAATACCCGCGAGAAATAGAGTTGATCGTCAAAACCAACA
TTGCGACCGACGGTGGCGATAGGCATCCGGGTGGTGCTCAAAAGCAGCTTCGCCTGGCTGATA
CGTTGGTCCTCGCGCCAGCTTAAGACGCTAATCCCTAACTGCTGGCGGAAAAGATGTGACAGA
CGCGACGGCGACAAGCAAACATGCTGTGCGACGCTGGCTATATCAAAATTGCTGTCTGCCAG
GTGATCGCTGATGTACTGACAAGCCTCGCGTACCCGATTATCCATCGGTGGATGGAGCGACTC
GTTAATCGCTTCCATGCGCCGCAGTAACAATTGCTCAAGCAGATTTATCGCCAGCAGCTCCGA
ATAGCGCCCTTCCCCTTGCCCGGCGTTAATGATTTGCCCAAACAGGTCGCTGAAATGCGGCTG
GTGCGCTTCATCCGGGCGAAAGAACCCCGTATTGGCAAAGATTGACGGCCAGTTAAGCCATTC
ATGCCAGTAGGCGCGCGGACGAAAGTAAACCCACTGGTGATACCATTCGCGAGCCTCCGGAT
GACGACCGTAGTGATGAATCTCTCCTGGCGGGAACAGCAAAATATCACCCGGTCGGCAAACA
AATTCTCGTCCCTGATTTTTCACCACCCCCTGACCGCGAATGGTGAGATTGAGAATATAACCTT
TCATTCCCAGCGGTCGGTCGATAAAAAAATCGAGATAACCGTTGGCCTCAATCGGCGTTAAAC
CCGCCACCAGATGGGCATTAAACGAGTATCCCGGCAGCAGGGGATCATTTTGCGCTTCAGCCA
TACTTTTCATACTCCCGCCATTCAGAGAAGAAACCAATTGTCCATATTGCATCAGACATTGCC
GTCACTGCGTCTTTTACTGGCTCTTCTCGCTAACCAAACCGGTAACCCCGCTTATTAAAAGCAT
TCTGTAACAAAGCGGGACCAAAGCCATGACAAAAACGCGTAACAAAAGTGTCTATAATCACG
GCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATC
CATAAGATTAGCGGATCTTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTCCATACCCG
TTTTTTTGGGCTAACATCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGAAA
TCTTCTCACCATCACCATCACCATGGTTCTTCTATGGCTAGCATGTCCAAAGGAGAAGAACTG
TTTACCGGTGTTGTGCCAATTTTGGTTGAACTCGATGGTGATGTCAACGGACATAAGTTCTCAG
TGAGAGGCGAAGGAGAAGGTGACGCCACCATTGGAAAATTGACTCTTAAATTCATCTGTACT
ACTGGTAAACTTCCTGTACCATGGCCGACTCTCGTAACAACGCTTACGTACGGAGTTCAGTGC
TTTTCGAGATACCCAGACCATATGAAAAGACATGACTTTTTTAAGTCGGCTATGCCTGAAGGT
TACGTGCAAGAAAGAACAATTTCGTTCAAAGATGATGGAAAATATAAAACTAGAGCAGTTGT
TAAATTTGAAGGAGATACTTTGGTTAACCGCATTGAACTGAAAGGAACAGATTTTAAAGAAG
ATGGTAATATTCTTGGACACAAACTCGAATACAATTTTAATAGTCATAACGTATACATCACTG
CTGATAAGCAAAAGAACGGAATTAAAGCGAATTTCACAGTACGCCATAATGTAGAAGATGGC
AGTGTTCAACTTGCCGACCATTACCAACAAAACACCCCTATTGGAGACGGTCCGGTACTTCTT
CCTGATAATCACTACCTCTCAACACAAACAGTCCTGAGCAAAGATCCAAATGAAAAAGGATC
CCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAG
AAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCAT
GCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAG

TAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTT
TATCTGTTGTTTGTCGGTGAACT

Protein:
MKSSHHHHHHGSSMASMSKGEE
LFTGVVPILVELDGDVNGHKFSVRGEGEGDATIGKLTLKFICTTGKLPVPWPTLVTTLTY
GVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGKYKTRAVVKFEGDTLVNRIELK
GTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQLADHYQQNT
PIGDGPVLLPDNHYLSTQTVLSKDPNEKGSLGCFGG*

3.12.3 pET-GFP$_{11}$-Stop

Resistance: Kanamycin

Use: Used to produce POI-GFP$_{11}$ for the split GFP assay

Summary:

.... -- rbs -- TATA -- NdeI -- NheI – Stop (3 frames) -- BamHI – GGGGS – GFP11- Stop -- TTAACTAAACGA -- GATC....

Sequence:

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCA
GCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTC
GCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTA
GTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCAT
CGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTT
GTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTG
CCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAAC
AAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATT
TGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCA
TCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAA
AGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTG
GTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAA
ATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAG
TTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTC
GCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTG
TTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATC
AACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATC
GCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGG
CATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCT
TTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCA
CCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAA
TTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTAC
TGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCC
ACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCG
TAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAG
AGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCC
TTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCG
CTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGG
ACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACA
CAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGA

AAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGA
ACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGG
GTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATG
GAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATG
TTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATA
CCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCG
CCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTC
TCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGA
CTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCT
GCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTT
TTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAA
GCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAA
TGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCC
TCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCT
CACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTA
ATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATA
ATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATT
CATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCG
GTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGA
GCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAAC
GTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCA
AGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAG
CGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGA
TAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGT
CGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCT
TTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGG
CGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGAT
TGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCA
GGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGT
CGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTG
CGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCA
TTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTG
AATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAAC
TTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCA
GTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAA
GAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGC
GGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAG
GCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGA
GATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCC
AATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTC
CGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACG
CGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGC
GCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAG
CCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGAT
GGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCA
TGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCA
ACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTC
GATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTA
GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGGCTAGCTAATAGATAAGTAGGG
GATCCGGTGGAGGTGGATCGCGTGATCACATGGTATTACATGAATACGTGAACGCTGCTGGG

ATTACATGATTAACTAAACGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGC
TGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGG
TTTTTTGCTGAAAGGAGGAACTATATCCGGAT

Protein:

M - AS –**ISR- GS - GGGGS –RDHMVLHEYVNAAGIT*

3.12.4 pET-β-lactamase

Resistance: Kanamycin

Use: Protein production in bacteria testing ampicillin resistance through the ability of the two halves of beta-lactamase combining (bla1 and bla2).

Summary:
....-NdeI–-bla1-–(GGGGS)2–-NheI–-stop–-BamHI–-(GGGGS)2–-bla2–-stop–-....

Sequence:
TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCA
GCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTC
GCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTA
GTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCAT
CGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTT
GTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTG
CCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAAC
AAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATT
TGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCA
TCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAA
AGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTG
GTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAA
ATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAG
TTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTC
GCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTG
TTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATC
AACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATC
GCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGG
CATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCT
TTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCA
CCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAA
TTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTAC
TGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCC
ACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCG
TAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAG
AGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCC
TTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCG
CTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGG
ACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACA
CAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGA
AAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGA
ACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGG
GTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATG
GAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATG
TTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATA
CCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCG
CCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTC
TCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGA

CTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCT
GCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTT
TTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAA
GCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAA
TGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCC
TCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCT
CACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTA
ATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATA
ATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATT
CATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCG
GTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGA
GCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAAC
GTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCA
AGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAG
CGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGA
TAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGT
CGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCT
TTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGG
CGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGAT
TGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCA
GGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGT
CGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTG
CGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCA
TTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTG
AATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAAC
TTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCA
GTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAA
GAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGC
GGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAG
GCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGA
GATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCC
AATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTC
CGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACG
CGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGC
GCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAG
CCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGAT
GGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCA
TGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCA
ACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTC
GATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTA
GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGATGAGTATTCAACATTTCCGTG
TCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTG
AAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAA
CAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAA
AGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCG
CATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGA
TGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCA
ACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGG
ATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAG
CGTGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGGTGG
TGGCGGAAGTGGAGGCGGAGGCAGTGCTAGCTAATAGATAAGTAGGGGATCCGGAGGAGGC
GGAAGTGGAGGAGGAGGTAGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTG
GATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTAT
TGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGA

TGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACG
AAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAATGATTAACTAAACG
AGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAAT
AACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAA
CTATATCCGGAT

Protein:
MMSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYIELDL
NSGKILESFRPEERFPMMSTFKVLLCGAVLSRVDAGQEQLGRRIHYSQNDLVEYSPVTEK
HLTDGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNE
AIPNDERDTTMPAAMATTLRKLLTGGGGSGGGGSAS**ISRGSGGGGSGGGGSELLTL
ASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVI
YTTGSQATMDERNRQIAEIGASLIKHW**

3.12.5 pET-V5-His6

Resistance: Kanamycin

Use: Protein production in bacteria

Purification with N-terminal V5 tag and a C-terminal His_6 tag

Summary:

.... -- rbs -- TATA – NdeI(start) -V5- NheI –STOP-- BamHI -- His6 -- Stop -- TTAACTAAACGA -- GATC....

Sequence:

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCA
GCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTC
GCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTA
GTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCAT
CGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTT
GTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTG
CCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAAC
AAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATT
TGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCA
TCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAA
AGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTG
GTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAA
ATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAG
TTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTC
GCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTG
TTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATC
AACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATC
GCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGG
CATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCT
TTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCA
CCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAA
TTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTAC
TGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCC
ACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCG
TAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAG
AGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCC

TTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCG
CTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGG
ACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACA
CAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGA
AAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGA
ACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGG
GTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATG
GAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATG
TTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATA
CCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCG
CCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTC
TCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGA
CTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCT
GCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTT
TTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAA
GCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAA
TGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCC
TCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCT
CACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTA
ATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATA
ATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATT
CATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCG
GTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGA
GCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAAC
GTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCA
AGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAG
CGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGA
TAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGT
CGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCT
TTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGG
CGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGAT
TGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCA
GGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGT
CGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTG
CGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCA
TTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTG
AATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAAC
TTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCA
GTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAA
GAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGC
GGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAG
GCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGA
GATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCC
AATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTC
CGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACG
CGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGC
GCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAG
CCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGAT
GGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCA
TGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCA

ACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTC
GATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTA
GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGGGCAAACCGATTCCTAATCCGC
TTTTAGGTTTGGATAGTACGGCTAGCTAATAGATAAGTAGGGGATCCCACCATCACCATCATC
ACTGATTAACTAAACGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCT
GCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTT
TTGCTGAAAGGAGGAACTATATCCGGAT

Protein:

MGKPIPNPLLGLDSTAS ** ISR GSHHHHHH

Chapter 4 - Predicting and Interpreting Protein Developability via Transfer of Convolutional Sequence Representation

This chapter describes the current progress of a study and is unpublished at time of defense. Plans for future experimental work are described within the text.

4.1 Abstract

Engineered proteins have emerged as novel diagnostics, therapeutics, and catalysts. Often, poor protein developability - quantified by expression, solubility, and stability - hinders commercialization. The ability to predict protein developability from amino acid sequence would reduce the experimental burden when selecting candidates. Recent advances in screening technologies enabled a high-throughput (HT) developability dataset for 10^5 of 10^{20} possible variants of protein scaffold Gp2. In this work, we evaluate the ability of neural networks to learn a developability representation from the HT dataset and transfer the knowledge to predict recombinant expression beyond the observed sequences. Mimicking protein theory, our model convolves learned amino acid properties to predict expression levels 42% closer to the experimental variance compared to a non-embedded control. Analysis of learned amino acid embeddings highlights the uniqueness of cysteine and the importance of hydrophobicity and charge, and unimportance of aromaticity, when aiming to improve developability. We identify clusters of similar sequences with increased developability through nonlinear dimensionality reduction (UMAP) and explore the inferred developability landscape via nested sampling. We identified a phase transition region where competing sequence motifs permit increased developability. The work aims to advance protein engineering by predicting and interpreting protein scaffold developability and advance data science by displaying the power of machine learning and sampling techniques to study a highly intricate system.

4.2 Introduction

Engineered proteins have broad utility as therapeutics[38], diagnostics[161], and targeted drug-delivery vehicles[5], and as commercial products including industrial enzymes[162] and agricultural processing[163,164]. Beyond the primary function (such as binding affinity or enzymatic activity), the utility of the protein is also dependent on the ability to be manufactured, transported, and stored while maintaining functionality. Commonly termed developability[48,49], this often-overlooked property - quantified by stability, solubility, and production yield - is not typically assessed until late in the commercialization pipeline[56,117]. Late-state developability failures: i) requires substantial time for engineering or discovery a new lead, ii) adds avoidable costs which are often passed on to the consumer, and iii) prevents the immediate use of proteins that would otherwise improve society[58]. The ability to predict protein developability and suggest beneficial mutations would ease the manufacturing process by reducing the experimental effort in selecting lead candidates for further evaluation[58,118].

Predicting protein developability from amino acid sequence is nontrivial due to a myriad of factors: i) the combinatorial space resulting from twenty conical amino acids possible at each position produces an astronomically large domain, ii) the domain is believed to be rugged where a single mutation has the ability to eliminate functionality[13], and iii) traditional developability assays often drastically subsample the domain due to experimental constraints[47]. The combination of these factors suggests the creation of a sequence-developability model, and the accurate determination the most beneficial mutations will require advanced models and sampling techniques.

Recent advances in protein modeling have suggested that machine learning possess the ability to accurately predict functionality[61,62,165]. However, it remains unclear which

embedding, or numeric representation, for proteins results in the most accurate and efficiently trained model. The traditional one-hot embedding for categorical variables creates a sparse embedding that lacks knowledge of physicochemical similarities between amino acids and is likely to result in poor training. Alternative approaches attempt to utilize precomputed amino acid properties, such as AAindex[166] or structurally-based properties, such as non-polar surface area or contact density[134], to embed sequences. However, determining the correct set of properties to use can lead to an exhaustive search. Another popular approach is to utilize an evolutionarily-based model trained from homologues[60,65,167]. However, properties that impact natural proteins (likely including primary function, natural mutational rates, and likelihood of experimental sampling) may not be the properties useful for identifying developability. As a result, we believe the most efficient method of training a sequence-developability model will be using more direct experimental developability proxies, collected in high-throughput (HT), that can be transferred to predict traditional developability metrics.

In this study we aim to train and test a sequence-based model to predict the developability for variants of the protein scaffold Gp2. While specific variants of this 45-49 amino acid protein scaffold have been shown to possess novel binding activity[26,71], serve as a diagnostic in PET imaging[168], and inhibit growth of breast cancer cells[131], many variants still possess poor developability. In a prior study, a series of three HT assays - on-yeast protease resistance, expression as a fusion with split green fluorescent protein (GFP), and modular insertion in split β-lactamase - were validated by mutual information and prediction of Gp2 variant yield[169]. Herein, we will assess the ability to first train a sequence-based machine learning model to predict HT assay performance and transfer the

developability **rep**resentation (DevRep) to improve the accuracy in prediction of a traditional developability metric. (Figure 1). After building a predictive model, we will then i) determine the importance of training data by altering the number of samples and HT assays for training, ii) analyze trained model variables to identify factors driving developability, and iii) use sampling techniques to explore and portray the developability landscape and identify high-yielding variants.

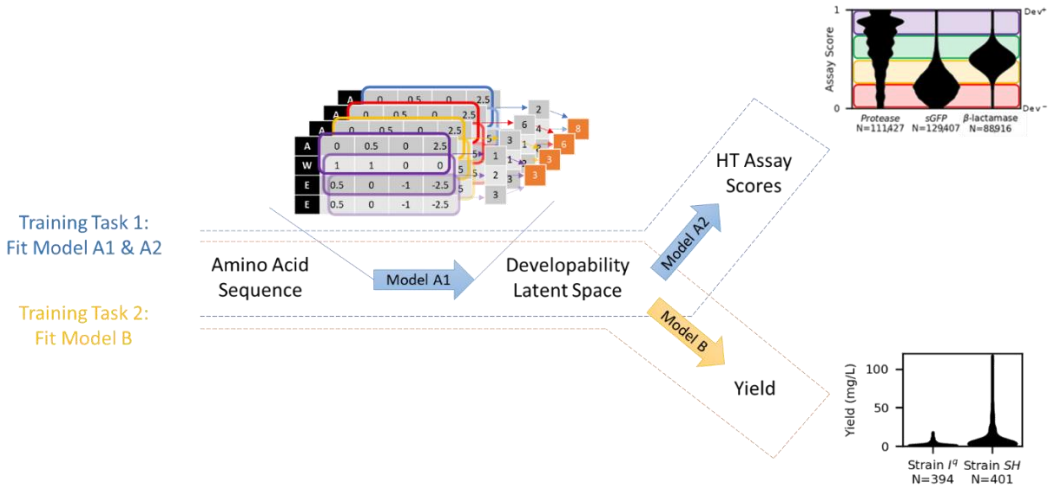

Figure 4.1 - Prediction of protein developability via transfer learning

Training a sequence-based model to predict developability takes place in two steps. Task 1 (blue, top): The large database of protein assay scores will be used to train a mapping (Model A1) from amino acid sequence to a developability latent space (DevRep). Task 2 (orange, bottom): By transferring the mapping, yield (a traditional metric of developability) can be predicted by training a top model with a smaller dataset.

4.3 Results

4.3.1 Training Representations via HT Assays

A protein's properties are determined by the interaction between amino acids, with various chemical properties, in a non-branching sequence. Thus, we constructed models that first learn amino acid properties, then combine the properties to create an embedding representative of the Gp2 paratope (Figure 2a). We considered three architectures: i) Flatten - where all amino acid properties at all positions can interact, ii) Recurrent - where

amino acid properties are fed one at a time into a memory-containing unit that is updated as a function of the previously seen positions, and iii) Convolutional - where amino acid properties are first summarized in a local region of the protein and then combined to obtain a full protein embedding.

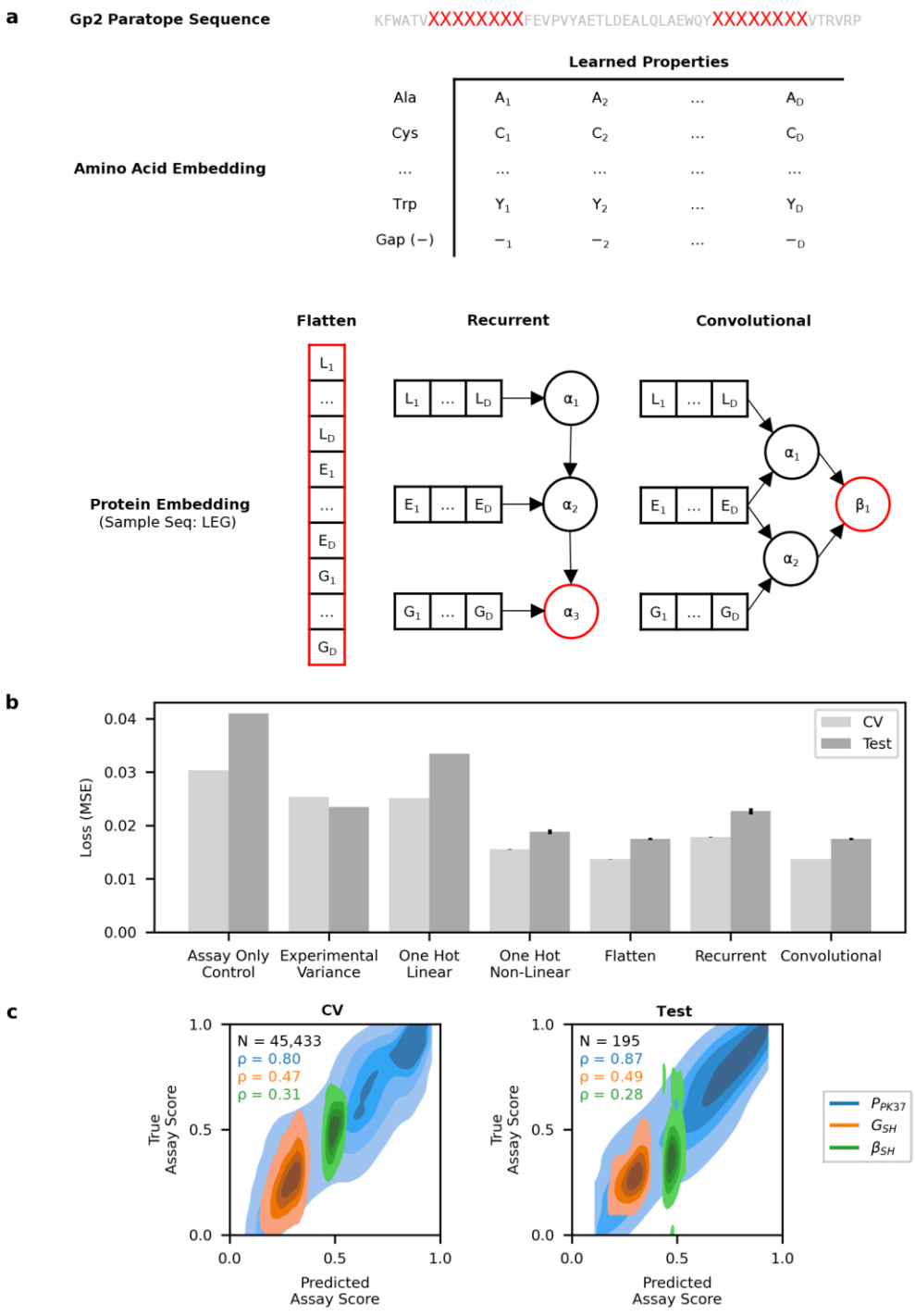

Figure 4.2 - Protein embedding strategies based on interacting amino acid properties predict HT developability assay scores

a) The Gp2 paratope residues are embedded via trained properties and are combined via three different strategies into a developability representation, identified via a red outline. **b)** Embedded and non-embedded (one-hot) architectures were trained to predict assay scores via cross-validation (CV) and evaluated on an independent test set of sequences. **c)** The convolutional architecture's predictions are compared to the true

assay scores as a kernel density plot. The number of unique Gp2 variants and the Spearman's rank correlation are displayed.

Multitask learning was applied to use all three HT assays to train a single developability embedding. Previous analysis revealed that the HT assays are nonlinearly related to each other. Accordingly, we allowed dense layers between the protein embedding and assay scores (exact number determined via hyperparameter optimization during cross-validation) after the concatenation of a one-hot encoded assay identifying vector.

The predictive performance of HT assay score prediction was compared to a series of controls as assessed by the mean squared error of the cross-validation set and an independent test set (Figure 2b). All three architectures (Flatten: 0.014/0.017, Recurrent: 0.018/0.023, Convolutional: 0.014/0.017 *{CV/Test}*) using sequence information were more accurate than the assay only-no sequence information-model (0.030/0.041). Interestingly, the protein-inspired architectures were also able to predict assay scores with lower error than the experimental variance (0.025/0.023), highlighting the previously noted low resolution of a single trial of the assays. We also compared the results to a traditional embedding with a flattened one-hot encoding of the amino acids of the Gp2 paratope. While a linear (ridge regression) model obtained moderate performance (0.025/0.033), a nonlinear model (flattened sequence with dense layers between sequence and assay score) was able to achieve equal performance with the linearly-embedded amino acid models (0.016/0.019). We then visualized the relative correlation of the convolutional model's predicted versus actual assay score (Figure 2c). We found that the model was not equally predictive across assays, with the most accurate performance for the on-yeast protease assay.

4.3.2 Testing Transferability to Traditional Developability Metric

Having trained a series of protein embeddings trained on and capable of predicting HT developability assay performance, we next asked if the same embedding could be transferred to predict a traditional metric of developability. Keeping the embedding steps constant (Figure 1, Model A1), we then fit a separate top-model (Figure 1, Model B) to predict the Gp2 yield in two *E. coli* strains via multitask learning using a one-hot encoded strain identifying vector. We used attempted both linear (ridge regression) and nonlinear models (support vector machine and random forest) to account for possible complex interactions between the embeddings and yield.

We found that transferring embeddings trained via assay scores resulted in the prediction of yield more accurately than a traditional one hot embedding. During cross validation, the recurrent embedding with a random forest top model (CV MSE: 0.46) and the convolutional model with an SVM top model (CV MSE: 0.47, Figure 3a) exhibited optimal performance. Upon evaluation of an independent test set (Figure 3b), the convolutional embedding with an SVM top model produced the most generalizable model (Test MSE: 0.53, Figure 3c) while the recurrent embedding suffered from overfitting (Test MSE: 0.68). Compared to the one hot model with a forest top model (CV MSE: 0.60, Test MSE: 0.67), the convolutional embedding reduced the gap to experimental variance (0.36) by 45%. Additionally, the convolutional embedding was also able to outperform a model trained on experimentally measured assay scores (CV MSE: 0.50, Test MSE: 0.56) suggesting the embedding can avoid errors that occur from the non-perfect representation of the proxy HT assays to yield.

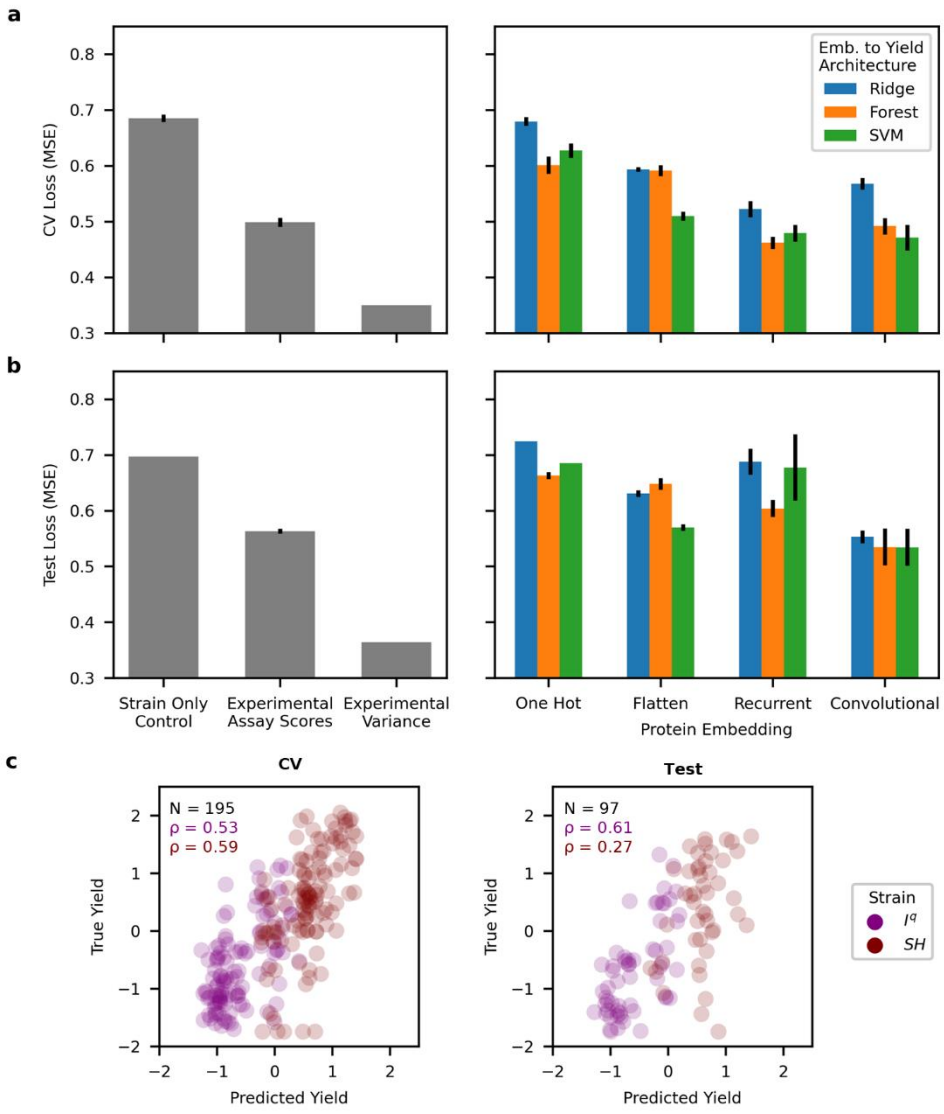

Figure 4.3 - Transferred convolutional embedding predicts yield more accurately than traditional embedding strategy

a) Cross validation and **b)** Test performances of predicting yield comparing a traditional one hot embedding to protein inspired embeddings trained by HT assay scores. **c)** The convolutional embedding with a support vector machine top model's prediction of yields versus experimentally measured yield.

4.3.3 Alternative Model Building Approaches

We next asked if the assay score predictions could be used to predict yield rather than using an intermediate hidden state (Figure 4a). Despite the accuracy of assay score predictions during training (see Figure 2B), none of the architectures' assay score

predictions were accurate enough when fed into a previously trained assay score to yield model (Figure 4b). Upon inspection of the predictions, while the on-yeast protease assay scores were predicted well (ρ = 0.84), the inaccuracy of the split GFP (ρ = 0.41) and split β-lactamase assay (ρ = 0.02) produced incorrect assay scores (Figure 4c). The incorrect scores, compounded with a non-perfect assay score to yield model, likely resulted in the poor performance in predicting yield.

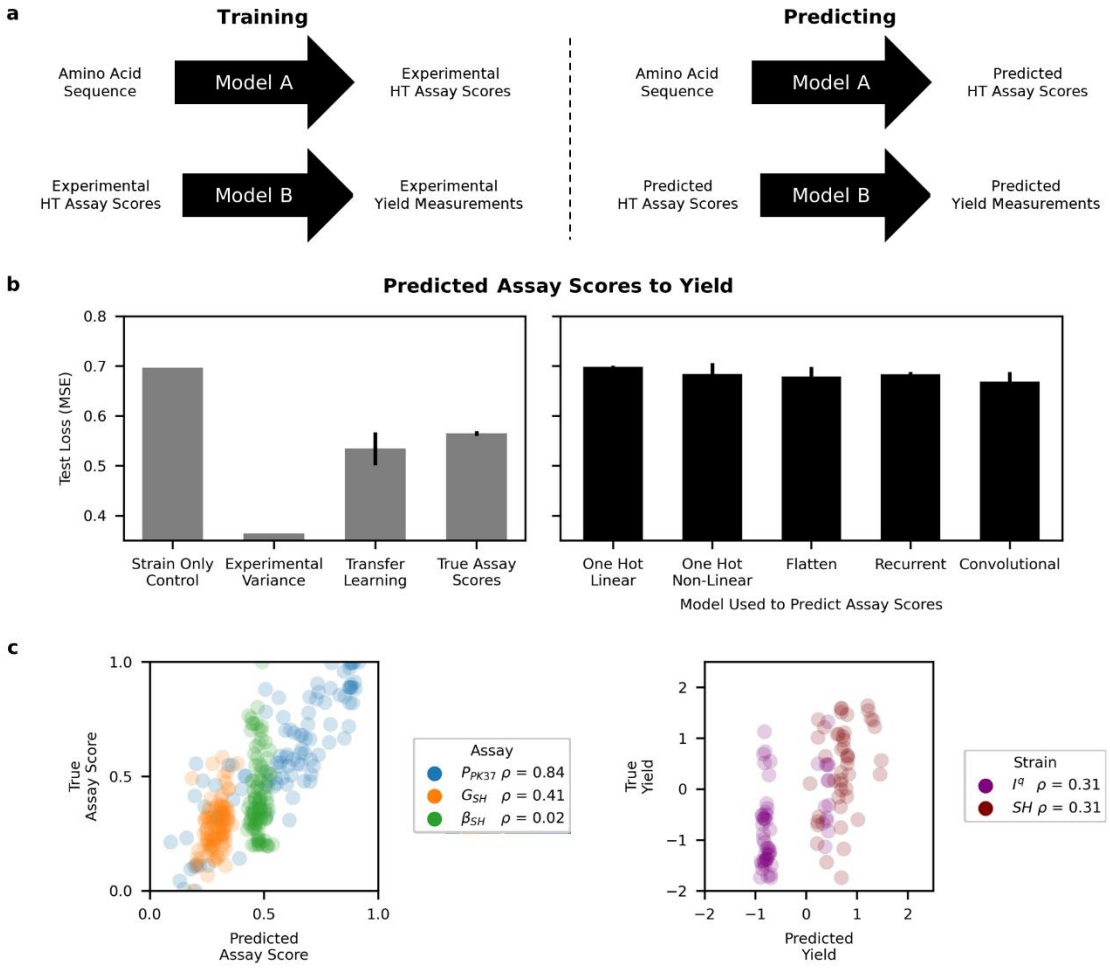

Figure 4.4 - Predicted assay scores not accurate enough for yield predictions

a) In this approach, two models are trained in parallel to predict yield from amino acid sequence. The predicted assay scores are used to predict yield, compared to an intermediate hidden state during transfer learning. **b)** Accuracy in yield measurements is compared to the transfer learning approach using the convolutional embedding and a model using experimentally measured assay scores. **c)** Comparison of predicted versus actual assay scores (left) and yield predicted from assay scores (right) for the convolutional embedding model.

We next asked if we could have fit a sequence-based model on the yields predicted by experimentally measured assay scores (Figure 5a). The set of 45,433 assay scores were converted to predicted yield in both bacterial strains and then used to train models with the same architectures used when predicting assay scores (Figure 5b). All architectures displayed a strong ability to learn the predicted yields with cross validation losses ranging from 0.093 (convolutional) to 0.107 (one hot non-linear). However, upon evaluation of the independent test set of sequences all models displayed high levels of overfitting with test losses ranging from 0.615 (convolutional) to 0.664 (one hot linear). Further visualization confirms the ability of the convolutional architecture to match the assay-score predicted yields, but the inaccuracies of the assay-score predicted yields to true yields caused overfitting (Figure 5c).

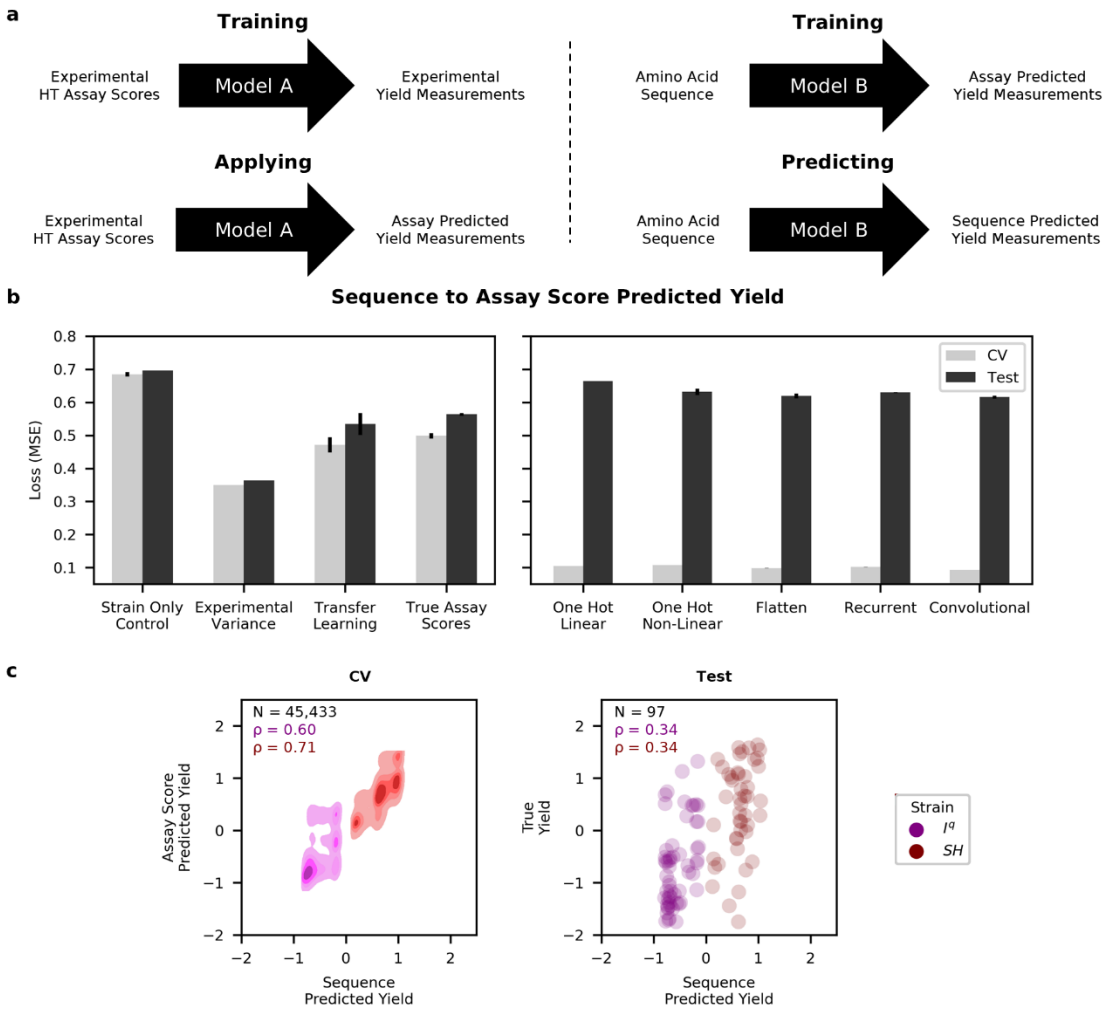

Figure 4.5 - Models trained on yields predicted from experimentally measured assay scores display overfitting

a) In this strategy, Model A is first trained to predict yield measurements from HT assay measurements. Then model A is used to predict the yield to build a database of assay predicted yield measurements. Model B is then trained to predict yield from the large database of assay predicted yield measurements from amino acid sequence. Once trained, the model B can then predict yield from sequence. **b)** A series of models were trained on set of 45,433 sequences' predicted yields from experimentally measured assay scores. The generalizability was then determined by the prediction of an independent test set of 97 sequences. **b)** The convolutional architecture could learn the assay score predicted yields but failed to generalize to experimentally measured yields.

The two alternative approaches for yield prediction differed from the transfer learning approach by directly building on the assay score to yield model. While the assays were related to yield, the inaccuracies of the model generated a bias that could not be corrected. The transfer learning model was able to overcome this by generating a

representation of proteins useful for developability metrics but training an assay-score independent top model. We believe that the differences in the approaches are largely dependent on the relationship between the assay scores and yield, where an accurate assay-score to yield model would likely produce similar results across strategies.

4.3.4 Dependence on sample size

We next desired to understand the relationship between the size of the training datasets and the accuracy of the model. To this end, we randomly subsampled unique sequences from the HT assay dataset to develop convolutional embeddings and compare performance to the one hot encoded architecture (Figure 6a). We found that the convolutional embedding always outperformed the one hot embedding at each sample size. However, the gap between the two models decreased with increasing training data, suggesting that with enough data, a simpler model can learn the same information of a more complex architecture. Conversely, at lower sample size, a more complex architecture was capable of learning information useful for protein developability predictions.

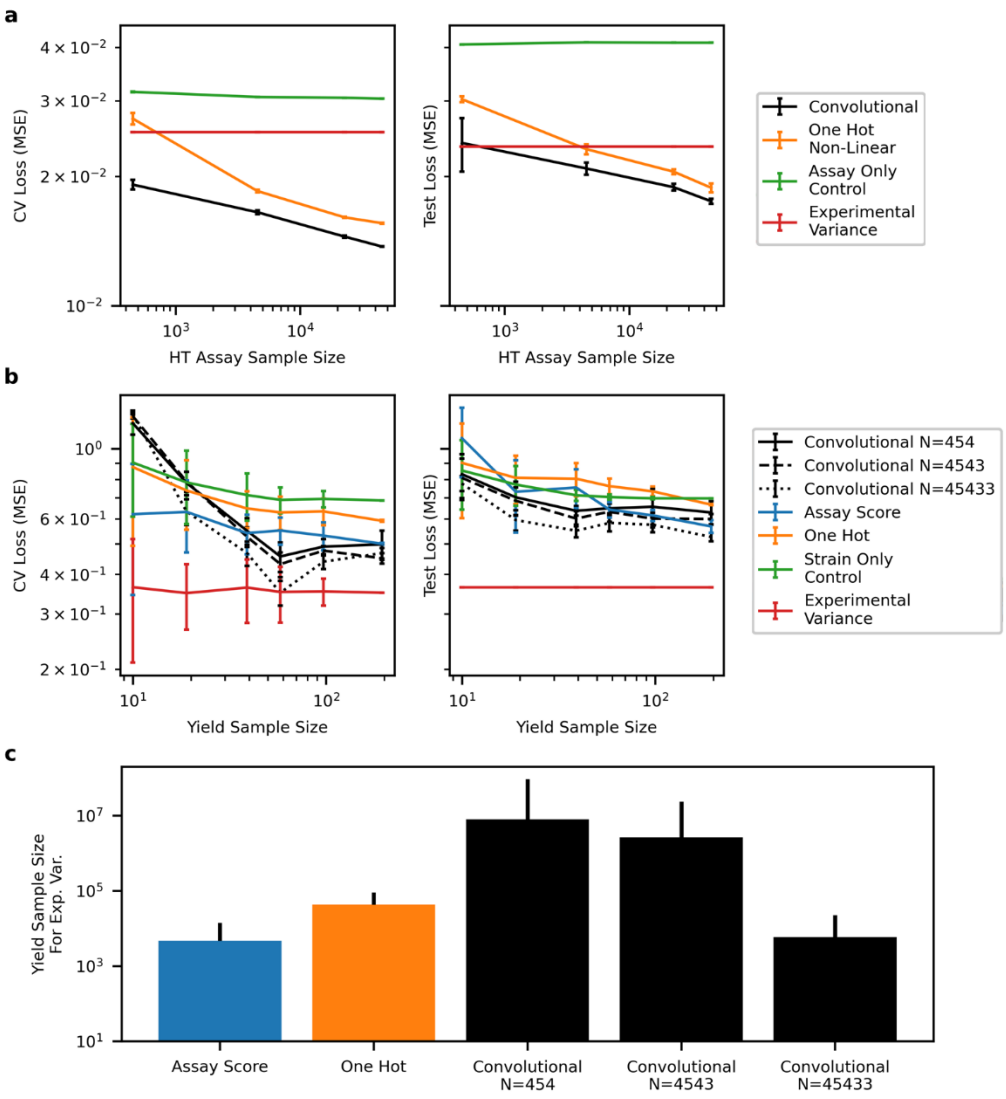

Figure 4.6 - Transfer model benefits from increase of sample size in both training steps
a) The convolutional embedding was trained on random subsets of the HT assay data. While performance improved with sample size, the relative performance over traditional embedding decreased. **b)** The convolutional embeddings from (a) were used to predict yield with top models trained via random subsets of available data. **c)** The predicted number of yield measurements to obtain a model with error matching experimental variance was extrapolated via log-log line of best fit weighted by inverse of the confidence at each sample size. Error bars are propagated from the standard error of slope and intercept.

The convolutional embeddings using 1%, 10%, and 100% of the HT assay data were transferred to the task of predicting yield (Figure 6b). Top models were then trained using 5%, 10%, 20%, 30%, 50%, and 100% of the yield training data. At all fractions of yield data, performance was improved with embeddings trained on more data suggesting

the larger HT assay dataset learned a more universal description. Finally, we assessed the efficiency of the embedding by estimating how many unique sequences it would require to achieve predictive accuracy within the experimental variance of yield measurements (Figure 6c). We found that the convolutional embedding trained by the full amount of HT assay data learns 90 ± 40 % more efficiently than the one hot embedding only requiring $(0.6 \pm 1.7)\times10^4$ unique sequences compared to the $(4 \pm 5)\times10^4$ sequences required for the one hot model. Interestingly, the embeddings trained by subsets of the HT assay data were less efficient than the one hot model, and all embeddings were less efficient than the model trained directly on assay scores which required only $(0.5 \pm 0.9)\times10^4$ sequences to achieve the same result.

4.3.5 Dependence on HT Assays

Having observed the success of transfer learning utilizing all three HT assays, we desired to i) understand the importance of each individual assay in creating a transferable embedding and ii) understand if the transfer learning approach can always create a more accurate embedding than the direct use of HT assay scores. Each combination of HT assays was used to fit the three embedding architectures utilized in this study (flatten, recurrent and convolutional). The three top model architectures (ridge, random forest, SVM) was trained on each HT assay combinations' embedding, where the results of the optimal top model for predicting yield was selected for comparison (Figure 7a). The combination of all three HT assays created the optimal model. Combinations utilizing the on-yeast protease assay resulted in losses lower than those without ($p < 0.01$, independent 2-way Student's t-test). In fact, the assay alone only increases error 2% from the model utilizing all three

HT assays. This suggests the on-yeast protease assay is the most informative assay and could potentially be used independently in future studies.

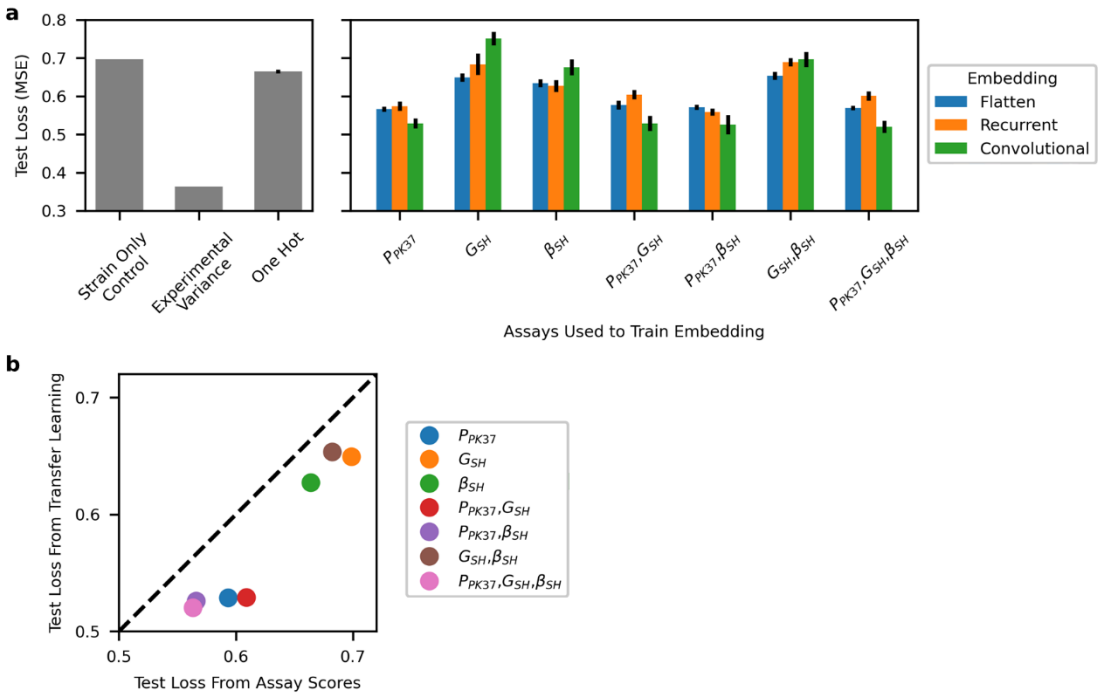

Figure 4.7 - On-yeast protease assay is most informative and transfer learning enables discovery of true signal from inaccurate HT assay proxies

a) A developability representation and top model to yield was trained with combinations of HT assays. The prediction error of sequence yield is grouped by assay combination and colored by embedding architecture. Error bars represent standard deviation of loss from N = 10 stochastically trained embeddings and top models. **b)** Yield predictions from assay scores and the most accurate trained embeddings for each combination of HT assays suggests transfer learning more accurate than direct representation of the assay output.

The ability of the transfer learning training strategy to identify developability trends and average out noisy signals from similar sequences enable predictions more accurate than direct use of HT assay outputs. Having seen the ability of the transfer model outperform prediction from experimentally measured assay scores (see Figure 3b), we desired to understand if transfer learning was successful because of the use of multiple assays and/or a large training database to learn a more generalizable representation. To answer this question, we plotted the model accuracy trained directly on experimentally

measured assay scores to the model accuracy that utilized the assay scores to train a representation that was transferred to predict yield (Figure 7b). We observed a correlation (Spearman's $\rho = 0.96$) between the losses, suggesting that the more relevant assay score combinations enable more accurate embeddings. We also observed a significant decrease of loss from transfer learning models to models trained directly on assay scores (paired t-test $p < 0.01$). The ability of transfer learning to always outperform assay score models, even when a single assay is used, suggests the model can utilize sequence information to denoise errors present in the assay output.

4.3.6 Model Interpretability

Having trained an accurate model (transfer learning utilizing a convolutional embedding with all three HT assays and an SVM top model to yield, now referred to as DevRep) for the Gp2 developability as characterized by soluble yield in *E. coli*, we desired to explore model parameters and visualize the landscape.

4.3.6.1 AA Embedding

First, we analyzed the trained amino acid embeddings to determine what properties were most relevant to Gp2 developability (Figure 8a). The 17 trained properties were distilled down to three principal components (PC) which explained 68% of the total variance. Upon inspection, we determined that cysteine was uniquely separated in PC 1 and 2. Additionally, PC 2 appeared to separate the remaining residues by hydrophobicity by placing aromatic and aliphatic residues separated from polar and charged residues. PC 3 further separated hydrophilic residues into negative, neutral, and positively charged. Interestingly, histidine (which possesses a pKa near experimental conditions) is located closer to neutral amino acids compared to arginine (R) and lysine (K), commenting on the

ability of the model to learn about both charged states. We then compared each PC to the AAindex[166] list of properties in an attempt to find the most correlative physicochemical property: PC 1- coefficient over single-domain globular proteins ($\rho = 0.91$), a measurement of hydrophobicity[170] again underscoring its importance on developability.; PC 2- normalized frequency of N-terminal non-beta region ($\rho = 0.86$), a measurement of residue frequency in nonstructured regions[171]; and PC3- helix termination parameter at position j-2,j-1,j ($\rho = 0.83$), a measurement of residue frequency in short helical structures[172]. Together, PC 2 and 3 suggest the paratopes may be balancing between a flexible loop and a short helical confirmation to provide stability.

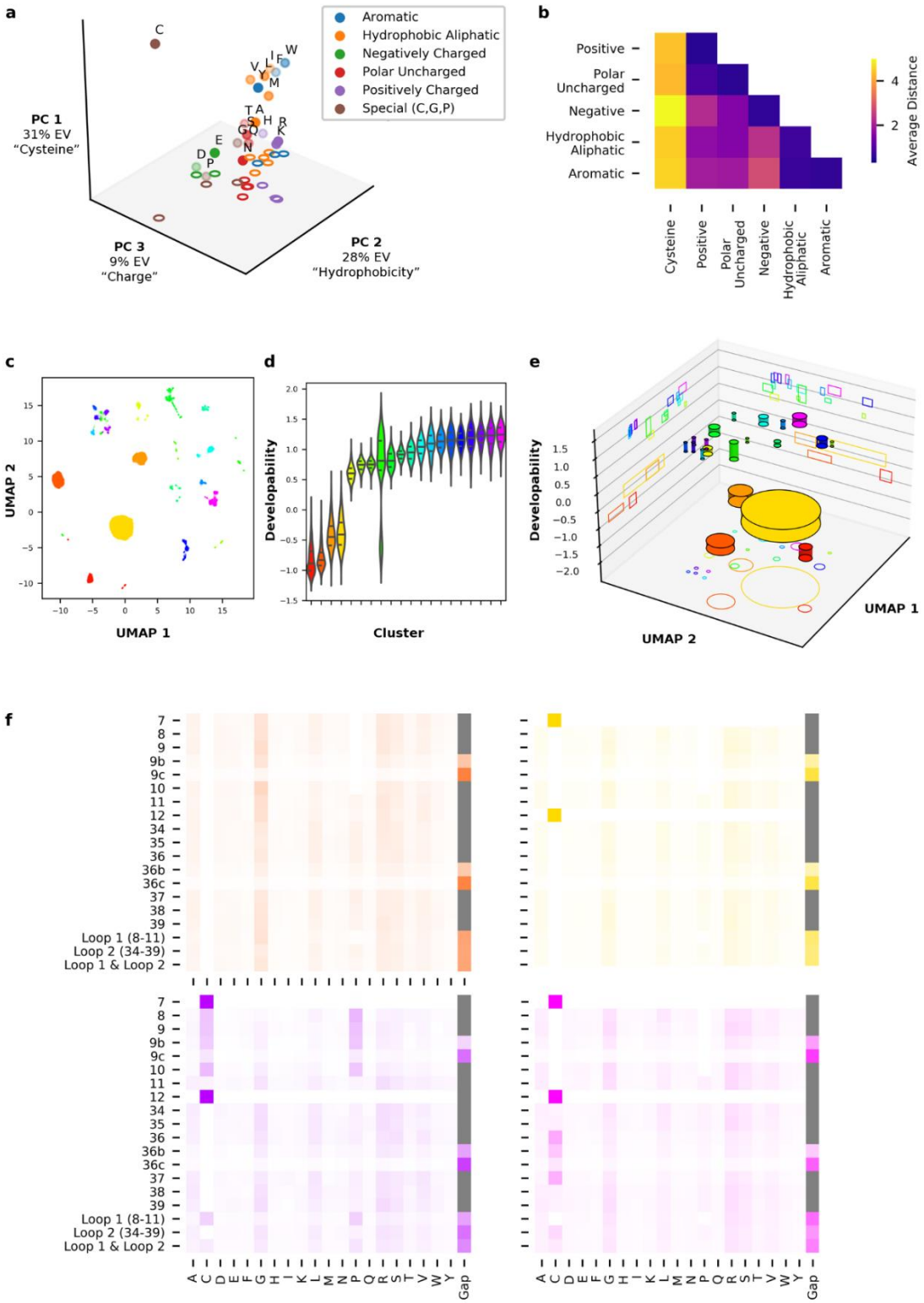

Figure 4.8 - Analysis of trained embeddings reveals properties related to developability

a) Principal component (PC) analysis of the 17 amino acid embedding table colored by category of residue. EV = explained variance. **b)** Inter- and intra- residue category distances highlighting the uniqueness of cysteine and lack of difference between aromatic and aliphatic residues**. c)** Clusters of sequences were identified via UMAP and hdbscan of the 45,000 sequences used for training. **d)** Developability, as predicted by yield, varies between clusters trained by on HT assay scores. **e)** Three-dimensional landscape visualized by cylinders centered at the sequence-mean UMAP location with height representing the interquartile range of developability and radius corresponding to the number of sequences. Low-developability clusters are in the front right portion of the landscape. **f)** Amino-acid distribution of two low-developability clusters (orange and yellow) and two high-developability clusters (purple and pink).

We further evaluated the average inter- and intra-residue PC distances (Figure 8b). Each identified cluster of residues had a lower intra-residue distance than inter-residue distance except for aromatic (F, W, Y) and hydrophobic aliphatic (A, I, L, M, V) residues suggesting the hydrophobic nature of these residues outweighed the relative size difference and additional interaction capabilities of aromatic rings.

4.3.6.2 Location of Training Sequences in DevRep

We next assessed the interaction of the residue embeddings by converting the 97-dimensionial embedding of the 45,000 training sequences via UMAP[173]. We then utilized hdbscan[174] which identified 19 clusters of sequences from the 2-dimensional UMAP space (Figure 8c). We then discovered that the clusters contained information about the variant developability by finding a significant difference in developability distributions as a function of cluster (Figure 8d, Kruskal-Wallis H-test, $p < 0.05$). Interestingly, the location of the 4 lowest clusters were also located close together in UMAP space (Figure 8e). As UMAP can cluster sequences locally and globally, this suggests DevRep places most low-developability sequences in a similar location within the embedding.

Finally, we analyzed the intra-cluster amino acid distribution for select clusters (Figure 8f): Orange - a low developability cluster not containing any cysteines, an increase in glycine, and loops of length 7; Yellow - a low developability cluster with cysteines at

sites 7 & 12 with fewer glycines; Purple - a high developability cluster with cysteines at sites 7 & 12 and proline in the first loop; and Pink - a high developability cluster with cysteines at sites 7 & 12 and also in the second loop. The differences in amino acid frequencies of the selected clusters, paired with the amino acid embedding analysis, suggest the model learned residues interaction, particularly with cysteines.

4.3.7 Comparison to Alternative Protein Embeddings

We next compared the HT assay trained embeddings to other state of the art protein embeddings. The AAindex[166] was used to create an embedding based upon physiochemical properties. As the index is known to contain several similar entries, PC analysis was used to isolate down 3 and 10 residue properties for which the paratope sequence was transformed and flattened. We also compared DevRep to three embeddings trained on evolutionary properties: TAPE[167]'s transformer embedding which was trained on the Pfam[175] database via predicting masked residues, UniRep[60] which was trained on the UniRef[176] database via predicting the next residue in a recurrent-style architecture, and evolutionarily (evo) tuned UniRep via isolating homologous sequences to Gp2 via HMMER[177] and updating via Jax-UniRep[178] software. All evolutionary embeddings were tested by averaging over either the full sequence or the paratope sequence.

Each embedding was trained to predict yield utilizing the same architectures and hyperparameter search strategy as DevRep (Figure 9a). We found that DevRep was able to predict yield more accurately than every other embedding. The evolutionary based embeddings (particularly UniRep paratope) were able to predict yield more accurately than the strain only and one hot controls, suggesting that developability contains similar information as contained in the embedding, but not as much information as the HT assays.

The poor performance of AAindex suggests traditional physicochemical properties are not the best way to describe Gp2 variants regarding developability.

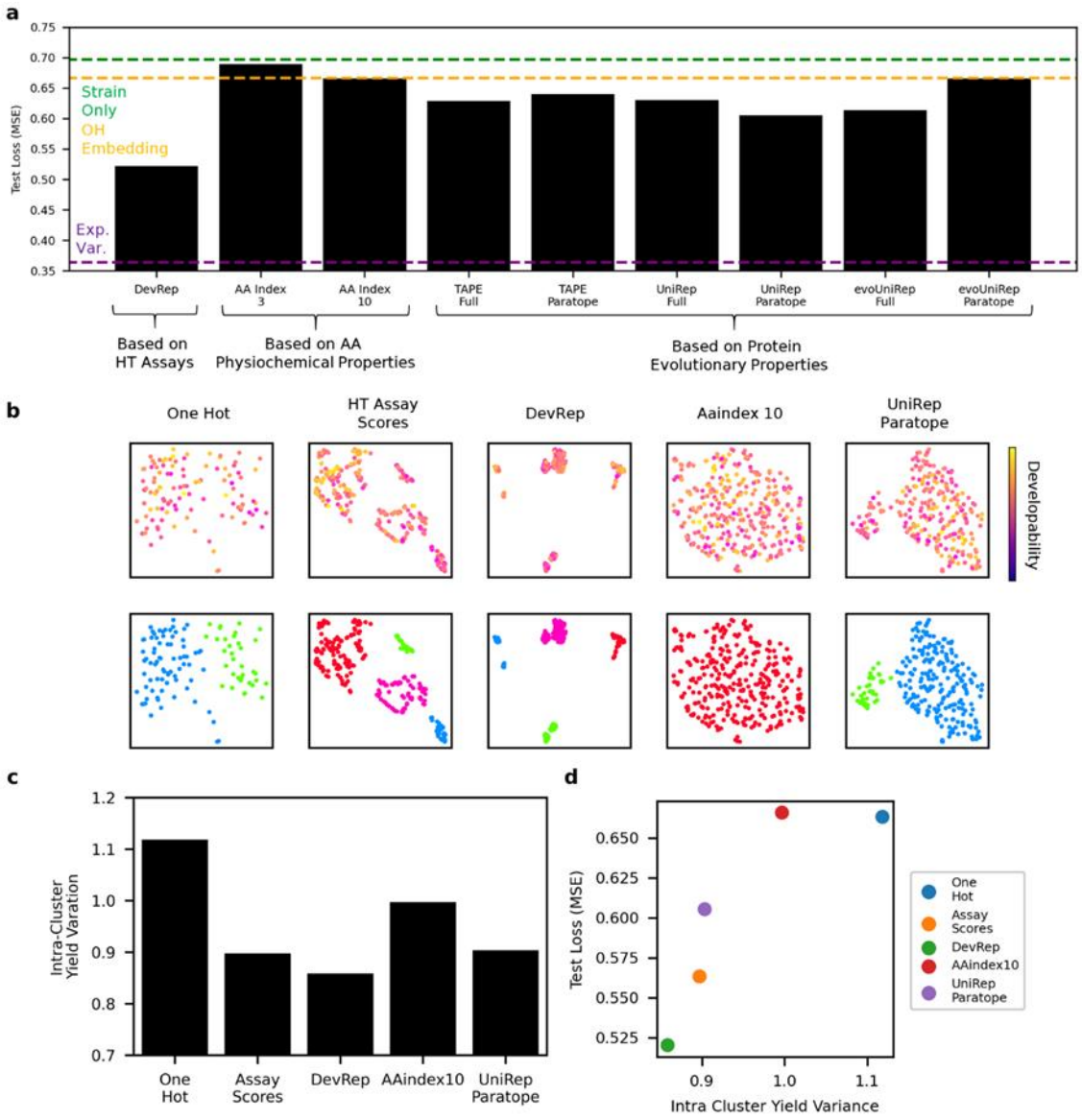

Figure 4.9 - HT assay trained embedding contains more developability information than alternative embeddings

a) Comparison of protein embeddings' ability to predict yield as represented by the loss of an independent set of sequences. **b)** Variants were plotted using UMAP for each embedding. *(top)* Color represents experimentally measured developability. *(bottom)* Sequences were clustered by UMAP coordinates. Color represents unique clusters. **c)** A low yield variance across sequences in clusters determined by embedding suggests shared information. **d)** The correlation between the embedding's developability information and the model's (trained using the same embedding) predictive performance confirms success of model training.

To ensure the superior performance of DevRep in predicting yield was not due to poor model development, we assessed the relationship between position of sequences in each embedding to the measured developability. The 195 unique sequences with experimentally measured yield were embedded and transformed for visualization via and UMAP (Figure 9b). Clustering via location was then performed and the average intra-cluster variance of yield was calculated to estimate how much information about developability was stored in the embedding (Figure 9c). We found that the HT assay scores and DevRep's UMAP representation contained the most information about yield (Figure 9d). The most yield-information containing embedding is expected to enable the most accurate predictions if the model can interpret the information. We found a correlation between the embedding's intra-cluster yield variance and predictive performance of trained models, suggesting the models were fully trained and the limitation was the information contained in the embedding.

4.3.8 Phase Space Analysis via Nested Sampling

Rather than rely on the skewed experimentally observed distribution of developability, we sought to use nested sampling to systematically characterize the structure of the fitness landscape while identifying highly developable sequences. At every iteration, nested sampling reduces the fraction of available sequence space "volume" by a constant amount. As a result, we can use the output of nested sampling (a list of threshold sequences and their associated yield) to compute the density of states (DOS) as a function of developability. Put simply, we can estimate the relative number of sequences available at any given developability. Computation of the DOS also allows us to determine analogs of thermodynamic properties such as entropy, mean developability and heat capacity (i.e.,

susceptibility with respect to a temperature-like quantity that in this context modulates the ability to mutate) to identify the occurrence of phase transitions[179,180]. We ran the algorithm with 100 live sequences, removing the lowest yield sequence and 0.99% of the phase volume at every loop until convergence to a single sequence (Figure 10a). We then utilized the DOS to identify two phase transitions, between which the sequences split into multiple subpopulations that compete at a critical temperature. Signaled by peaks in the heat capacity, the expected values of developability corresponding to the critical temperatures have a developability of ~1.25 and 2.0 (Figure 10b). This phase transition occurs with only 10^{-5} - 10^{-10} of all sequencies predicted to have a higher yield.

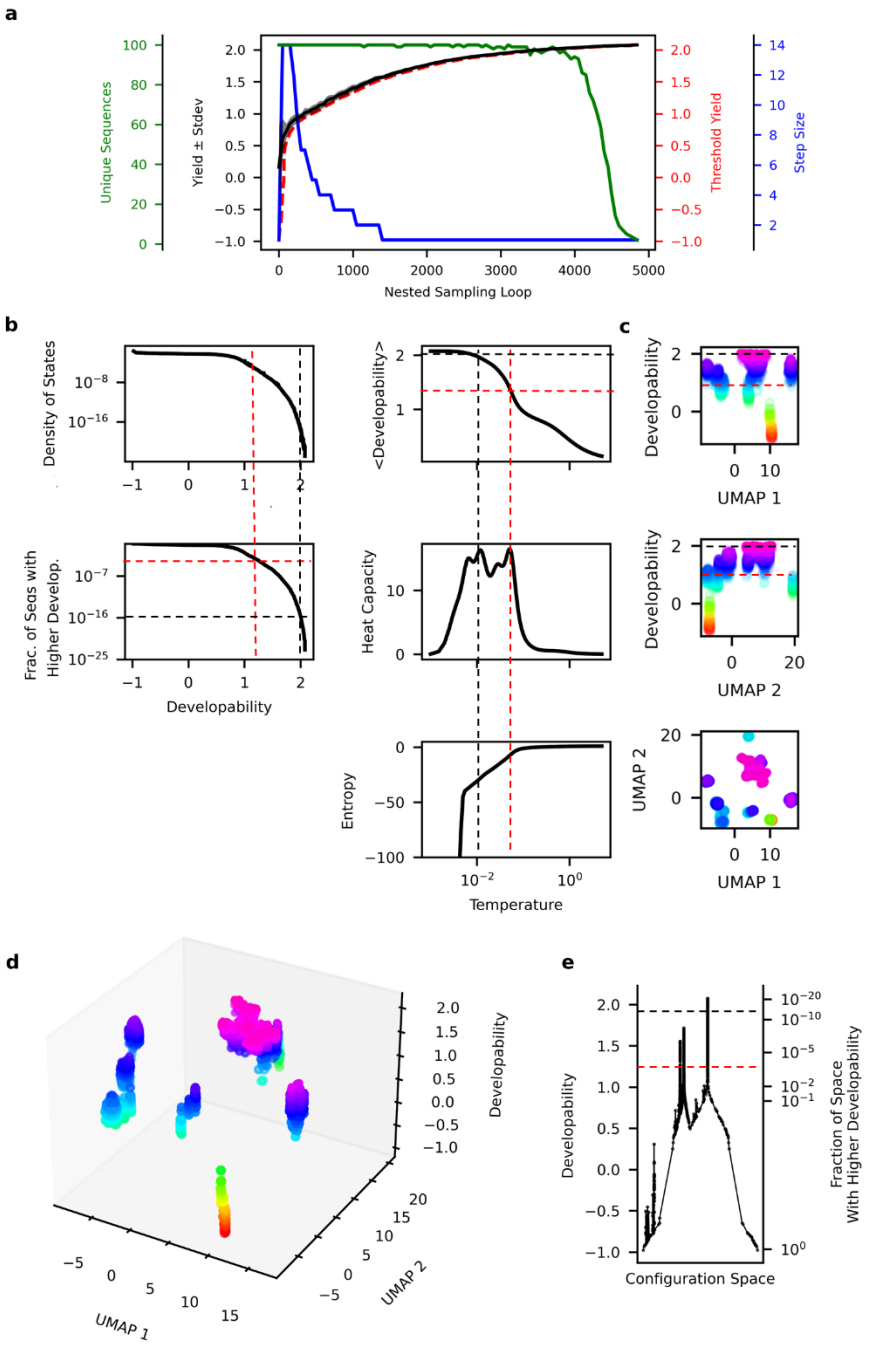

Figure 4.10 - Nested sampling explores the developability-sequence landscape

a) Nested sampling was performed using 100 active sequences while accepting mutations with corresponding yields above the threshold per round. The threshold sequence, and corresponding yield, was determined by the lowest yield of the live sequences. **b)** The density of state for each level of developability was determined and used to estimate the expected developability, heat capacity, and entropy at various temperatures corresponding to the ability to mutate. Phase transitions are identified with a dashed line. **c,d)** The UMAP representation displays the landscape splitting into distinct clusters of DevRep space above the transition. **e)** The disconnectivity plot displays a sudden contraction with a rugged landscape at the phase transition.

The output of nested sampling can also be used to visualize the phase space by plotting the relationships between sampled variants[181,182]. Plotting the sequences in UMAP space shows a single stalk of low developability sequences up to the phase transition where several high developability clusters exist (Figure 10c,d). The split suggests that beyond the phase transition, there exists several distinct modes of achieving high developability. A disconnectivity plot was synthesized by creating a graph of nearest sequences of higher yield based upon the UMAP transformation of the DevRep embedding (Figure 10e). The phase transition at developability corresponds to a sharp decrease in configuration space with disconnected subgraphs of sequences.

We compared disconnectivity plots and UMAP landscape to the one hot and UniRep Paratope models' embeddings (Figure 11). Every model suggests a steep contraction of configuration space. The DevRep landscape is the only embedding to show a large split of sequence space. The one hot UMAP landscape appears to have sequences of various predicted developability located at every UMAP location, confirming the one hot embedding lacks easily interpretable developability information. The UniRep paratope landscape does show correlation between UMAP 1 and developability, suggesting there is some shared information.

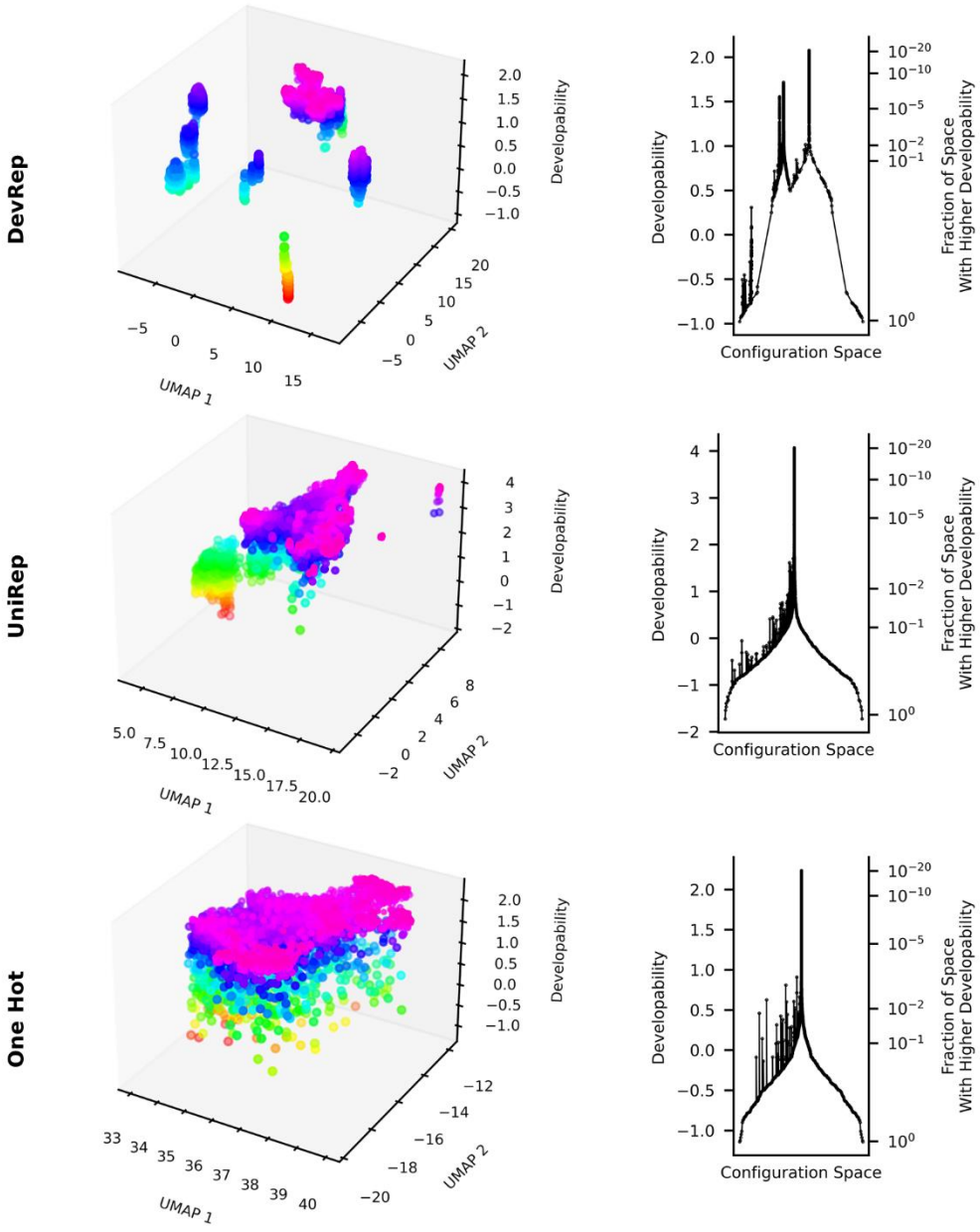

Figure 4.11 - Comparison of developability information contained in embedding

Nested sampling was performed with each model. *(left)* UMAP representations of sequences sampled during sampling have varying relationships with predicted developability. *(right)* Disconnectivity plots from different models agree with a sudden constraint in configuration space of the top 10% of sequences but disagree in the ability to identify a large split of sequence space.

4.3.9 Identification of Top Developability Variants

As a final test of the transfer model approach to predict protein developability, we desired to measure the ability to predict high developability variants. Because we found

that the Gp2 library splits into many subgroups of sequences that can achieve high developability, we also focused on generating diverse sequences. We also identified sequences using simulated annealing[183] to compare search strategies. The embeddings from each sampling approach were reduced via UMAP and clustered via hdbscan to identify sequences from clusters diverse in DevRep space. We chose 100 variants equally sampled across the top of each identified cluster (Figure 12a). The same process was repeated with the one hot model and UniRep and Paratope models and embeddings. A randomly generated set of sequences was also tested for comparison. The predicted yields and different location within each embedding for the isolated sequences suggests each model has its own maximum (Figure 12b).

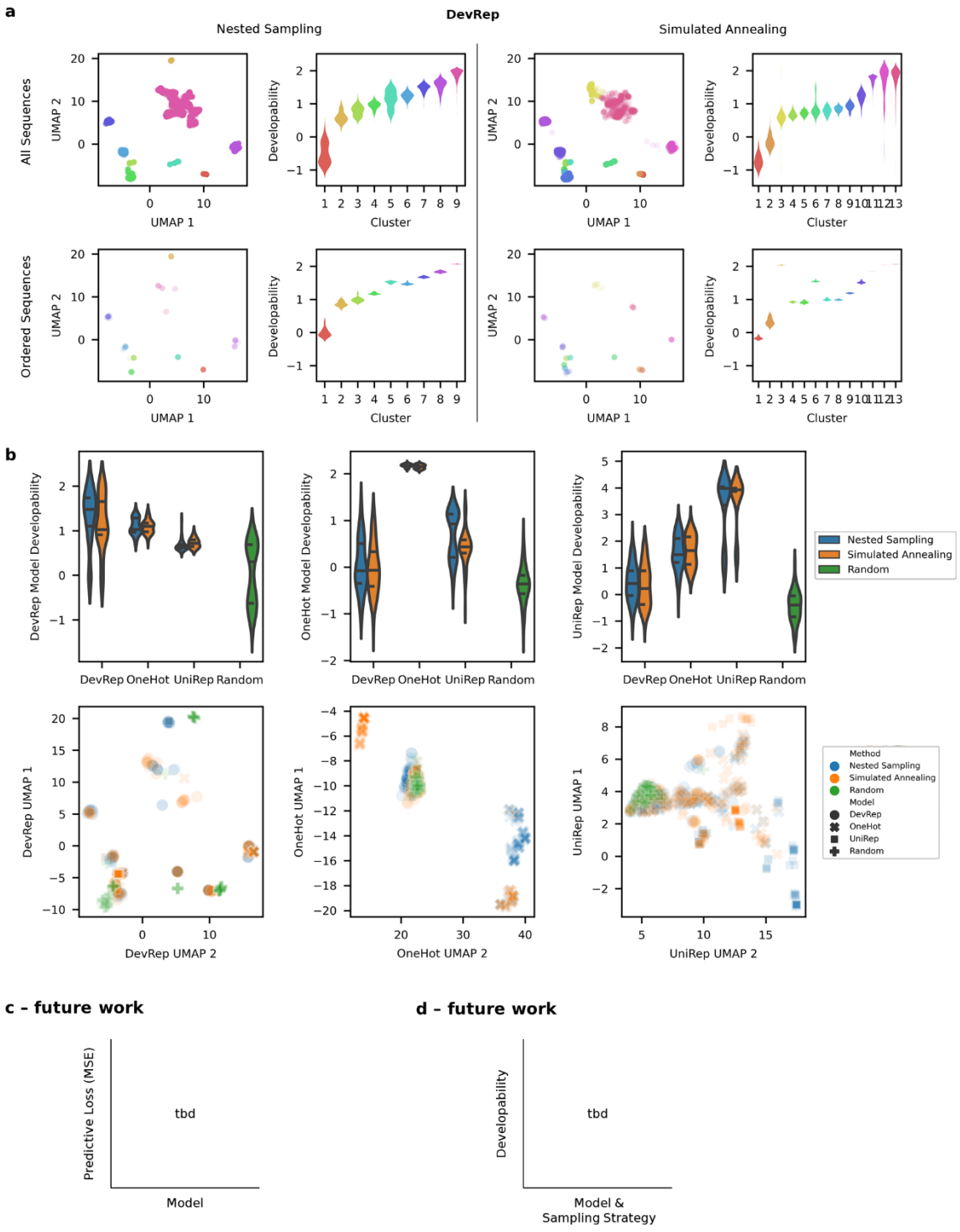

Figure 4.12 - Assessment of models to predict high developability variants

a) Sequences identified through each sampling strategy were clustered via embedding. The top sequences within each cluster were ordered to obtain a diverse set of sequences for experimentation. **b)** Sequences ordered from each model were not predicted to be optimal in other models and were unique. **c)** The accuracy of each model to predict novel sequences will be evaluated as a final test of predictability. **d)** The distribution

of experimentally measured developability will be compared across model and sampling strategy to determine the best approach to identify high developability sequences.

It was observed that including sequence diversity in the selection scheme introduced lower developability variants. Additionally, large clusters of high developability sequences were observed in both DevRep and UniRep embeddings during nested sampling (Figure 4.13). Thus, for each model, the large cluster was split into subclusters where 100 additional variants were ordered equally spread across the high-developability subclusters.

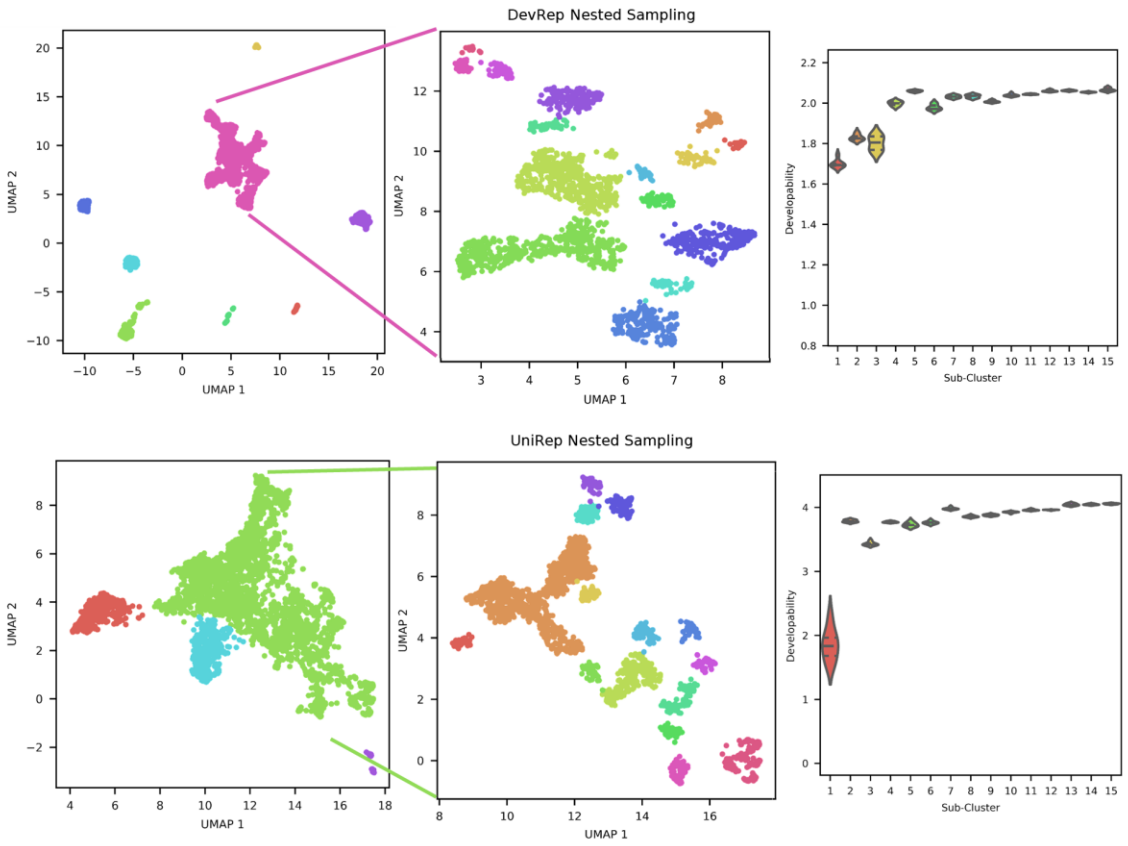

Figure 4.13- Selection of additional high developability variants

Large clusters of highly developable variants were observed while clustering all recorded sequences during nested sampling. Thus, the clusters (left panel, DevRep - Purple, UniRep - Green) were broken into subclusters (middle panel) using the same methodology. The top sequences within each subcluster were selected for additional screening.

4.3.10 Work in Progress

Variants will be produced in a similar method to the training data. To assess the accuracy of the models, we will again compare the error in the prediction of newly produced sequences (Figure 12c). Should DevRep not be the most accurate in prediction, a suggestion for improvement would be training additional model architectures or attempting to improve the UniRep embedding by tuning parameters with HT assay scores. We will also access each model and sampling strategies' ability to identify unique and highly developable sequences by comparing the distribution of yields (Figure 12d). While nested sampling enables phase space visualization, simulated annealing theoretically allows for more exploration. The model that identifies the highest yielding sequences is hypothesized to best represent the landscape in terms of the ability to explore and find the maximum.

4.3.11 Preliminary Results

The following section contains data resulting from a single trial of yield measurements. We assayed 280 variants in the I^q strain and 269 variants in the SH strain. Using both metrics, the DevRep model was the most accurate in prediction of unseen sequences (Figure 4.14a, MSE: 0.71, ρ I^q: 0.13, ρ SH: 0.21). Interestingly, the One Hot encoded model (MSE: 0.93, ρ I^q: 0.03, ρ SH: 0.25) outperformed the UniRep encoded model (MSE: 1.23, ρ I^q: 0.04, ρ SH: 0.15). All models displayed increased predictive performance over the Strain Only control model (MSE: 1.40).

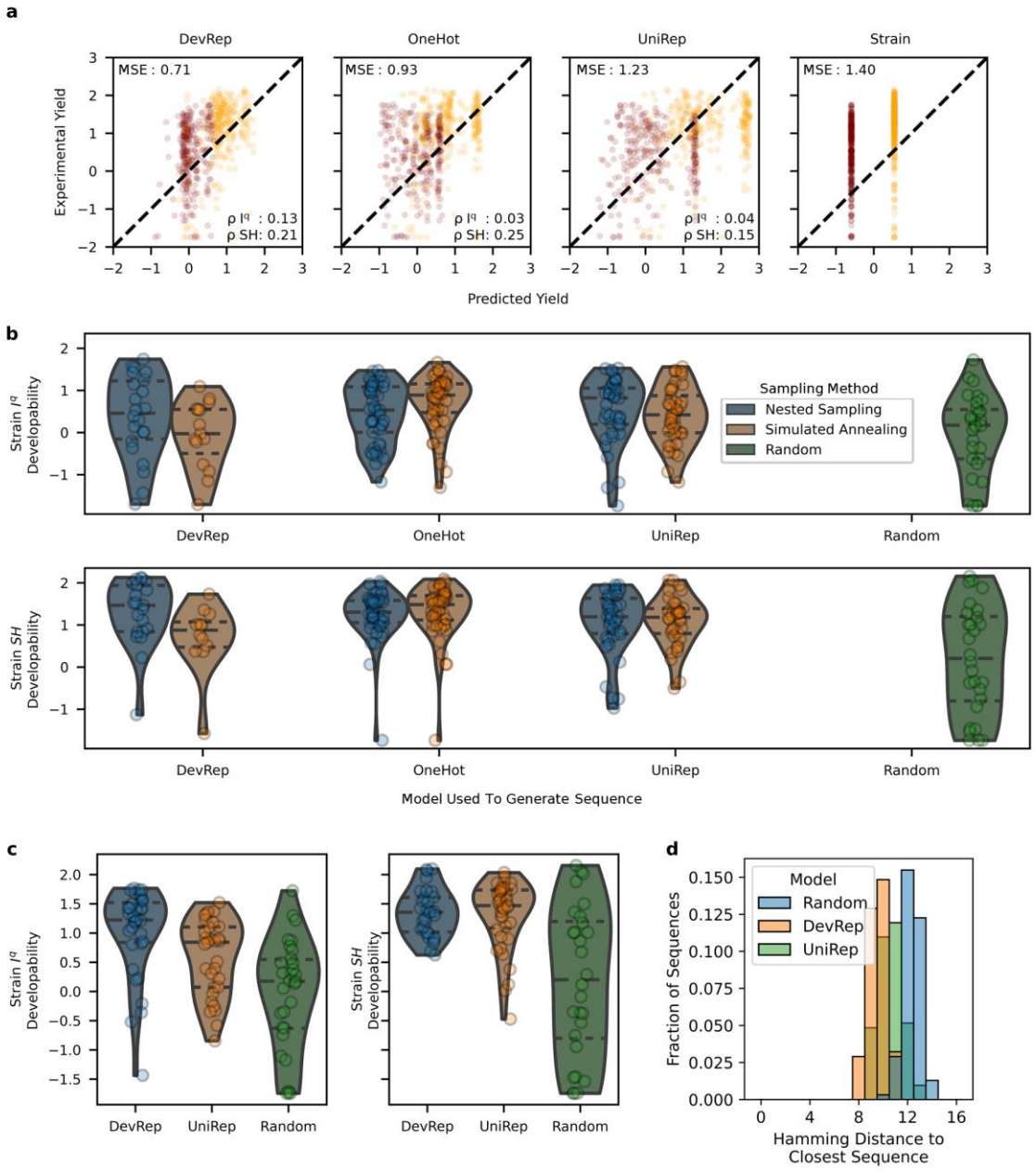

Figure 4.14 - Preliminary results display the abilities of sequence embeddings

a) The predicted versus actual developability of 280 I^q and 269 SH variants identified via sampling strategies (see Figures 4.12 and 4.13). **b)** Sequences generated by each embedding and sampling strategy are compared to each other and to a selection of randomly generated sequences. **c)** An additional set of sequences identified via nested sampling of DevRep and UniRep were also compared. These sequences were designed to be more developable and more similar in embedding space. **d)** Each sequence in (c) was compared to the set of sequences with measured yield that was used during model training. The distribution shown is broken down by the model used to generate the sequences.

We next assessed which model and sampling technique identified the top performing variants with additional focus on diversity (Figure 4.14b). While nested

sampling outperformed simulated annealing for DevRep by identify higher yielding sequences (t-test, $p < 0.05$ for both strains), this trend was not true across all models and strains. All models and sampling strategies were able to produce a sample of sequences with a higher median developability than when samples were chosen randomly. However, when comparing the embedding strategies to each other, there is no clear winner in terms of producing the highest yield variants.

We then assessed the distribution of yields obtained with a higher focus on developability than diversity (see Figure 4.13). Again, both DevRep and UniRep embeddings were able to select sequences with higher developability than a random selection (Figure 4.14b). Additionally, DevRep was able to identify the sequence with the highest developability in I^q (Loop 1: CWCPXRPC, Loop 2: NRGAXGGG) and had the highest minimum yielding sequence in SH. Of final note, we found the sequences identified in this final evaluation were significantly far (in terms of Hamming distance) from variants evaluated during model training. Although DevRep's sequences were closer than UniRep and randomly identified, they were still 9.5 (on average) amino acid mutations away from the closest sequence during training.

These preliminary results display a promising utility of DevRep in terms of both predictive accuracy and the ability to identify highly developable variants. Additionally, the performance of both One Hot and UniRep embedding and sampling strategies suggest these techniques could be a useful first step in sequence identification, even prior to experimentation.

4.4 Conclusion

This work evaluated the ability of HT developability assays to train an embedding that is transferable to a traditional metric. We determined that this strategy can smooth out erroneous information found in the proxy assays and train more efficiently than a one-hot control embedding. We then analyzed the model's parameters and predictions to identify unique modes of achieving high developability based upon the location of cysteine and the importance of hydrophobicity and charge. The configuration space was explored via nested sampling which identified a range of developability where the sequences are highly clustered and unique, suggesting a series of sub-libraries may outperform a single design. At present, the transfer learning approach outperforms models based on physiochemical or evolutionary properties confirming developability is a complex and unique property. Further assessment on the ability to accurately identify high developability sequences will further validate the utility of this approach.

Chapter 5 - Concluding Remarks and Future Work

The ability to predict protein function is one of the most studied and unsolved problems in biology. Benefits of an accurate and interpretable model include improving pharmaceutical treatments, laboratory reagents, and enzymes found in a variety of consumer and industrial applications. Engineering these properties is complicated by the immensely large domain of protein space and the complexity and barrenness of sequence-function relationships. The presented work focused on protein evolvability (ability to easily modify functionality) and developability (ability to manufacture while maintaining function). These abstract functions are believed to comprise a multitude of factors impacting performance. Improvements would improve the efficiency of commercialization across vast protein applications. This work evaluated a data-driven engineering approach by experimentally obtaining an information-rich and deep dataset for each function and testing various machine learning approaches for the ability to provide accurate insight.

5.1 Aim 1: Interpreting and Predicting Protein Evolvability

Parameterizing proteins using biophysical properties is problematic as there are practically unlimited number of ways to describe a protein. Even utilizing previously optimized high-throughput evolvability techniques, only seventeen different proteins could be evaluated. The disparity between metrics and samples was addressed by employing dimension reducing techniques that isolated a smaller number of independent signals in the data, lowering the number of trainable parameters. Additionally, a model with a focus on generalizability was created through cross-validated evaluation. The importance of each biophysical metric in the final model was evaluated and determined that a large, spatially separated paratope is ideal for protein scaffold evolvability.

There exist other scenarios which there are far more potential input values than datapoints. The machine learning techniques employed in this paper can be used to ensure the generation of a generalizable and interpretable model. Future studies should always attempt the simplest (linear) models first. The benefit of interpretability generally far outweighs a potentially slight improvement in performance, especially with only 10's of datapoints.

It was found that when testing for evolvability, developability is not guaranteed. Future evaluation of predicted evolvable scaffolds was halted when the most evolvable scaffold from the study exhibited poor developability, limiting characterization capabilities. A threshold in developability to possess evolvability was observed, as was a decrease in stability with functionality. Taken together, it seems advantageous that scaffolds should be developable. At present, the relative importance of the biophysical properties versus developability remains unknown.

The completion of this aim, signified by identifying beneficial properties of evolvable protein scaffolds, supports the hypothesis that a data driven approach can provide accurate insight to protein engineering problems. Future experimentation is likely to increase in scale with increasing molecular technologies with decreasing costs which will aid in resolution of driving properties. Future work may also benefit by editing the metric of evolvability, as the balance between the level of function of the proteins (e.g. binding affinity) and the unique number of functioning proteins remains heuristic. We hypothesize that the next substantial advance in predicting evolvability will result from identifying biophysical properties of high developability scaffolds. Thus, we advocate that the community expand on the database of evolvability in terms of number of scaffolds and the

number of targeted proteins for which novel specific binding is the function of interest. These scaffolds should be diverse in terms of biophysical properties and developability - although upsampling highly developable libraries would be of interest. A more generalized understanding of evolvability can aid in the selection - or *de novo* creation - of protein scaffolds that can readily produce variants with high affinity towards the target of interest.

5.2 Aim 2: Interpreting and Predicting Protein Developability

Without explicitly implementing geometric or evolutionary background information, we attempted to train a model where only the amino acid sequence is used to predict developability via data-driven learning of relevant properties. The data suggests that it would take 10^4 - 10^5 variant measurements (for a library of 10^{20} variants) to train an accurate model for a traditional metric of developability, well beyond traditional capacities. Rather, we gathered developability data by establishing three high-throughput developability assays while simultaneously reducing a major bottleneck in the candidate selection process. Future work should focus on the validity of these assays with a variety of proteins as well as their predictive performance towards other traditional developability metrics.

The ability to transfer the information of developability from the high-throughput assays to predict the traditional metric of interest was nontrivial due to biologic noise and the limitations in high-throughput assays' relevance. We showed that an intermediate representation in a deep-learning model can remove inaccurate signals by predicting with 50% more accuracy than traditional modeling approaches. Finally, we examined parameters of the model to identify various sequence motifs with high developability,

including those utilizing the stability of disulfide bond by including the amino acid cysteine.

We also showed that developability models trained on more relevant functions can outperform models trained on evolutionary or physiochemical information. This further underscores the importance on the relevance of the dataset. It is popularized that deep-learning networks require sufficient depth of data. However, I argue that the quality of information stored in the data is much more important. Future work on validating assays or utilizing the numeric representation of the sequence for a prediction of a new task must consider the relevance of their task to the developability assessed on the high-throughput developability assays.

Completion of this aim, signified by the creation of an accurate deep-learning model which provides insights into the motifs driving developability, again support the hypothesis of data-driven protein engineering. Future laboratory experimentation should be aimed towards identifying nonredundant high-throughput measurements of developability while computational experiments should aim to discover the most datapoint efficient model architectures. A final limitation of deep-learning for protein engineering is the language-barrier between topics. However, as more and more studies begin combining the fields, this obstacle is likely to vanish. Nevertheless, it is vital that scientists of all backgrounds focus on the clarity of communications to accelerate collaborative and synergistic research.

5.3 Final Statements

There are several methods to engineer proteins. We have shown that data-driven engineering is a successful technique when modifying abstract functionality such as

evolvability and developability. At present, the largest limitation to the success of data driven protein engineering remains the ability to obtain a large information-rich dataset and extract interpretable information from the model. However, the techniques presented in this document were able to improve performance and provide insightful feedback with the amount of data collected in a university laboratory within five years.

Continued validation and implementation of the methodology presented here could revolutionize the protein commercialization pipeline. With an understanding of protein scaffold evolvability, novel scaffolds can be selected to increase the hit rate of finding a functional molecule. With an understanding of protein scaffold developability, functional variants can be computationally sorted to reduce experimental efforts. There even exists the potential to switch the order of operations by first creating a library of high developability candidates which can subsequently be sorted for a variety of functions without need for further developability assessment. Future work in determining the relationship between evolvability and developability should address such possibilities.

References

1. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science (80-.).* **181**, 223 LP – 230 (1973).
2. Rick, W. Trypsin. in (ed. Bergmeyer, H. U. B. T.-M. of E. A. (Second E.) 1013–1024 (Academic Press, 1974). doi:https://doi.org/10.1016/B978-0-12-091302-2.50099-2
3. Saraswat, R., Verma, V., Sistla, S. & Bhushan, I. Evaluation of alkali and thermotolerant lipase from an indigenous isolated Bacillus strain for detergent formulation. *Electron. J. Biotechnol.* **30**, 33–38 (2017).
4. Zhang, L. & Gallo, R. L. Antimicrobial peptides. *Curr. Biol.* **26**, R14–R19 (2016).
5. Kennedy, P. J., Oliveira, C., Granja, P. L. & Sarmento, B. Antibodies and associates: Partners in targeted drug delivery. *Pharmacol. Ther.* **177**, 129–145 (2017).
6. Elgert, K. D. *Immunology : understanding the immune system*. (Wiley-Blackwell, 2009).
7. Li, C., Zhang, R., Wang, J., Wilson, L. M. & Yan, Y. Protein Engineering for Improving and Diversifying Natural Product Biosynthesis. *Trends Biotechnol.* **38**, 729–744 (2020).
8. Lutz, S. & Iamurri, S. M. Protein Engineering: Past, Present, and Future. *Methods Mol. Biol.* **1685**, 1–12 (2018).
9. Kintzing, J. R., Filsinger Interrante, M. V & Cochran, J. R. Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment. *Trends Pharmacol. Sci.* **37**, 993–1008 (2016).
10. Wurth, C., Demeule, B., Mahler, H.-C. & Adler, M. Quality by Design Approaches to Formulation Robustness—An Antibody Case Study. *J. Pharm. Sci.* **105**, 1667–1675 (2016).
11. Zhang, J. Protein-length distributions for the three domains of life. *Trends Genet.* **16**, 107–109 (2000).
12. Yuksel, A. *et al.* An Overview of Thermal and Mechanical Design, Control, and Testing of the World's Most Powerful and Fastest Supercomputer. *J. Electron. Packag.* **143**, (2020).
13. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
14. Kazlauskas, R. J. & Bornscheuer, U. T. Finding better protein engineering strategies. *Nat. Chem. Biol.* **5**, 526–529 (2009).
15. Schirò, A. *et al.* On the complementarity of X-ray and NMR data. *J. Struct. Biol. X* **4**, 100019 (2020).
16. Milne, J. L. S. *et al.* Cryo-electron microscopy – a primer for the non-microscopist. *FEBS J.* **280**, 28–45 (2013).
17. Wells, J. A., Powers, D. B., Bott, R. R., Graycar, T. P. & Estell, D. A. Designing substrate specificity by protein engineering of electrostatic interactions. *Proc. Natl. Acad. Sci.* **84**, 1219 LP – 1223 (1987).
18. Bränd´en, C.-I. & Alwyn Jones, T. Between objectivity and subjectivity. *Nature* **343**, 687–689 (1990).
19. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
20. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem.*

Int. Ed. Engl. **57**, 4143–4148 (2018).
21. Güven, G., Prodanovic, R. & Schwaneberg, U. Protein Engineering – An Option for Enzymatic Biofuel Cell Design. *Electroanalysis* **22**, 765–775 (2010).
22. Lane, M. D. & Seelig, B. Advances in the directed evolution of proteins. *Curr. Opin. Chem. Biol.* **22**, 129–136 (2014).
23. Chen, R. Enzyme engineering: rational redesign versus directed evolution. *Trends Biotechnol.* **19**, 13–14 (2001).
24. Stern, L. A. *et al.* Cellular-Based Selections Aid Yeast-Display Discovery of Genuine Cell-Binding Ligands: Targeting Oncology Vascular Biomarker CD276. *ACS Comb. Sci.* **21**, 207–222 (2019).
25. Barbas, C. F. *et al.* In vitro evolution of a neutralizing human antibody to human immunodeficiency virus type 1 to enhance affinity and broaden strain cross-reactivity. *Proc. Natl. Acad. Sci.* **91**, 3809 LP – 3813 (1994).
26. Kruziki, M. A., Sarma, V. & Hackel, B. J. Constrained Combinatorial Libraries of Gp2 Proteins Enhance Discovery of PD-L1 Binders. *ACS Comb. Sci.* **20**, 423–435 (2018).
27. Lipovšek, D. *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.* **368**, 1024–1041 (2007).
28. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382-8 (2005).
29. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32**, (2004).
30. Alford, R. F. *et al.* An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol.* **11**, (2015).
31. Usmanova, D. R. *et al.* Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics* **34**, 3653–3658 (2018).
32. Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science (80-.).* **324**, 203 LP – 207 (2009).
33. Banta, S., Dooley, K. & Shur, O. Replacing antibodies: engineering new binding proteins. *Annu. Rev. Biomed. Eng.* **15**, 93–113 (2013).
34. Stern, L., Case, B. & Hackel, B. Alternative Non-Antibody Scaffolds for Molecular Imaging of Cancer. *Curr. Opin. Chem. Eng.* **2**, 425–432 (2013).
35. Holliger, P. & Hudson, P. J. Engineered antibody fragments and the rise of single domains. *Nat. Biotechnol.* **23**, 1126–1136 (2005).
36. Hackel, B. J. Alternative Protein Scaffolds for Molecular Imaging and Therapy. in *Engineering in Translational Medicine* 343–364 (Springer London, 2014). doi:10.1007/978-1-4471-4372-7_13
37. Škrlec, K., Štrukelj, B. & Berlec, A. Non-immunoglobulin scaffolds: A focus on their targets. *Trends Biotechnol.* **33**, 408–418 (2015).
38. Gebauer, M. & Skerra, A. Engineered protein scaffolds as next-generation therapeutics. *Annu. Rev. Pharmacol. Toxicol.* **60**, 391–415 (2020).
39. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLOS Comput. Biol.* **4**, e1000002 (2008).
40. Bloom, J. D., Wilke, C. O., Arnold, F. H. & Adami, C. Stability and the evolvability

of function in a model protein. *Biophys. J.* **86**, 2758–2764 (2004).
41. Hackel, B. J., Ackerman, M. E., Howland, S. W. & Wittrup, K. D. Stability and CDR Composition Biases Enrich Binder Functionality Landscapes. *J. Mol. Biol.* **401**, 84–96 (2010).
42. 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* **56**, 1656–1671 (2017).
43. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci.* **103**, 5869 LP – 5874 (2006).
44. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9205 LP – 9210 (2004).
45. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).
46. Vazquez-Lombardi, R. *et al.* Challenges and opportunities for non-antibody scaffold drugs. *Drug Discov. Today* **20**, 1271–1283 (2015).
47. Pérez, A. W. *et al.* In vitro and in silico assessment of the developability of a designed monoclonal antibody library. *MAbs* **11**, 388–400 (2019).
48. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 944–949 (2017).
49. Raybould, M. I. J. J. *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* **116**, 4025 LP – 4030 (2019).
50. Liu, Y. *et al.* High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. *MAbs* **6**, 483–92 (2014).
51. Estep, P. *et al.* An alternative assay to hydrophobic interaction chromatography for high-throughput characterization of monoclonal antibodies. *MAbs* **7**, 553–561 (2015).
52. Ochoa, J.-L. Hydrophobic (interaction) chromatography. *Biochimie* **60**, 1–15 (1978).
53. Queiroz, J. A., Tomaz, C. T. & Cabral, J. M. S. Hydrophobic interaction chromatography of proteins. *J. Biotechnol.* **87**, 143–159 (2001).
54. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
55. Lauer, T. M. *et al.* Developability index: A rapid in silico tool for the screening of antibody aggregation propensity. *J. Pharm. Sci.* **101**, 102–115 (2012).
56. Bailly, M. *et al.* Predicting Antibody Developability Profiles Through Early Stage Discovery Screening. *MAbs* **12**, 1743053 (2020).
57. Kuroda, D. & Tsumoto, K. Engineering Stability, Viscosity, and Immunogenicity of Antibodies by Computational Design. *J. Pharm. Sci.* **109**, 1631–1651 (2020).
58. Lobo, S. A. *et al.* Stability liabilities of biotherapeutic proteins: Early assessment as mitigation strategy. *J. Pharm. Biomed. Anal.* **192**, 113650 (2021).
59. Lv, Z., Ao, C. & Zou, Q. Protein Function Prediction: From Traditional Classifier to Deep Learning. *Proteomics* **19**, 1900119 (2019).
60. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified

rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
61. Xu, Y. *et al.* Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* **60**, 2773–2790 (2020).
62. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
63. Allen-Zhu, Z., Li, Y. & Liang, Y. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. (2020).
64. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–42 (2000).
65. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
66. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *bioRxiv* 2020.01.23.917682 (2020). doi:10.1101/2020.01.23.917682
67. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).
68. Bershtein, S. & Tawfik, D. S. Advances in laboratory evolution of enzymes. *Curr. Opin. Chem. Biol.* **12**, 151–158 (2008).
69. Scott, A. M., Wolchok, J. D. & Old, L. J. Antibody therapy of cancer. *Nat. Rev. Cancer* **12**, 278–287 (2012).
70. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).
71. 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.* **22**, 946–56 (2015).
72. Frejd, F. Y. & Kim, K.-T. Affibody molecules as engineered protein drugs. *Exp. Mol. Med.* **49**, e306–e306 (2017).
73. Binz, H. K. *et al.* Design and characterization of MP0250, a tri-specific anti-HGF/anti-VEGF DARPin® drug candidate. *MAbs* **9**, 1262–1269 (2017).
74. Schiff, D. *et al.* 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* **33**, 247–253 (2015).
75. Souied, E. H. *et al.* 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.* **158**, 724–732 (2014).
76. Rothe, C. & Skerra, A. Anticalin® Proteins as Therapeutic Agents in Human Diseases. *BioDrugs* **32**, 233–243 (2018).
77. Stern, L. A., Case, B. A. & Hackel, B. J. Alternative Non-Antibody Protein Scaffolds for Molecular Imaging of Cancer. *Curr. Opin. Chem. Eng.* **2**, 425–432 (2013).
78. Kobe, B. & Deisenhofer, J. The leucine-rich repeat: a versatile binding motif. *Trends Biochem. Sci.* **19**, 415–421 (1994).
79. Koide, A., Bailey, C. W., Huang, X. & Koide, S. The fibronectin type III domain as a scaffold for novel binding proteins. *J. Mol. Biol.* **284**, 1141–1151 (1998).
80. Binz, H. K. *et al.* High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.* **22**, 575–582 (2004).

81. 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.* **96**, 1898–903 (1999).
82. Nord, K. *et al.* Binding proteins selected from combinatorial libraries of an α-helical bacterial receptor domain. *Nat. Biotechnol.* **15**, 772–777 (1997).
83. Grabulovski, D., Kaspar, M. & Neri, D. A Novel, Non-immunogenic Fyn SH3-derived Binding Protein with Tumor Vascular Targeting Properties. *J. Biol. Chem.* **282**, 3196–3204 (2007).
84. Kolmar, H. Alternative binding proteins: Biological activity and therapeutic potential of cystine-knot miniproteins. *FEBS J.* **275**, 2684–2690 (2008).
85. Correa, A. *et al.* Potent and Specific Inhibition of Glycosidases by Small Artificial Binding Proteins (Affitins). *PLoS One* **9**, e97438 (2014).
86. 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.* **50**, 417–25 (2009).
87. Chen, J., Sawyer, N. & Regan, L. Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* **22**, 510–515 (2013).
88. Engh, R. A. & Bossemeyer, D. Structural aspects of protein kinase control—role of conformational flexibility. *Pharmacol. Ther.* **93**, 99–111 (2002).
89. Novotný, J. *et al.* Molecular anatomy of the antibody binding site. *J. Biol. Chem.* **258**, 14433–7 (1983).
90. Nord, K., Nilsson, J., Nilsson, B., Uhlén, M. & Nygren, P. A. A combinatorial library of an alpha-helical bacterial receptor domain. *Protein Eng.* **8**, 601–8 (1995).
91. 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.* **415**, 393–405 (2012).
92. Searle, M. S. & Williams, D. H. The cost of conformational order: entropy changes in molecular associations. *J. Am. Chem. Soc.* **114**, 10690–10697 (1992).
93. Cole, C. & Warwicker, J. Side-chain conformational entropy at protein-protein interfaces. *Protein Sci.* **11**, 2860–70 (2002).
94. Yu, H., Yan, Y., Zhang, C. & Dalby, P. A. Two strategies to engineer flexible loops for improved enzyme thermostability. *Sci. Rep.* **7**, 41212 (2017).
95. Nagarajan, R. *et al.* PDBparam: Online Resource for Computing Structural Parameters of Proteins. *Bioinform. Biol. Insights* **10**, BBI.S38423 (2016).
96. Eyal, E. & Bahar, I. Toward a Molecular Understanding of the Anisotropic Response of Proteins to External Forces: Insights from Elastic Network Models. *Biophys. J.* **94**, 3424–3435 (2008).
97. Schrödinger, LLC. *The {PyMOL} Molecular Graphics System, Version~1.8*. (2015).
98. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
99. Kowalsky, C. A. *et al.* Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. *J. Biol. Chem.* **290**, 26457–70 (2015).
100. Woldring, D. R., Holec, P. V. & Hackel, B. J. ScaffoldSeq: Software for

characterization of directed evolution populations. *Proteins Struct. Funct. Bioinforma.* **84**, 869–874 (2016).
101. 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* **10**, e0138956 (2015).
102. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
103. Birtalan, S., Fisher, R. D. & Sidhu, S. S. The functional capacity of the natural amino acids for molecular recognition. *Mol. Biosyst.* **6**, 1186 (2010).
104. Birtalan, S. *et al.* The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies. *J. Mol. Biol.* **377**, 1518–1528 (2008).
105. Koide, S. & Sidhu, S. S. The Importance of Being Tyrosine: Lessons in Molecular Recognition from Minimalist Synthetic Binding Proteins. *ACS Chem. Biol.* **4**, 325–334 (2009).
106. Eijsink, V. G. H. *et al.* Rational engineering of enzyme stability. *J. Biotechnol.* **113**, 105–120 (2004).
107. Ma, J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* **13**, 373–80 (2005).
108. Skjaerven, L., Hollup, S. M. & Reuter, N. Normal mode analysis for proteins. *J. Mol. Struct. THEOCHEM* **898**, 42–48 (2009).
109. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–56 (1987).
110. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1**, 755–768 (2006).
111. Hood, M. T. & Stachow, C. Influence of Polyethylene Glycol on the Size of Schizosaccharomyces pombe Electropores. *Appl. Environ. Microbiol.* **58**, 1201–6 (1992).
112. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).
113. Le, Q. V, Karpenko, A., Ngiam, J. & Ng, A. Y. *ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning*.
114. Jolliffe, I. Principal component analysis. in *International encyclopedia of statistical science* 1094–1096 (Springer, 2011).
115. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **67**, 301–320 (2005).
116. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
117. Xu, Y. *et al.* Structure, heterogeneity and developability assessment of therapeutic antibodies. *MAbs* **11**, 239–264 (2019).
118. Yang, X. *et al.* Developability studies before initiation of process development. *MAbs* **5**, 787–794 (2013).
119. Almagro, J. C., Pedraza-Escalona, M., Arrieta, H. I. & Pérez-Tapia, S. M. Phage Display Libraries for Antibody Therapeutic Discovery and Development. *Antibodies (Basel, Switzerland)* **8**, 44 (2019).

120. Delmar, J. A., Wang, J., Choi, S. W., Martins, J. A. & Mikhail, J. P. Machine Learning Enables Accurate Prediction of Asparagine Deamidation Probability and Rate. *Mol. Ther. - Methods Clin. Dev.* **15**, 264–274 (2019).
121. Lu, X. *et al.* Deamidation and isomerization liability analysis of 131 clinical-stage antibodies. *MAbs* **11**, 45–57 (2019).
122. Buck, P. M. *et al.* Computational methods to predict therapeutic protein aggregation. *Methods Mol. Biol.* **899**, 425–451 (2012).
123. Magnan, C. N., Randall, A. & Baldi, P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* **25**, 2200–2207 (2009).
124. Agrawal, N. J. *et al.* Computational tool for the early screening of monoclonal antibodies for their viscosities. *MAbs* **8**, 43–48 (2016).
125. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci.* **106**, 11937 LP – 11942 (2009).
126. 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.* **22**, 553–560 (2009).
127. Nisthal, A., Wang, C. Y., Ary, M. L. & Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci.* **116**, 16367 LP – 16377 (2019).
128. Chai, Q. *et al.* Development of a high-throughput solubility screening assay for use in antibody discovery. *MAbs* **11**, 747–756 (2019).
129. Carter, P. J. & Lazar, G. A. Next generation antibody drugs : pursuit of the ' high-hanging fruit '. *Nat. Rev. Drug Discov.* **17**, 197–223 (2018).
130. Miersch, S. & Sidhu, S. S. Synthetic antibodies: Concepts, potential and practical considerations. *Methods* **57**, 486–498 (2012).
131. Chan, J. Y., Hackel, B. J. & Yee, D. Targeting Insulin Receptor in Breast Cancer Using Small Engineered Protein Scaffolds. *Mol. Cancer Ther.* **16**, 1324 LP – 1334 (2017).
132. Case, B. A., Kruziki, M. A., Johnson, S. M. & Hackel, B. J. Engineered Charge Redistribution of Gp2 Proteins through Guided Diversity for Improved PET Imaging of Epidermal Growth Factor Receptor. *Bioconjug. Chem.* **29**, 1646–1658 (2018).
133. Du, F. *et al.* Engineering an EGFR-binding Gp2 domain for increased hydrophilicity. *Biotechnol. Bioeng.* **116**, 526–535 (2019).
134. Golinski, A. W., Holec, P. V, Mischler, K. M. & Hackel, B. J. Biophysical Characterization Platform Informs Protein Scaffold Evolvability. *ACS Comb. Sci.* **21**, 323–335 (2019).
135. Ritter, S. C. & Hackel, B. J. Validation and stabilization of a prophage lysin of Clostridium perfringens by using yeast surface display and coevolutionary models. *Appl. Environ. Microbiol.* **85**, (2019).
136. Klesmith, J. R. *et al.* Retargeting CD19 Chimeric Antigen Receptor T Cells via Engineered CD19-Fusion Proteins. *Mol. Pharm.* **16**, 3544–3558 (2019).
137. Cabantous, S. & Waldo, G. S. In vivo and in vitro protein solubility assays using split GFP. *Nat. Methods* **3**, 845–854 (2006).
138. Foit, L. *et al.* Optimizing Protein Stability In Vivo. *Mol. Cell* **36**, 861–871 (2009).

139. Ebo, J. S. *et al.* An in vivo platform to select and evolve aggregation-resistant proteins. *Nat. Commun.* **11**, 1–12 (2020).
140. Overton, T. W. Recombinant protein production in bacterial hosts. *Drug Discov. Today* **19**, 590–601 (2014).
141. Tian, G. *et al.* Quantitative dot blot analysis (QDB), a versatile high throughput immunoblot method. *Oncotarget* **8**, 58553–58562 (2017).
142. Chen, J. *et al.* Chaperone activity of DsbC. *J. Biol. Chem.* **274**, 19601–19605 (1999).
143. Galarneau, A., Primeau, M., Trudeau, L.-E. & Michnick, S. W. β-Lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein–protein interactions. *Nat. Biotechnol.* **20**, 619–622 (2002).
144. Mansell, T. J., Linderman, S. W., Fisher, A. C. & DeLisa, M. P. A rapid protein folding assay for the bacterial periplasm. *Protein Sci.* **19**, 1079–1090 (2010).
145. Yeo, I.-K. & Johnson, R. A. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* **87**, 954–959 (2000).
146. Jany, K.-D., Lederer, G. & Mayer, B. Amino acid sequence of proteinase K from the mold Tritirachium album Limber: Proteinase K—a subtilisin-related enzyme with disulfide bonds. *FEBS Lett.* **199**, 139–144 (1986).
147. Hall, M. A. Correlation-based feature selection for machine learning. (1999).
148. Zutz, A. *et al.* A dual-reporter system for investigating and optimizing protein translation and folding in E. coli. *bioRxiv* 2020.09.18.303453 (2020). doi:10.1101/2020.09.18.303453
149. Lesley, S. A., Graziano, J., Cho, C. Y., Knuth, M. W. & Klock, H. E. Gene expression response to misfolded protein as a screen for soluble recombinant protein. *Protein Eng. Des. Sel.* **15**, 153–160 (2002).
150. Hoffmann, F. & Rinas, U. Stress Induced by Recombinant Protein Production in Escherichia coli BT - Physiological Stress Responses in Bioprocesses: -/-. in 73–92 (Springer Berlin Heidelberg, 2004). doi:10.1007/b93994
151. Fellouse, F. A. *et al.* High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J. Mol. Biol.* **373**, 924–940 (2007).
152. Seeger, M. A. *et al.* Design, construction, and characterization of a second-generation DARPin library with reduced hydrophobicity. *Protein Sci.* **22**, 1239–1257 (2013).
153. Bergstra, J., Yamins, D. & Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* I–115–I–123 (JMLR.org, 2013).
154. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. & Cox, D. D. Hyperopt: a python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **8**, 14008 (2015).
155. Lebigot, E. O. Uncertainties: a Python package for calculations with uncertainties. *URL http//pythonhosted. org/uncertainties* (2010).
156. Scanlon, T. C., Gray, E. C. & Griswold, K. E. Quantifying and resolving multiple vector transformants in S. cerevisiae plasmid libraries. *BMC Biotechnol.* **9**, 95 (2009).
157. Kamiyama, D. *et al.* Versatile protein tagging in cells with split fluorescent protein. *Nat. Commun.* **7**, 11046 (2016).

158. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 81257 (2016). doi:10.1101/081257
159. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
160. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
161. Borrebaeck, C. A. K. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nat. Rev. Cancer* **17**, 199–204 (2017).
162. Arbige, M. V, Shetty, J. K. & Chotani, G. K. Industrial Enzymology: The Next Chapter. *Trends Biotechnol.* **37**, 1355–1366 (2019).
163. Engqvist, M. K. M. & Rabe, K. S. Applications of Protein Engineering and Directed Evolution in Plant Research. *Plant Physiol.* **179**, 907–917 (2019).
164. Kapoor, S., Rafiq, A. & Sharma, S. Protein engineering and its applications in food industry. *Crit. Rev. Food Sci. Nutr.* **57**, 2321–2329 (2017).
165. Narayanan, H. *et al.* Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. *Trends Pharmacol. Sci.* **42**, 151–165 (2021).
166. Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **28**, 374 (2000).
167. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. *Adv. Neural Inf. Process. Syst.* **32**, 9689–9701 (2019).
168. Kruziki, M. A. *et al.* 64Cu-Labeled Gp2 Domain for PET Imaging of Epidermal Growth Factor Receptor. *Mol. Pharm.* **13**, 3747–3755 (2016).
169. Golinski, A. W. *et al.* High-Throughput Developability Assays Enable Library-Scale Identification of Producible Protein Scaffold Variants. *bioRxiv* 2020.12.14.422755 (2020). doi:10.1101/2020.12.14.422755
170. Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins* **58**, 22–30 (2005).
171. Chou, P. Y. & Fasman, G. D. Secondary structural prediction of proteins from their amino acid sequence. *Trends Biochem. Sci.* **2**, 128–131 (1977).
172. Finkelstein, A. V, Badretdinov, A. Y. & Ptitsyn, O. B. Physical reasons for secondary structure stability: α-Helices in short peptides. *Proteins Struct. Funct. Bioinforma.* **10**, 287–299 (1991).
173. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2020).
174. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, (2017).
175. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
176. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
177. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
178. Ma, E. J. & Kummer, A. Reimplementing Unirep in JAX. *bioRxiv* (2020). doi:10.1101/2020.05.11.088344
179. Skilling, J. Nested sampling for general Bayesian computation. *Bayesian Anal.* **1**,

833–859 (2006).
180. Martiniani, S., Stevenson, J. D., Wales, D. J. & Frenkel, D. Superposition Enhanced Nested Sampling. *Phys. Rev. X* **4**, 31034 (2014).
181. Pártay, L. B., Bartók, A. P. & Csányi, G. Efficient Sampling of Atomic Configurational Spaces. *J. Phys. Chem. B* **114**, 10502–10512 (2010).
182. Burkoff, N. S., Várnai, C., Wells, S. A. & Wild, D. L. Exploring the energy landscapes of protein folding simulations with Bayesian computation. *Biophys. J.* **102**, 878–886 (2012).
183. Pardalos, P. M. & Mavridou, T. D. Simulated annealingSimulated Annealing. in *Encyclopedia of Optimization* (eds. Floudas, C. A. & Pardalos, P. M.) 3591–3593 (Springer US, 2009). doi:10.1007/978-0-387-74759-0_617