High-throughput developability assays enable library-scale identification of producible protein scaffold variants

Alexander W. Golinski[a], Katelynn M. Mischler[a], Sidharth Laxminarayan[a], Nicole L. Neurock[a], Matthew Fossing[a], Hannah Pichman[a], Stefano Martiniani[a], and Benjamin J. Hackel[a,1]

[a]Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455

Edited by David Baker, University of Washington, Seattle, WA, and approved April 7, 2021 (received for review December 28, 2020)

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 10 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 site-wise 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.

developability | protein engineering | predictive modeling

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 (1–5). Despite this universal importance, developability studies are performed late in the commercialization pipeline (2, 4) and limited by traditional experimental capacity (6). This is problematic because 1) proteins with poor developability limit practical assay capacity for measuring primary function, 2) optimal developability is often not observed with proteins originally found in alternative formats [such as display or two-hybrid technologies (7)], and 3) engineering efforts are limited by the large gap between observation size ($\sim 10^2$) and theoretical mutational diversity ($\sim 10^{20}$). 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 posttranslational modifications (8, 9), solubility (10, 11), viscosity (12), and overall developability (13). Structural approaches have informed stability (14) and solubility (10, 15). However, many in silico models require an experimentally solved structure or suffer from computational structure prediction inaccuracies (16). Additionally, limited developability information allows for limited predictive model accuracy (17). In vitro methods have identified several experimental protocols to mimic practical developability requirements [e.g., affinity-capture self-interaction nanoparticle spectroscopy (18) and chemical precipitation (19) 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 (1, 5).

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 (20–24). While significant progress has been achieved with regards to identifying paratopes for optimal binding strength and specificity (25, 26), isolating highly developable variants remains plagued. One particular protein scaffold, Gp2, has been evolved into specific binding variants toward multiple targets (27–29). Continued study improved charge distribution (30), hydrophobicity (31), and stability (28). While these studies have suggested improvements for future framework and paratope residues (including a disulfide-stabilized loop), a poor developability distribution is still observed (32) (Fig. 1 *A* and *B*). Assuming the randomized paratope library will lack similar primary functionality, the Gp2 library will simulate the universal

Significance

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 green fluorescent protein 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. 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.

Author contributions: A.W.G., K.M.M., and B.J.H. designed research; A.W.G., K.M.M., S.L., N.L.N., M.F., and H.P. performed research; A.W.G., S.M., and B.J.H. analyzed data; and A.W.G., S.M., and B.J.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: hackel@umn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2026658118/-/DCSupplemental.

Published June 2, 2021.

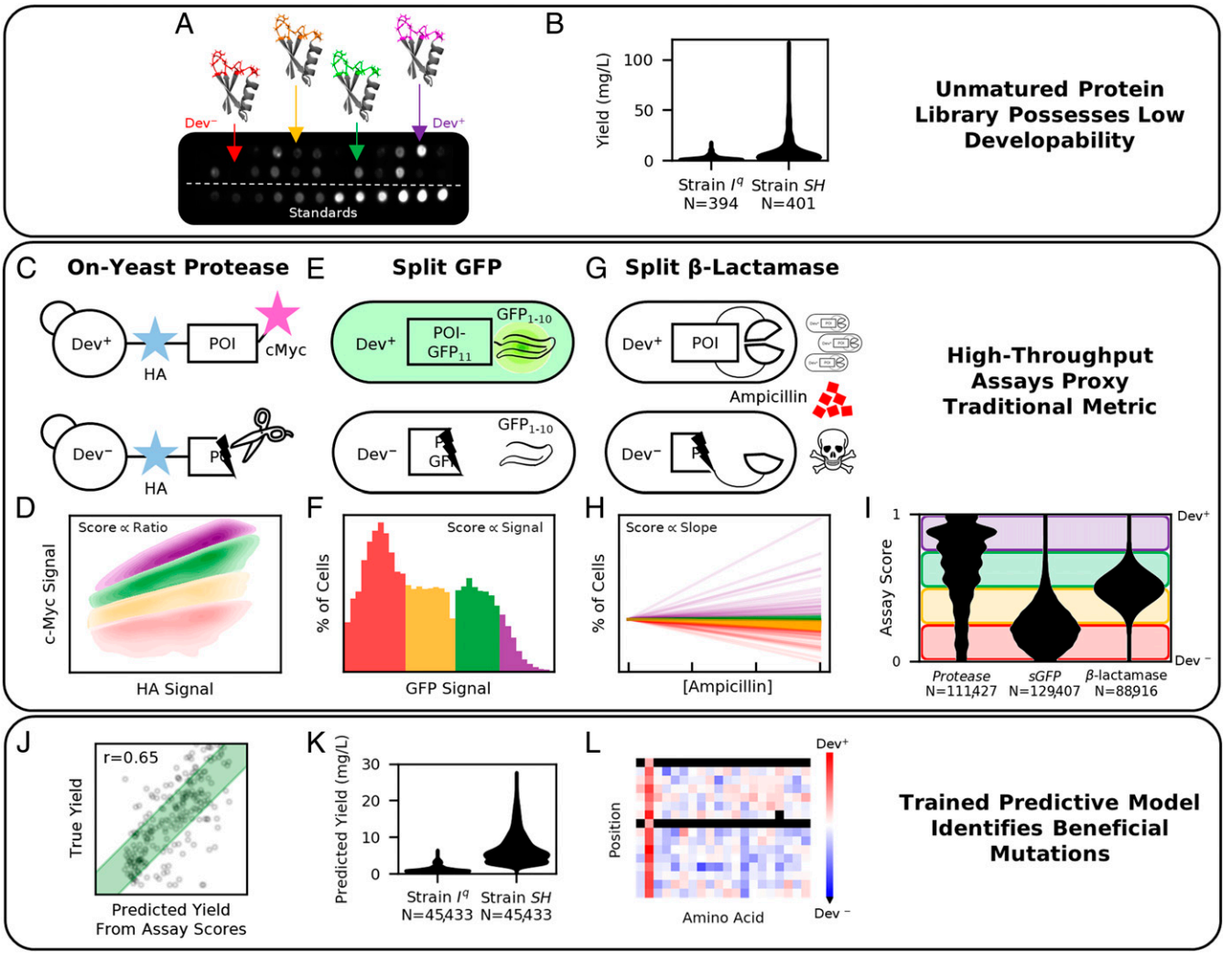

Fig. 1. HT assays were evaluated for the ability to identify protein scaffold variants with increased developability. (*A* and *B*) Gp2 variant expression, commonly measured via low-throughput techniques such as the dot blot shown, highlights the rarity of ideal developability. (*C* and *D*) The HT on-yeast protease assay measures the stability of the POI by proteolytic extent. (*E* and *F*) The HT split-GFP assay measures POI expression via recombination of a genetically fused GFP fragment. (*G* and *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* and *J*) Assay scores, assigned to each unique sequence via deep sequencing, were evaluated by predicting expression (Fig. 3). (*K* and *L*) HT assay capacity enables large-scale developability evaluation and can be used to identify beneficial mutations (Fig. 4).

applicability of the proposed high-throughput (HT) developability assays.

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 (Fig. 1 *C* and *D*)—previously validated to improve the stability of de novo proteins (33), antimicrobial lysins (34), and immune proteins (35)—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, and protease specificity) can be resolved; 2) Split green fluorescent protein (GFP, Fig. 1 *E* and *F*)—previously used to determine soluble protein concentrations (36)—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; and 3) Split β-lactamase (Fig. 1 *G* and *H*)—previously used to improve thermodynamic stability (37) and solubility (38)—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 (Fig. 1*I*). We then converted the assay scores into a traditional developability metric by building a model that predicts recombinant yield (Fig. 1*J*). The assays' capacity enabled yield evaluations for >100-fold traditional assay capacity (Fig. 1*K*, compared to Fig. 1*B*) and provide an introductory analysis of factors driving protein developability by observing beneficial mutations via predicted developable proteins (Fig. 1*L*).

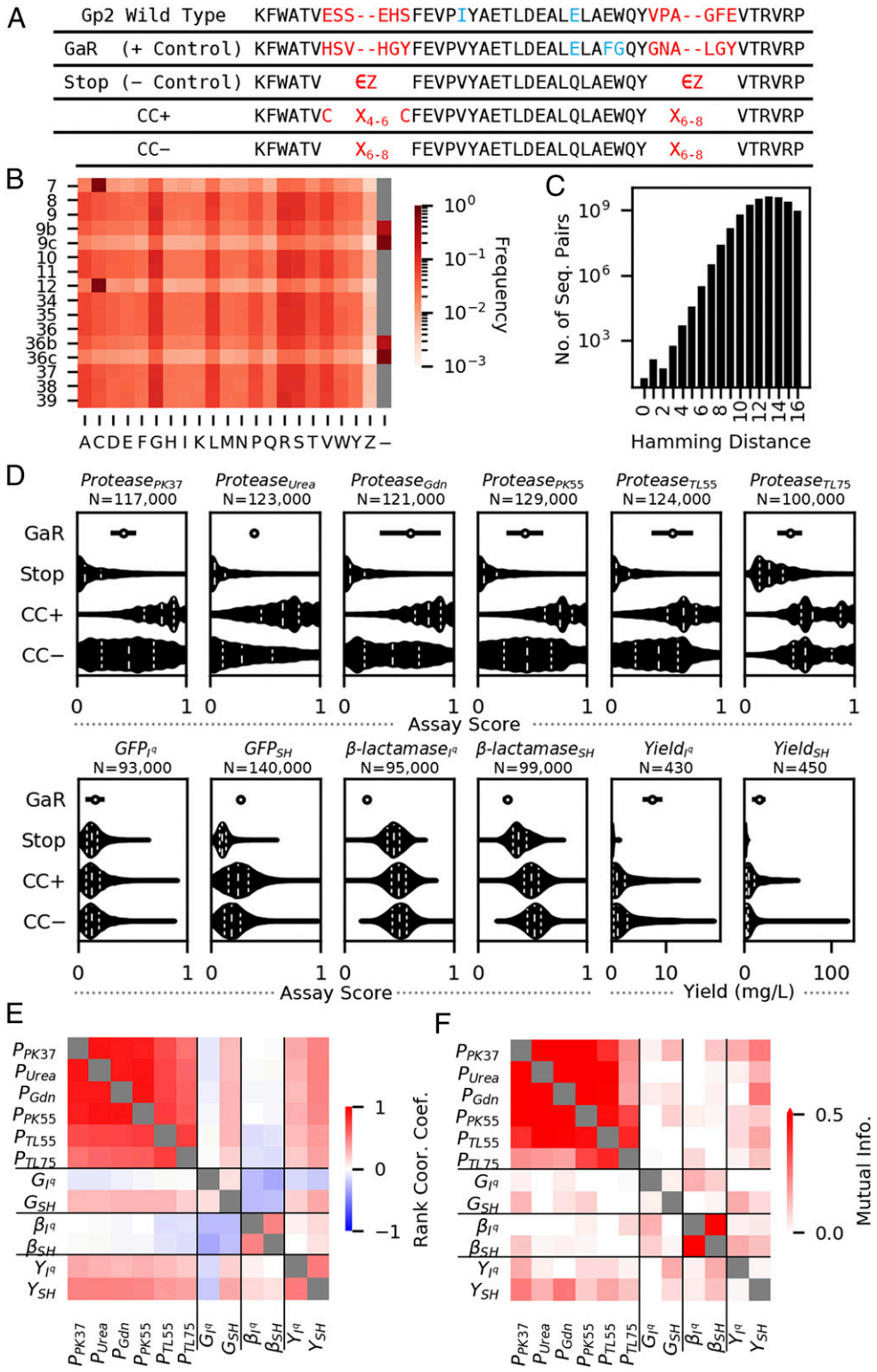

Fig. 2. Developability characterization of loop-diversified Gp2 library. (*A*) Sequence alignment of assayed sequence classes: *GaR* (single-variant control), *Stop*: (sequences with stop codon, Z), *CC+*: (hypothesized to be more developable), *CC−*: (hypothesized to be less developable). (*B*) Diversified paratope frequency heatmap. (*C*) Histogram depicting the pairwise distances between 190,483 full-length and genotypically unique variants. (*D*) Assay performance distributions divided by class. (*Top*) Various on-yeast protease assay reaction conditions. (*Bottom*) Bacterial assays performed in strain I^q and strain SH. GaR error bars represent the SD (n = 3 trials). Total unique variants for Stop, CC+, and CC− range 93,178 to 140,229 for HT assays and 431 to 447 for yield (reference *SI Appendix*, Fig. S2). (*E*) The Spearman's rank correlation coefficient and (*F*) MI between HT assays and yield.

Results

Gp2 Paratope Library Quantification. We first evaluated the assays' ability to separate sequence classes with a hypothesized difference in developability. A total of 204,174 observed Gp2 variants belonged to one of four classes (Fig. 2*A*): *GaR*: a thermostable variant (27), *Stop*: 13,690 nonfunctional truncated variants; *CC+*: 128,854 variants with a hypothesized (28) 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 (28) and two paratope loops, each with 6 to 8 "NNK" degenerate codons encoding all 20 amino acids (Fig. 2*B*). The library was widely diversified, averaging 13.1 differences between observed sequence pairs (Fig. 2*C*).

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 (39). 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 (40) via a C-terminal His_6 tag (Fig. 1 *A* and *B*).

Different bacterial strains have been evolved containing additional machinery to obtain increased yield. We chose to include two *Escherichia 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 (41). 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 (Fig. 2*D*): *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* SD, $n = 3$ 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.

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*, Fig. S1). Below we motivate score calculation, followed by assay score distribution analysis (Fig. 2*D* and *SI Appendix*, Fig. S2).

On-yeast stability. The on-yeast stability assay evaluates protein stability by measuring proteolytic cleavage (Fig. 1*C*). Using yeast surface display technology (42), the POI is expressed between two tags (N-terminal HA and 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 (Fig. 1*D*). Sequencing scored every collected variant on a cell-weighted average: 1 (intact), 2/3, 1/3, and 0 (fully cleaved).

We performed the proteolysis using various conditions to determine if additional stability metrics could be obtained (*SI Appendix*, Fig. 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} versus P_{PK55} and P_{TL55} versus P_{TL75}).

Assay scores were calculated for $>10^5$ unique Gp2 variants in each of the six reaction conditions. The assay score distributions per class (Fig. 2*D*) matched hypothesized developability in all conditions except P_{TL75}. SDs were small (0.17 to 0.20, except P_{TL75}: 0.29). *Stop* variants scored low (0.04 to 0.08, except P_{TL75}: 0.23). *GaR* scored higher than most *Stop* variants (67 to 81%, except P_{TL75}: 35%). One potential hypothesis for P_{TL75} is the increased temperature may lead to nonspecific binding of surface-aggregated proteins. Nevertheless, all reaction conditions, displayed a significantly higher distribution of assay scores for *CC+* versus *CC−* (one-way Mann–Whitney U test, $P < 0.001$), validating each condition's utility.

Split GFP. The split GFP assay measures POI concentration with a C terminus–fused 11th strand of GFP (Fig. 1*E*). Upon recombination with GFP strands 1 to 10, which was separately induced following POI production and a 1-h gap, the POI fusion remaining soluble in the cytosol will produce a fluorescent signal detectable by FACS (Fig. 1*F*). 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, and 0 (background signal).

The assay score distributions (Fig. 2*D*) 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.

Split β-lactamase. In the split β-lactamase assay, the POI is inserted in a loop distal to the active site [final construct: β-$lac_{1\text{-}194}$-$(G_4S)_2$-AS-POI-GS-$(G_4S)_2$-β-$lac_{197\text{-}287}$, location previously observed to retain 40% activity (43)]. Functional enzyme, hypothesized to be paired with POI solubility and folding robustness (44), provides ampicillin resistance allowing cell reproduction (Fig. 1*G*). The change in growth rates was measured as the change in POI amplicon abundance in cultures grown to saturation with varying antibiotic concentrations (Fig. 1*H*). For comparison to other assays and improved modeling efficiency, slopes were normalized and scaled (*Materials and Methods*).

The split β-lactamase assay produced assay scores that were contradictory toward hypothesized developability yet were able to differentiate classes, suggesting potential utility despite an unsolved mechanism. We obtained assay scores for 10^5 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 nonzero ampicillin concentrations) performed as expected (*SI Appendix*, Fig. 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 Fig. 5 and *Discussion* for further explanation.

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 (Fig. 2*E*), the assays have nonzero mutual information (MI) with yield. This suggests utility as long as the predictive model is capable of exploiting the nonlinear relationships captured by MI

(Fig. 2*F*). 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*, Fig. S4); thus, a multitask model (Fig. 3*A*) 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.

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 (45) power transform and normalization was applied to remove correlation between error and yield ($\lambda = -0.324$, *SI Appendix*, Fig. 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 1,023 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 (Fig. 3*B*), suggesting all conditions possess utility. There were seven assay combinations (using 7 of the 10 assays) that performed optimally and equally (*SI Appendix*, Fig. 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}, and β_{SH} (Fig. 3*C*, 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 (Fig. 3*D* and *SI Appendix*, Fig. S6). The assay model (MSE: 0.565) was able to significantly (one-way Student's *t* test, $P < 0.05$) outperform the OH sequence model (MSE: 0.667) and

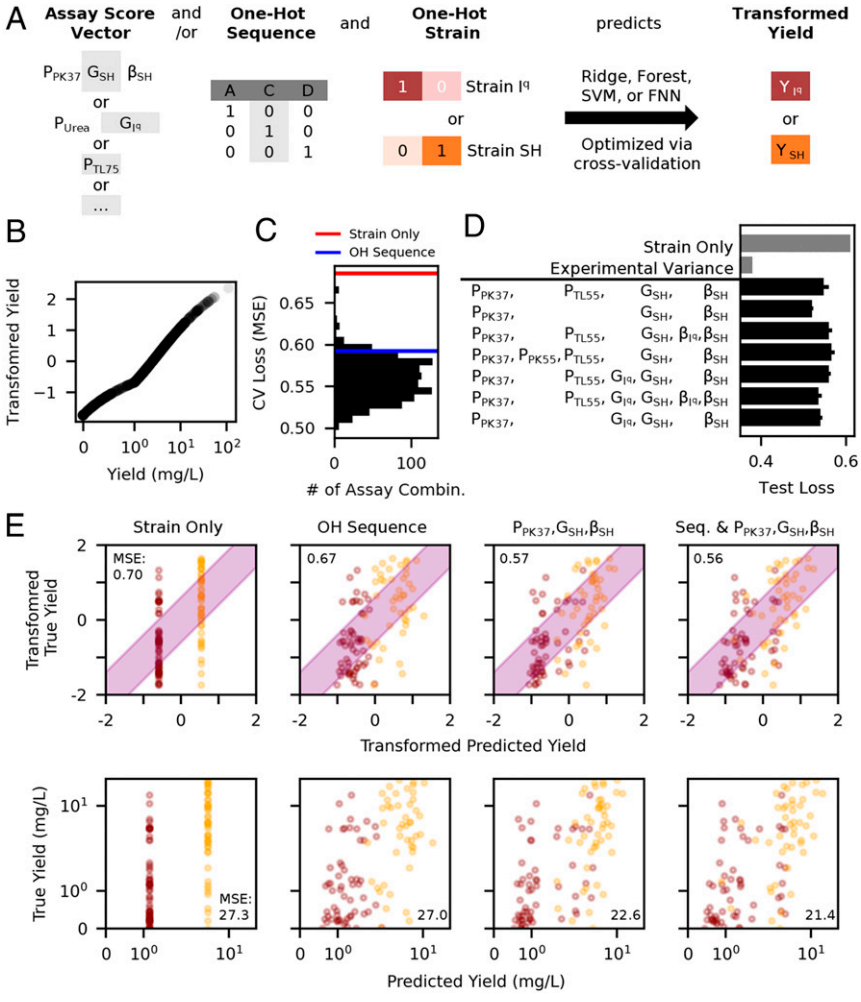

Fig. 3. Determination of predictive HT developability assays. (*A*) Model visualization utilizing HT assay scores, an OH paratope sequence, and an OH strain identifier to predict the recombinant yield in both cell types. (*B*) Power transformation and standardization of yields to remove correlation between yield and error (*SI Appendix*, Fig. S5). (*C*) Model loss distribution for 1,023 HT assay combinations of MSE between predicted and actual yields. (*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:* nontransformed) during final evaluation on set of 97 sequences. The purple-shaded area represents true yield ± square root of sequence-averaged experimental variance.

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 50th percentile of each strain (*SI Appendix*, Fig. 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.

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}, and β_{SH} assay scores and OH sequence was used to predict the yield for 45,433 unique sequences in both strains (Figs. 1*K* and 4*A*). 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.

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 (Fig. 4*B*). 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*, Fig. S8). Regarding epistasis, we analyzed the probability of Dev^+ as conditioned by number of cysteines in the sequence, finding three or four cysteines most optimal (Fig. 4*C*). There also appears to be a benefit of seven 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 (Fig. 4*D*). It should be noted that 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*, Fig. 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 (46)]. However, both the OH model and a model utilizing only assays G_{SH} and β_{SH} also

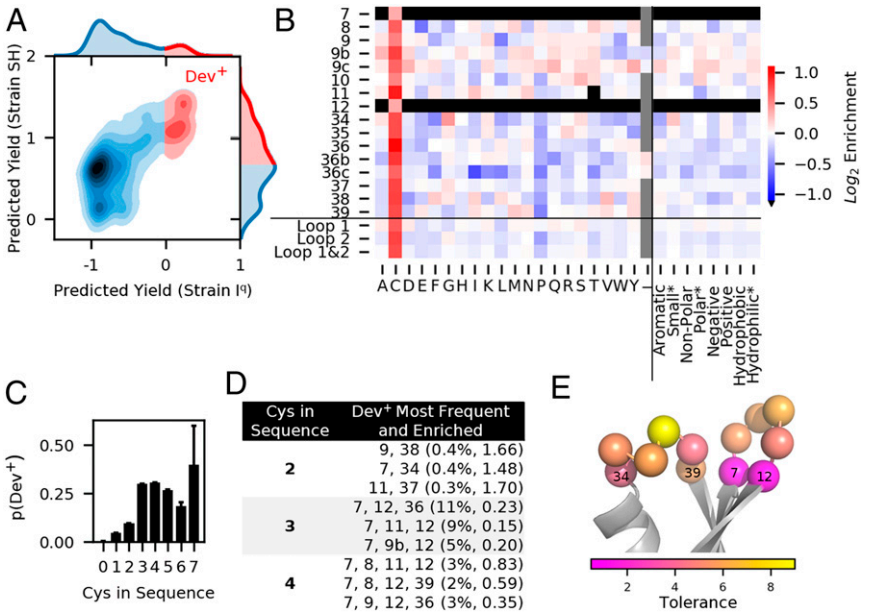

Fig. 4. HT assays enable prediction of Gp2 variants with high developability. (*A*) Kernel density plot of the predicted yield of 45,433 unique sequences in each bacterial strain. A total of 6,394 sequences with high predicted yield in both strains were isolated as Dev^+ (red). (*B*) Site-wise enrichment 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), nonpolar 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 to 11. Loop 2: positions 34 to 39. (*C*) The proportion of sequences predicted identified as Dev^+ as a function of the number of cysteines in the sequence. Error bars: 1 divided by 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 (Protein Data Bank: 2WMN) colored by the mutational tolerance calculated as the inverse of the average magnitude of amino acid enrichment.

indicate a stabilizing effect of additional cysteines (*SI Appendix*, Fig. S10 *C–F*). Moreover, recombinant yield increased at higher cysteine frequencies of synthesized variants (ρ: I^q = 0.28, SH = 0.48, *SI Appendix*, Fig. S12 *A* and *B*).

Additional analyses 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 and 36c (enriching sequences of wild-type length) suggest that 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 (Fig. 4*E*). 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 that either 1) the second loop is a poor paratope in terms of allowing broad diversity with favorable developability or 2) the stabilizing disulfide bond offsets unfavorable mutations within the first loop.

β_{SH} Assay Predictive Performance Explained by MI. 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 (Fig. 5*A*). Matching the sequence class distributions (Fig. 2), P_{PK37} and G_{SH} assay scores of Dev^+ sequences were significantly higher, and β_{SH} assay scores were significantly lower than the initial distribution (Fig. 5*A*, 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.

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 (Fig. 5*B*). While the P_{PK37} and G_{SH} assays, alone and together, performed better with a linear model, four of five models using the β_{SH} assay performed best with a nonlinear model.

The correlation-based feature selection (47) (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 nonredundant information with respect to other HT assays. The CFS calculated by MI was significantly higher, and CV loss was significantly lower for HT assay combinations containing β_{SH} than assay combinations without (Fig. 5*C*, 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*, Fig. S11). As a result, the top CFS combination via rank correlation (P_{PK37}, P_{Urea}, P_{PK55}, G_{Iq}, and 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}, and 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 (48, 49), should be considered if it is hypothosized that the assays will provide nonredundant metrics of developability.

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 (Fig. 6*A*). With only 10 sequences (5% of data), the P_{PK37}, G_{SH}, and β_{SH} model achieves this goal (one-way paired Student's *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 (Fig. 6*B*), the models using assays required only 59 (P_{PK37}, G_{SH}, and β_{SH}, 30% of data, $P < 0.05$) or 78 (Sequence and P_{PK37}, G_{SH}, and 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 to 70% over sequence information alone.

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 (Fig. 6*C*). 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}, and β_{SH}) and 81 ± 24% (sequence and P_{PK37}, G_{SH}, and 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.

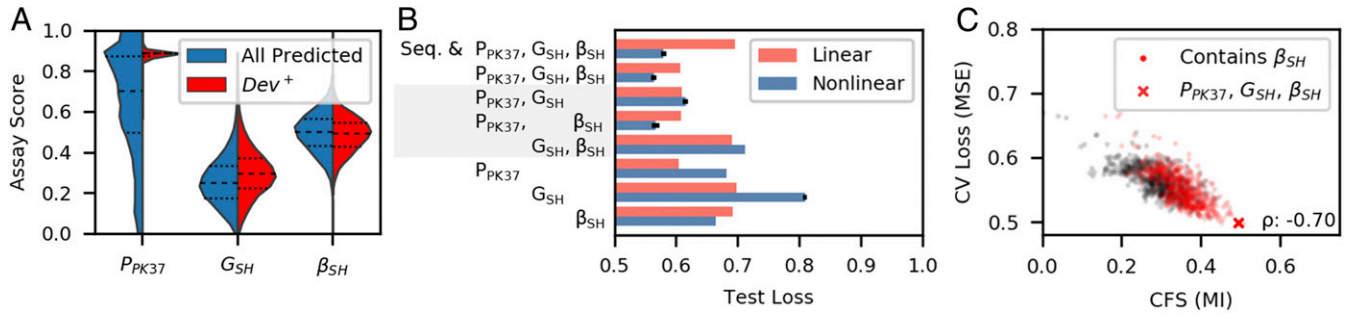

Fig. 5. Nonlinear models can extract nonlinear developability MI from the split β-lactamase assay. (*A*) Comparison of assay score distributions between 45,433 unique sequences with observed P_{PK37}, G_{SH}, and β_{SH} assay scores (blue) versus 6,394 of the sequences with high predicted developability (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). The error bars in nonlinear models represent SD in MSE from $n = 10$ stochastically trained models. (*C*) The CFS as calculated by MI for 1,023 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.

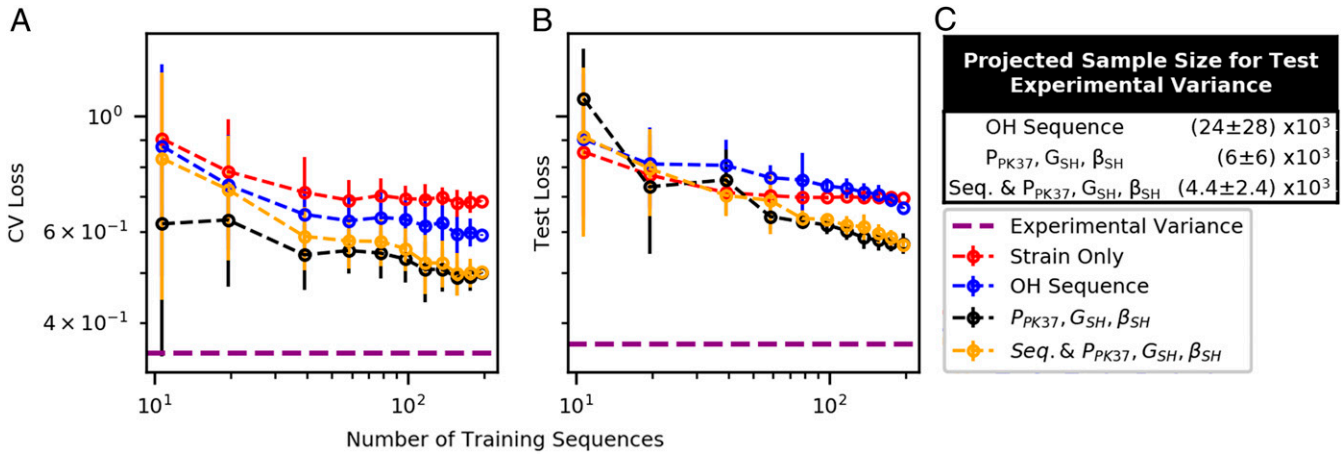

Fig. 6. HT developability assays reduce training size requirement. A total of 10 bootstrapped samples for each sample size (ranging 5 to 100% of available data) were individually trained by CV and evaluated on 97 independent test sequences. The error bars represent SD across models. (*A*) The performance during CV describing the model's ability to predict developability. (*B*) The predictive performance against the independent sequence set describing the model's 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 SEs of the parameter estimates.

Error Analysis. While the trio of assays provide valuable developability assessment, we sought to identify factors that limit performance. Due to the sampling strategy (*Materials and 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*, Fig. S12*A*). However, we observed nonsignificant correlation (ρ: $I^q = 0.02$, SH = 0.07, *SI Appendix*, Fig. S12*B*) 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*, Fig. 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*, Fig. 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: $P_{PK37} = 0.66$ to 0.71, $G_{SH} = 0.26$ to 0.29, and $\beta_{SH} = 0.39$ to 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 HT error identified in this section paired with moderate MI between assays and yield (*SI Appendix*, Fig. S11*D*) 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 nonredundant 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.

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 40,000 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 assay modifications were unable to capture alternative stability metrics or that a single condition is sufficient to predict developability. Additional conditions may be useful for predicting other traditional developability metrics, such as thermostability. For example, P_{TL55} was found in five of seven 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 (37). 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 (50), lowering the frequency of highly produced variants. Library plating on agar plates could reduce this mechanism but may introduce throughput limitations to achieve sufficient physical spacing to avoid bystander ampicillin reduction. 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 ($\sim 10^{20}$) 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 OH-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 site-wise amino acid biases based upon predicted yield of 40,000 unique paratopes, which can be used to design more effective libraries (25, 51–54). However, the analysis utility is limited by multisite 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.

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 *SI Appendix*.

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 10 assays. We found the limiting factors to be the capacity of HT sequencing and bacterial ligation efficiency. Given that an Illumina NovaSeq SP flowcell can achieve 400×10^6 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.

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 phosphate-buffered saline with 0.1% albumin (PBSA) (P_{Urea} was diluted using 3 M urea in PBSA and 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*, Fig. S1).

A total of 10 million yeast cells expressing the subsampled library were centrifuged at 5,000 *g* for 1 min, aspirated, resuspended in 1 mL cold PBSA, centrifuged, resuspended in 50 μL PBSA, and transferred to a 0.2-mL PCR tube on ice. A total of 50 μL 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 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.

Split GFP Assay. Frozen aliquots of cells were thawed and grown in 5 mL lysogeny broth (LB) + Amp + Kan overnight. Part of the overnight culture was added to 5 mL fresh LB + Amp + Kan at an optical density at 600 nm (OD_{600}) of 0.1 and grown for 90 min. Gp2-GFP_{11} production was induced by the addition of 0.5 mM isopropyl β-d-1-thiogalactopyranoside (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 (3,000 *g* 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. $GFP_{1\text{-}10}$ 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 (3,000 *g* for 10 min) and frozen at −80 °C to inhibit growth. The cells were then thawed and miniprepped to obtain the Gp2-encoding plasmids.

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 6×300 μL wells per concentration of ampicillin in a 96-well plate. A total of 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 5 min. 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.

HT Assay Score Calculations.

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 6.7× (on-yeast protease) and 7.9× (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 (Eq. **1**).

$$cells\ of\ sequence\ i\ in\ gate\ j = \frac{reads\ of\ sequence\ i}{total\ filtered\ reads\ in\ gate\ j} * \text{number of cells collected in gate j.} \quad [1]$$

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 (Eq. **2**). 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.

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

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.

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 ∼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, or 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.

Dot Blots to Quantify Expression.

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 overnight culture was added to 1 mL/well 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 (I^q) or 4 h (SH) followed by centrifugation (3,000 *g* for 5 min) 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 h. The plates were centrifuged (3,000 *g* for 5 min), and 25 μL/well soluble fraction was added to 25 μL/well denaturing buffer. Protein lysates from SH were diluted an additional 5× 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.

Dot blot protocol. A section of 0.2-μm pore polyvinylidene fluoride (1620177, BioRad) was cut to size and placed in a box (15.2 × 10.2 × 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% vol/vol Tween 20 in Tris-buffered saline) for 5 min. The membrane was then placed on a TBST-soaked filter paper and padded dry with a Kimwipe. Using a multichannel pipet, 2 μL/well 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% (wt/vol) nonfat dry milk in TBST) and rocked overnight at 4 °C. The membrane was then labeled with 50 mL 0.2 μg/mL anti His_6–horseradish peroxidase (ab1187, Abcam) in blocking solution for 30 min at room temperature. Excess antibody was washed via three washes of 50 mL TBST for 10 min at room temperature. The membrane was then soaked in 25 mL SuperSignal West Pico PLUS Chemiluminescent Substrate (ThermoFisher) for 5 min. Then membrane was then placed inside a transparency and exposed 10 to 30 s on a ChemiDoc MP Imaging System (BioRad).

Identification of HT Assay Predictiveness.

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.

CV 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 × 10 repeated K-fold CV 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 (55) library to determine the optimal hyperparameters for each architecture. We allowed 50 trials (or a maximum of 24 h of computational time for a feedforward neural network [FNN]) and recorded the trial with the lowest predictive error.

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

CFS. CFS identifies the optimal feature set by maximizing the relationship between features (x, HT assays) and target (y, yield) while minimizing the interfeature relationships (47). We calculated the CFS for every set (S_x) of 1,023 HT assay combinations. We defined the relationship (r) as the absolute value of Spearman's rank correlation coefficient (ρ) or the MI to capture linear and nonlinear relationships (Eq. **3**).

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

Subsampling training data. When evaluating the predictive performance of assays with varying number of training datapoints, we bootstrapped the dataset for CV 10 times. Each random dataset had separately optimized architectures and hyperparameters determined by CV. Due to the computational constraints, FNN architecture was not evaluated when subsampling the training dataset.

Propagation of uncertainty. Calculations involving propagation of uncertainty for predicted sample size were performed using the *uncertainties* (56) Python package.

Data Availability. Sequences, models, and analytics data have been deposited in GitHub (https://github.com/HackelLab-UMN/DevRep).

Table 1. Description of model architectures utilized when evaluating HT assay predictive performance

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]

"Uniform" and "quniform" refer to stochastic search spaces defined in the Python hyperopt library (57).

ACKNOWLEDGMENTS. This work was funded by the NIH (R01 EB023339) and an NSF 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 at the University of Minnesota.

1. T. Jain *et al*., Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 944–949 (2017).
2. Y. Xu *et al*., Structure, heterogeneity and developability assessment of therapeutic antibodies. *MAbs* **11**, 239–264 (2019).
3. M. I. J. Raybould *et al*., Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 4025–4030 (2019).
4. X. Yang *et al*., Developability studies before initiation of process development: Improving manufacturability of monoclonal antibodies. *MAbs* **5**, 787–794 (2013).
5. M. Bailly *et al*., Predicting antibody developability profiles through early stage discovery screening. *MAbs* **12**, 1743053 (2020).
6. A. M. Wolf Pérez *et al*., In vitro and in silico assessment of the developability of a designed monoclonal antibody library. *MAbs* **11**, 388–400 (2019).
7. J. C. Almagro, M. Pedraza-Escalona, H. I. Arrieta, S. M. Pérez-Tapia, Phage display libraries for antibody therapeutic discovery and development. *Antibodies (Basel)* **8**, 44 (2019).
8. J. A. Delmar, J. Wang, S. W. Choi, J. A. Martins, J. P. Mikhail, Machine learning enables accurate prediction of asparagine deamidation probability and rate. *Mol. Ther. Methods Clin. Dev.* **15**, 264–274 (2019).
9. X. Lu *et al*., Deamidation and isomerization liability analysis of 131 clinical-stage antibodies. *MAbs* **11**, 45–57 (2019).
10. P. M. Buck *et al*., Computational methods to predict therapeutic protein aggregation. *Methods Mol. Biol.* **899**, 425–451 (2012).
11. C. N. Magnan, A. Randall, P. Baldi, SOLpro: Accurate sequence-based prediction of protein solubility. *Bioinformatics* **25**, 2200–2207 (2009).
12. N. J. Agrawal *et al*., Computational tool for the early screening of monoclonal antibodies for their viscosities. *MAbs* **8**, 43–48 (2016).
13. M. I. J. Raybould *et al*., Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.U.S.A.* **116**, 4025–4030 (2019).
14. N. Chennamsetty, V. Voynov, V. Kayser, B. Helk, B. L. Trout, Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci.U.S.A.* **106**, 11937–11942 (2009).
15. T. M. Lauer *et al*., Developability index: A rapid in silico tool for the screening of antibody aggregation propensity. *J. Pharm. Sci.* **101**, 102–115 (2012).
16. V. Potapov, M. Cohen, G. Schreiber, 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).
17. A. Nisthal, C. Y. Wang, M. L. Ary, S. L. Mayo, Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci.U.S.A.* **116**, 16367–16377 (2019).
18. P. Estep *et al*., An alternative assay to hydrophobic interaction chromatography for high-throughput characterization of monoclonal antibodies. *MAbs* **7**, 553–561 (2015).
19. Q. Chai, J. Shih, C. Weldon, S. Phan, B. E. Jones, Development of a high-throughput solubility screening assay for use in antibody discovery. *MAbs* **11**, 747–756 (2019).
20. P. J. Carter, G. A. Lazar, Next generation antibody drugs: Pursuit of the 'high-hanging fruit'. *Nat. Rev. Drug Discov.* **17**, 197–223 (2018).
21. S. Banta, K. Dooley, O. Shur, Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.* **15**, 93–113 (2013).
22. B. J. Hackel, "Alternative protein scaffolds for molecular imaging and therapy" in *Engineering in Translational Medicine*, W. Cai, Ed. (Springer, London, 2014), pp. 343–364.
23. L. A. Stern, B. A. Case, B. J. Hackel, Alternative non-antibody protein scaffolds for molecular imaging of cancer. *Curr. Opin. Chem. Eng.* **2**, 425–432 (2013).
24. R. Vazquez-Lombardi *et al*., Challenges and opportunities for non-antibody scaffold drugs. *Drug Discov. Today* **20**, 1271–1283 (2015).
25. D. R. Woldring, P. V. Holec, L. A. Stern, Y. Du, B. J. Hackel, A gradient of sitewise diversity promotes evolutionary fitness for binder discovery in a three-helix bundle protein scaffold. *Biochemistry* **56**, 1656–1671 (2017).
26. S. Miersch, S. S. Sidhu, Synthetic antibodies: Concepts, potential and practical considerations. *Methods* **57**, 486–498 (2012).
27. M. A. Kruziki, S. Bhatnagar, D. R. Woldring, V. T. Duong, B. J. A. Hackel, A 45-amino-acid scaffold mined from the PDB for high-affinity ligand engineering. *Chem. Biol.* **22**, 946–956 (2015).
28. M. A. Kruziki, V. Sarma, B. J. Hackel, Constrained combinatorial libraries of Gp2 proteins enhance discovery of PD-L1 binders. *ACS Comb. Sci.* **20**, 423–435 (2018).
29. J. Y. Chan, B. J. Hackel, D. Yee, Targeting insulin receptor in breast cancer using small engineered protein scaffolds. *Mol. Cancer Ther.* **16**, 1324–1334 (2017).
30. B. A. Case, M. A. Kruziki, S. M. Johnson, B. J. Hackel, Engineered charge redistribution of Gp2 proteins through guided diversity for improved PET imaging of epidermal growth factor receptor. *Bioconjug. Chem.* **29**, 1646–1658 (2018).
31. F. Du *et al*., Engineering an EGFR-binding Gp2 domain for increased hydrophilicity. *Biotechnol. Bioeng.* **116**, 526–535 (2019).
32. A. W. Golinski, P. V. Holec, K. M. Mischler, B. J. Hackel, Biophysical characterization platform informs protein scaffold evolvability. *ACS Comb. Sci.* **21**, 323–335 (2019).
33. G. J. Rocklin *et al*., Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
34. S. C. Ritter, B. J. Hackel, Validation and stabilization of a prophage lysin of Clostridium perfringens by using yeast surface display and coevolutionary models. *Appl. Environ. Microbiol.* **85**, e00054-19 (2019).
35. J. R. Klesmith *et al*., Retargeting CD19 chimeric antigen receptor T cells via engineered CD19-fusion proteins. *Mol. Pharm.* **16**, 3544–3558 (2019).
36. S. Cabantous, G. S. Waldo, In vivo and in vitro protein solubility assays using split GFP. *Nat. Methods* **3**, 845–854 (2006).
37. L. Foit *et al*., Optimizing protein stability in vivo. *Mol. Cell* **36**, 861–871 (2009).
38. J. S. Ebo *et al*., An in vivo platform to select and evolve aggregation-resistant proteins. *Nat. Commun.* **11**, 1816 (2020).
39. T. W. Overton, Recombinant protein production in bacterial hosts. *Drug Discov. Today* **19**, 590–601 (2014).
40. G. Tian *et al*., Quantitative dot blot analysis (QDB), a versatile high throughput immunoblot method. *Oncotarget* **8**, 58553–58562 (2017).
41. J. Chen *et al*., Chaperone activity of DsbC. *J. Biol. Chem.* **274**, 19601–19605 (1999).
42. E. T. Boder, K. D. Wittrup, Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).
43. A. Galarneau, M. Primeau, L.-E. Trudeau, S. W. Michnick, β-lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein protein interactions. *Nat. Biotechnol.* **20**, 619–622 (2002).
44. T. J. Mansell, S. W. Linderman, A. C. Fisher, M. P. DeLisa, A rapid protein folding assay for the bacterial periplasm. *Protein Sci.* **19**, 1079–1090 (2010).
45. I.-K. Yeo, R. A. Johnson, A new family of power transformations to improve normality or symmetry. *Biometrika* **87**, 954–959 (2000).
46. K.-D. Jany, G. Lederer, B. Mayer, 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).
47. Hall, M. A. "Correlation-based feature selection for machine learning," PhD thesis, The University of Waikato, Hamilton, New Zealand. (1999).
48. A. Zutz *et al*., A dual-reporter system for investigating and optimizing protein translation and folding in E. coli. *bioRxiv* [Preprint] (2020). https://doi.org/10.1101/2020.09.18.303453 (Accessed 11 Feburary 2021).
49. S. A. Lesley, J. Graziano, C. Y. Cho, M. W. Knuth, H. E. Klock, Gene expression response to misfolded protein as a screen for soluble recombinant protein. *Protein Eng.* **15**, 153–160 (2002).
50. F. Hoffmann, U. Rinas, Stress induced by recombinant protein production in Escherichia coli. *Adv. Biochem. Eng. Biotechnol.* **89**, 73–92 (2004).
51. B. J. Hackel, M. E. Ackerman, S. W. Howland, K. D. Wittrup, Stability and CDR composition biases enrich binder functionality landscapes. *J. Mol. Biol.* **401**, 84–96 (2010).
52. F. A. Fellouse *et al*., High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J. Mol. Biol.* **373**, 924–940 (2007).
53. M. A. Seeger *et al*., Design, construction, and characterization of a second-generation DARP in library with reduced hydrophobicity. *Protein Sci.* **22**, 1239–1257 (2013).
54. A. Koide, J. Wojcik, R. N. Gilbreth, R. J. Hoey, S. Koide, Teaching an old scaffold new tricks: Monobodies constructed using alternative surfaces of the FN3 scaffold. *J. Mol. Biol.* **415**, 393–405 (2012).
55. J. Bergstra, D. Yamins, D. D. Cox, "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* (JMLR.org, 2013), vol. 28, pp. 115–123.
56. E. O. Lebigot, Uncertainties: A Python Package for Calculations with Uncertainties, Version 3.1.5 (2010). https://pythonhosted.org/uncertainties/. Accessed 24 May 2021.
57. J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, D. D. Cox, Hyperopt: A python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **8**, 14008 (2015).