Mega-scale experimental analysis of protein folding stability in biology and protein design

Authors: Kotaro Tsuboyama[1,2,3], Justas Dauparas[4,5], Jonathan Chen[1,2], Niall M. Mangan[2,7], Sergey Ovchinnikov[8], Gabriel J. Rocklin[1,2] *

Affiliations:

[1] Department of Pharmacology, Northwestern University Feinberg School of Medicine; Chicago, IL, 60611 USA

[2] Center for Synthetic Biology, Northwestern University; Evanston, IL, 60208 USA

[3] PRESTO, Japan Science and Technology Agency; Chiyoda-ku, Tokyo, 102-0076, Japan

[4] Department of Biochemistry, University of Washington; Seattle, WA, 98195 USA

[5] Institute for Protein Design, University of Washington; Seattle, WA, 98195 USA

[6] Master of Biotechnology Program, McCormick School of Engineering, Northwestern University; Evanston, IL, 60208 USA

[7] Department of Engineering Sciences and Applied Mathematics, Northwestern University; Evanston, Illinois 60208, USA

[8] John Harvard Distinguished Science Fellowship Program, Harvard University; Cambridge, MA, 02138 USA

*Corresponding author. Email: grocklin@gmail.com

Abstract:

Advances in DNA sequencing and machine learning are illuminating protein sequences and structures on an enormous scale. However, the energetics driving folding are invisible in these structures and remain largely unknown. The hidden thermodynamics of folding can drive disease, shape protein evolution, and guide protein engineering, and new approaches are needed to reveal these thermodynamics for every sequence and structure. We present cDNA display proteolysis, a new method for measuring thermodynamic folding stability for up to 900,000 protein domains in a one-week experiment. From 1.8 million measurements in total, we curated a set of ~850,000 high-quality folding stabilities covering all single amino acid variants and selected double mutants of 354 natural and 188 de novo designed protein domains 40-72 amino acids in length. Using this immense dataset, we quantified (1) environmental factors influencing amino acid fitness, (2) thermodynamic couplings (including unexpected interactions) between protein sites, and (3) the global divergence between evolutionary amino acid usage and protein folding stability. We also examined how our approach could identify stability determinants in designed proteins and evaluate design methods. The cDNA display proteolysis method is fast, accurate, and uniquely scalable, and promises to reveal the quantitative rules for how amino acid sequences encode folding stability.

One-Sentence Summary:

Massively parallel measurement of protein folding stability by cDNA display proteolysis

Main Text:

Protein sequences vary by more than ten orders of magnitude in thermodynamic folding stability (the ratio of unfolded to folded molecules at equilibrium) (Dill, 1990; Goldenzweig and Fleishman, 2018). Even single point mutations that alter stability can have profound effects on health and disease (Stein et al., 2019; Wang and Moult, 2001; Yue et al., 2005), pharmaceutical development (Rodríguez-Rodríguez et al., 2012; Stutz and Blein, 2020; Wang et al., 2021), and protein evolution (Agozzino and Dill, 2018; Bloom et al., 2006; Drummond and Wilke, 2008; Gong et al., 2013). Thousands of point mutants have been individually studied over decades to quantify the determinants of stability (Nikam et al., 2021; Xavier et al., 2021), but these studies highlight a challenge: similar mutations can have widely varying effects in different protein contexts, and these subtleties remain difficult to predict despite substantial effort (Laimer et al., 2015; Schymkowitz et al., 2005). In fact, even as deep learning models have achieved transformative accuracy at protein structure prediction (Baek et al., 2021; Chowdhury et al., 2022; Jumper et al., 2021; Wu et al., 2022) progress in modeling folding stability has arguably stalled (Broom et al., 2020; Pak et al., 2021; Pucci et al., 2022). New high-throughput experiments have the potential to transform our understanding of stability by quantifying the effects of mutations across a vast number of protein contexts, revealing new biophysical insights and empowering modern machine learning methods.

Here, we present a powerful new high-throughput stability assay along with a uniquely massive dataset of 851,552 folding stability measurements. Our new method - cDNA display proteolysis - combines the strengths of cell-free molecular biology and next-generation sequencing and requires no on-site equipment larger than a qPCR machine. Assaying one library (up to 900,000 sequences in our experiments) requires one week and only ~$2,000 in reagents, excluding the cost of DNA synthesis and sequencing. Compared to mass spectrometry-based high-throughput stability assays (Jarzab et al., 2020; Savitski et al., 2014; Van Vranken et al., 2021; Walker et al., 2019), cDNA display proteolysis achieves a 100-fold larger scale and can easily be applied to study mutational libraries that pose difficulties for proteomics. Compared to our earlier yeast display proteolysis method (Rocklin et al., 2017), cDNA display proteolysis resolves a wider dynamic range of stability and is more reproducible even at a 50-fold larger experimental scale. Large-scale proteolysis data have already played a key role in the development of machine learning methods for protein design and protein biophysics (Alley et al., 2019; Biswas et al., 2021; Ingraham et al.; Rao et al., 2019; Strokach et al., 2020, 2021; Zhou et al., 2020). The cDNA display proteolysis method massively expands this capability and has the potential to expand our knowledge of stability to the scale of all known small domains.

Our new dataset of 851,552 absolute folding stabilities is unique in size and character. Current thermodynamic databases contain a skewed assortment of mutations measured under many varied conditions (Nikam et al., 2021). In contrast, our new dataset comprehensively measures all single mutants for 354 natural domains and 188 designed proteins - including single deletions and two insertions at each position - all under identical conditions. Our dataset also includes comprehensive double mutations at 595 site pairs spread across 208 domains (a total of 222,265 double mutants). By maintaining uniform experimental conditions, our data can be used to examine the determinants of *absolute* folding stability in addition to the effects of mutations. Using our unique dataset, we investigated how individual amino acids and pairs of amino acids contribute to folding stability (Figs. 3 and 4) as well as how selection for stability interacts with other selective pressures in natural protein domains (Figs. 5 and 6). We also explored how our unique scale of data can be applied in protein design (Fig. 7).

Massively parallel measurement of folding stability by cDNA display proteolysis

Proteases typically cleave unfolded proteins more quickly than folded ones, and proteolysis assays have been used for decades to measure folding stability (Park and Marqusee, 2005) and select for high stability proteins (Park et al., 2007; Sieber et al., 1998). In 2017, we introduced the high-throughput yeast display proteolysis method for measuring folding stability using next generation sequencing (Basanta et al., 2020; Dou et al., 2018; Golinski et al., 2019, 2021; Maguire et al., 2021; Norn et al., 2021; Rocklin et al., 2017; Singer et al., 2022). To improve the scale, precision, speed, and cost of stability measurements, we developed cDNA display proteolysis. Each experiment begins with a DNA library. Here, we employ synthetic DNA oligo pools where each oligo encodes one test protein. The DNA library is transcribed and translated using cell-free cDNA display (Yamaguchi et al., 2009), based on mRNA display (Nemoto et al., 1997; Roberts and Szostak, 1997), resulting in proteins that are attached at the C-terminus to their cDNA. We then incubate the protein-cDNA complexes with different concentrations of protease, quench the reactions, and pull down the proteins using an N-terminal PA tag (Fig. 1A). Intact (protease-resistant) proteins will also carry their C-terminal cDNA. Finally, we determine the relative amounts of all proteins in the surviving pool at each protease concentration by deep sequencing (Fig. 1D). To control for any effects of protease specificity, we perform separate experiments with two orthogonal proteases: trypsin (targeting basic amino acids) and chymotrypsin (targeting aromatic amino acids).

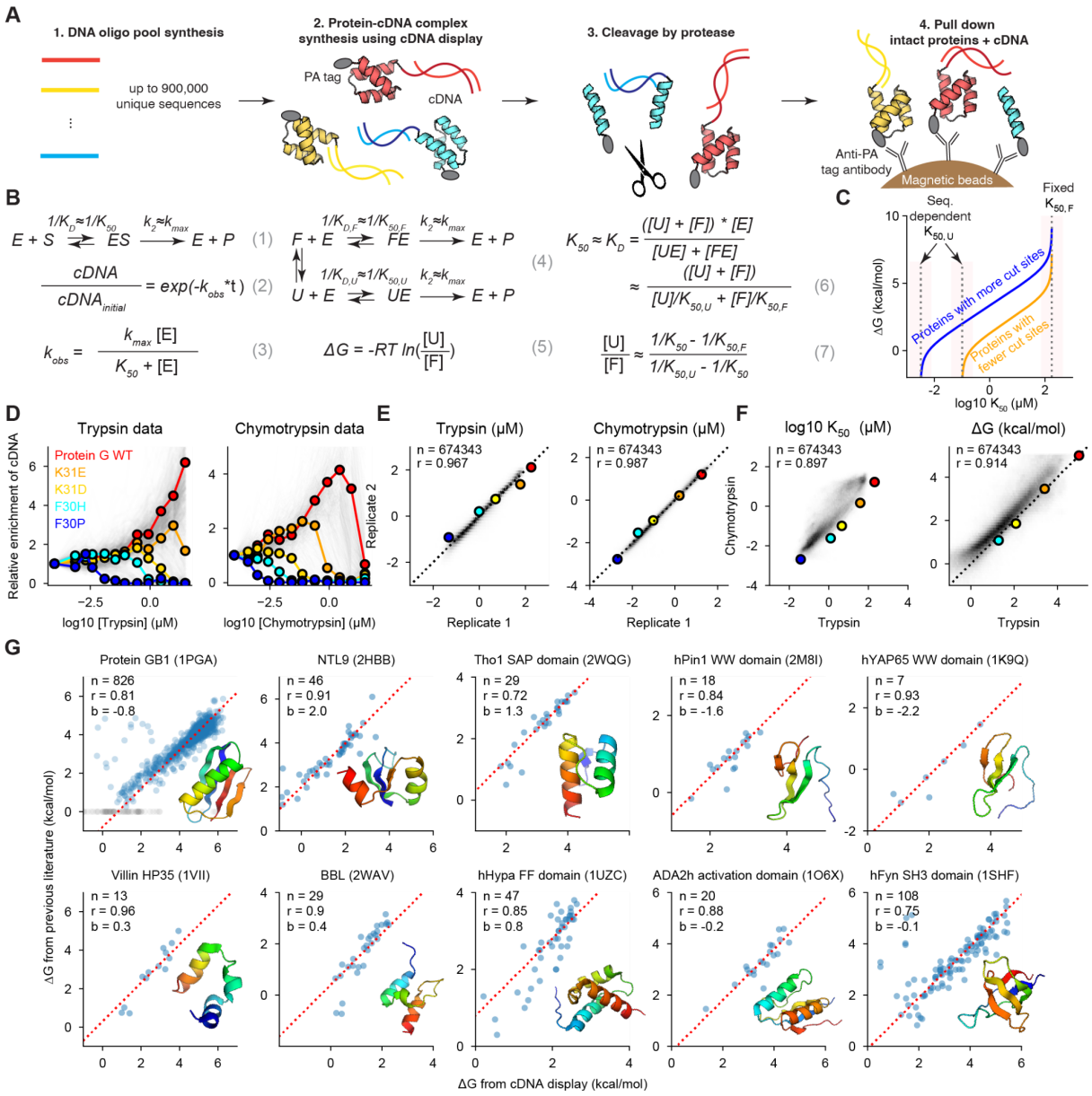


Fig. 1. cDNA display enables massively parallel measurement of protein folding stability.

(A) A DNA oligo library is expressed using cell-free cDNA display, producing proteins with an N-terminal PA tag and C-terminal covalent attachment to cDNA. Protease cleavage separates the cDNA from the PA tag. After protease challenge, magnetic beads with anti-PA antibodies pull down protein N-termini and intact proteins carry along their cDNA. cDNA is then amplified and sequenced to quantify the intact fraction of each protein.

(B) Thermodynamic model of proteolysis based on single turnover kinetics. Protease enzymes (E) and protein substrates (S) form an ES complex to produce cleaved protein products (P) (1). We model the cleavage as a first-order reaction (2) according to single turnover kinetics (3). We use an identical k_{max} for all sequences and fit each sequence's K_{50} concentration to our data. Proteins are normally cleaved in the unfolded (U) state but can also be cleaved in the folded (F) state (e.g. by cleaving the PA tag) (4). We determine the folding equilibrium using each sequence's measured K_{50}, a predicted sequence-specific K_{50} for the unfolded state ($K_{50,U}$), and a universal K_{50} for the folded state ($K_{50,F}$).

(C) Relationship between K_{50} and ΔG for a protein with fewer cut sites (yellow) and a protein with more cut sites (blue). When K_{50} approaches $K_{50,U}$ or $K_{50,F}$ (red shaded regions), ΔG becomes very sensitive to K_{50} and its uncertainty increases relative to the uncertainty in K_{50}.

(D) PA tag pulldown at increasing protease concentrations separates proteins by stability. Each sequence of Protein GB1 variants in a library is shown as a gray line tracking its change in population fraction relative to that in the pre-selection library (enrichment). Enrichment traces for the wild-type and four mutants are highlighted in color.

(E) Reproducibility of K_{50} from two replicates of the proteolysis procedure, after filtering for data quality and range (see Methods). The K_{50} density is shown in gray with the proteins from (D) highlighted in color.
(F) Consistency of K_{50} (left) and ΔG (right) between trypsin and chymotrypsin for one library (black), highlighting the five proteins shown in (D).
(G) Our high-throughput ΔG measurements are consistent with previously published stability data from purified protein samples for wild types and mutants of 10 domains. The red dashed line represents the Y = X+b (intercept) line. Gray points (Protein GB1) indicate 'no data' in the previous paper. See Table. S2 and Fig. S3 for analysis of the intercepts

---

We inferred the protease stability of all sequences from our sequencing counts using a Bayesian model of the experimental procedure. We modeled protease cleavage using single turnover kinetics (Coey and Drohat, 2017; Yourik et al., 2019) (Fig. 1B eqs. 1 to 3, Fig. S1, and Supplementary Text for the derivation) because we assume the enzyme is in excess over all substrates (up to ~20 pM of substrate based on previous estimates (Yamaguchi et al., 2009) versus 141 pM for the lowest concentration of protease). To parameterize the model, we used a universal k_{max} cleavage rate for all sequences (Fig. S1) and used our sequencing data to infer a unique K_{50} for each sequence (the protease concentration at which the cleavage rate is one-half k_{max}, see Methods). The K_{50} values inferred by the model were consistent between two replicates of the proteolysis procedure (R = 0.97 for trypsin and 0.99 for chymotrypsin for ~84% of sequences in a pool of 806,640 sequences after filtering based on confidence and dynamic range; Fig. 1E).

To infer each sequence's thermodynamic folding stability (ΔG for unfolding), we used a kinetic model that separately considers idealized folded (F) and unfolded (U) states (Fig. 1B eq. 4). We model both states using the same single-turnover equations as before (Fig. 1B eq. 3), with separate K_{50} protease concentrations for each state ($K_{50,F}$ and $K_{50,U}$) and a shared k_{max}. We assume that cleavage in the folded state exclusively occurs outside the folded domain (e.g. in the N-terminal PA tag added to all sequences), so we use an identical $K_{50,F}$ for all sequences. In contrast, $K_{50,U}$ reflects an individual sequence's unique protease susceptibility in the unfolded state, which depends on its potential cleavage sites. We inferred $K_{50,U}$ for each sequence using a position-specific scoring matrix (PSSM) model of protease cleavage parameterized using measurements of 64,238 scrambled sequences (sequences with a high probability of being fully unfolded, Fig. S2; see also Methods). Finally, we assume that folding, unfolding, and enzyme binding are all in rapid equilibrium relative to cleavage, implying that $K_{50,U}$, $K_{50,F}$, and the overall K_{50} can be approximated by the enzyme-substrate equilibrium dissociation constants for each state (Fig. 1B eq. 6). Although these approximations will not be universally accurate, they are appropriate for the small domains examined here and facilitate consistent analysis of all test sequences. With these approximations, we can express a sequence's ΔG in terms of the universal $K_{50,F}$, its inferred $K_{50,U}$, and its experimentally measured K_{50} (Fig. 1B eq. 5 and 7, and Supplementary Text for the derivation). For most analysis we combine our independent trypsin and chymotrypsin data into a single overall ΔG estimate (See Methods). Based on our kinetic model, (1) stability (ΔG) will be underestimated if significant cleavage occurs inside the test domain in the folded state, (2) stability can be over- or under-estimated depending on the accuracy of $K_{50,U}$ (independent measurements with trypsin and chymotrypsin help correct this), and (3) ΔG values become unreliable if K_{50} approaches $K_{50,F}$ or $K_{50,U}$ (Fig. 1C).

Folding stabilities from cDNA display proteolysis are consistent with traditional experiments on purified proteins

In Fig. 1G, we compare stabilities measured by cDNA display proteolysis to previous results from experiments on purified protein samples for 1,143 variants of ten proteins (Araya et al., 2012; Dodson and Arbely, 2015; Jäger et al., 2009; Jemth et al., 2005; Jiang et al., 2001; Maxwell and Davidson, 1998; Neuweiler et al., 2009; Nisthal et al., 2019; Northey et al., 2002; de los Rios et al., 2005; Sato et al., 2017; Villegas et al., 1998; Xiao et al., 2009, 2013). All Pearson correlations are above 0.7. Our stability measurements for these 1,143 sequences were all performed in libraries of 244,000–900,000 total sequences. Although several sets of mutants show systematic offsets (y-intercept values) between literature values and our measurements, these offsets correlate with temperature differences between experimental conditions (with the exception of the N-terminal domain of Ribosomal Protein L9 (2HBB), Fig. S3, see Table S2 for all experimental conditions). We also noticed several variants of Protein GB1 appear unstable in our data but stable in the previous experiments (Nisthal et al., 2019). Our structural analysis of these mutations suggests that our measurements are more likely to be correct (Fig. S4). Overall, the consistency between our cDNA display proteolysis data and traditional biophysical measurements establishes that (1) small domains are cleaved mainly in the globally unfolded state, (2) our method can reliably measure these cleavage rates on a massive scale, and (3) our unfolded state model can remove protease-specific effects to attain accurate quantitative folding stability measurements.

Comprehensive mutational analysis across designed and natural protein domains

To systematically examine how individual residues influence folding stability, we used cDNA display proteolysis to measure stability for all single substitutions, deletions, and Gly and Ala insertions in 983 natural and designed domains. We chose our natural domains to cover almost all of the small monomeric domains in the Protein Data Bank (30-72 amino acids in length). Our designed domains included (1) previous Rosetta designs with ααα, αββα, βαββ, and ββαββ topologies (40-43 a.a.) (Kim et al., 2021; Rocklin et al., 2017), (2) new ββαα proteins designed using Rosetta (47 a.a.), and (3) new domains designed by trRosetta hallucination (46 to 69 a.a.) (Anishchenko et al., 2021; Norn et al., 2021). We collected these data using four giant synthetic DNA oligonucleotide libraries and obtained K_{50} values for 2,520,337 sequences; 1,844,548 of these measurements are included here. K_{50} values were reproducible across libraries (Fig. S5). Oligo pools were synthesized by Agilent Technologies (one

244,000-sequence library, length 170 nt) and Twist Bioscience (three libraries of 696,000 - 900,000 sequences, length 250-300 nt).

Deep mutational scanning of hundreds of domains revealed several overall patterns. The largest fraction of these domains showed clear, biophysically reasonable sequence-stability relationships that were consistent between separate experiments with trypsin and chymotrypsin. However, other domains were completely unfolded, too stable to resolve, insensitive to mutation, or inconsistent between the proteases. For 42 domains that were too stable to resolve, we introduced single mutations to destabilize the wild-type sequence, then performed new mutational scanning experiments in these 121 new "wild-type" backgrounds (Fig. S6). In four domains, mutational scanning revealed trypsin-sensitive loops that could be cleaved in the folded state, leading to inconsistent stabilities between trypsin and chymotrypsin (Fig. S7). In these cases, we introduced one to two substitutions into the wild-type sequences to remove trypsin-sensitive sites, then performed new mutational scanning experiments in these alternative backgrounds. This led to consistent results between the two proteases. In total, we performed deep mutational scanning for 983 domain sequences, including both original and revised wild-type backgrounds.

Our overall categorization of all domains is shown in Fig. 2B (see Fig. S8 for inclusion criteria). Based on these categories, we assembled three curated datasets for machine learning (Fig. 2A). Our ΔΔG dataset (Dataset #1) includes 586,938 sequences (single and double mutants) from 251 natural domains and 145 designs. In this dataset, the wild-type sequence is 1.25-4.5 kcal/mol in stability so that most ΔΔG values (including for stabilizing mutants) are correctly resolved. Our ΔG dataset (Dataset #2) includes all 851,552 single and double mutants from 354 natural domains and 188 designs. In this dataset, the large majority of mutant ΔGs are accurately resolved, but the wild-type ΔG may lie outside the dynamic range, preventing accurate ΔΔG calculations. Finally, Dataset #3 includes all ~1.8 million confidently estimated K_{50} values, even when trypsin and chymotrypsin measurements produced inconsistent ΔG estimates. The main domain classes in Dataset #1 are shown in Fig. 2C; all natural domains included in Dataset #1 are listed by category in Fig. S9 (see Supplementary Materials for all wild-type sequences).

Mutational scanning results for nine domains are shown in Fig 2D and E. Like all mutational scans in Datasets 1 and 2, these examples show a strong consistency between independent ΔG measurements with trypsin and chymotrypsin (Pearson correlation 0.94 ± 0.04 for 542 domains in Dataset 2, median ± std.). In each structure, sites are colored according to the average effect of an amino acid substitution, with the most critical sites (where mutations are very destabilizing) colored dark blue. Most of these critical sites are in the hydrophobic core. However, our data also reveal numerous other critical interactions, such as a side chain hydrogen bond between S23 and D42 in the U-box domain of human E4B Ubiquitin ligase and a cation-π interaction between R10 and W32 in the chromodomain of human chromobox protein homolog 7 (residues have been re-numbered based on the exact sequence included in our experiments). These unique stabilizing interactions reveal the rich biophysical diversity found in our systematic exploration of stability across hundreds of domains.

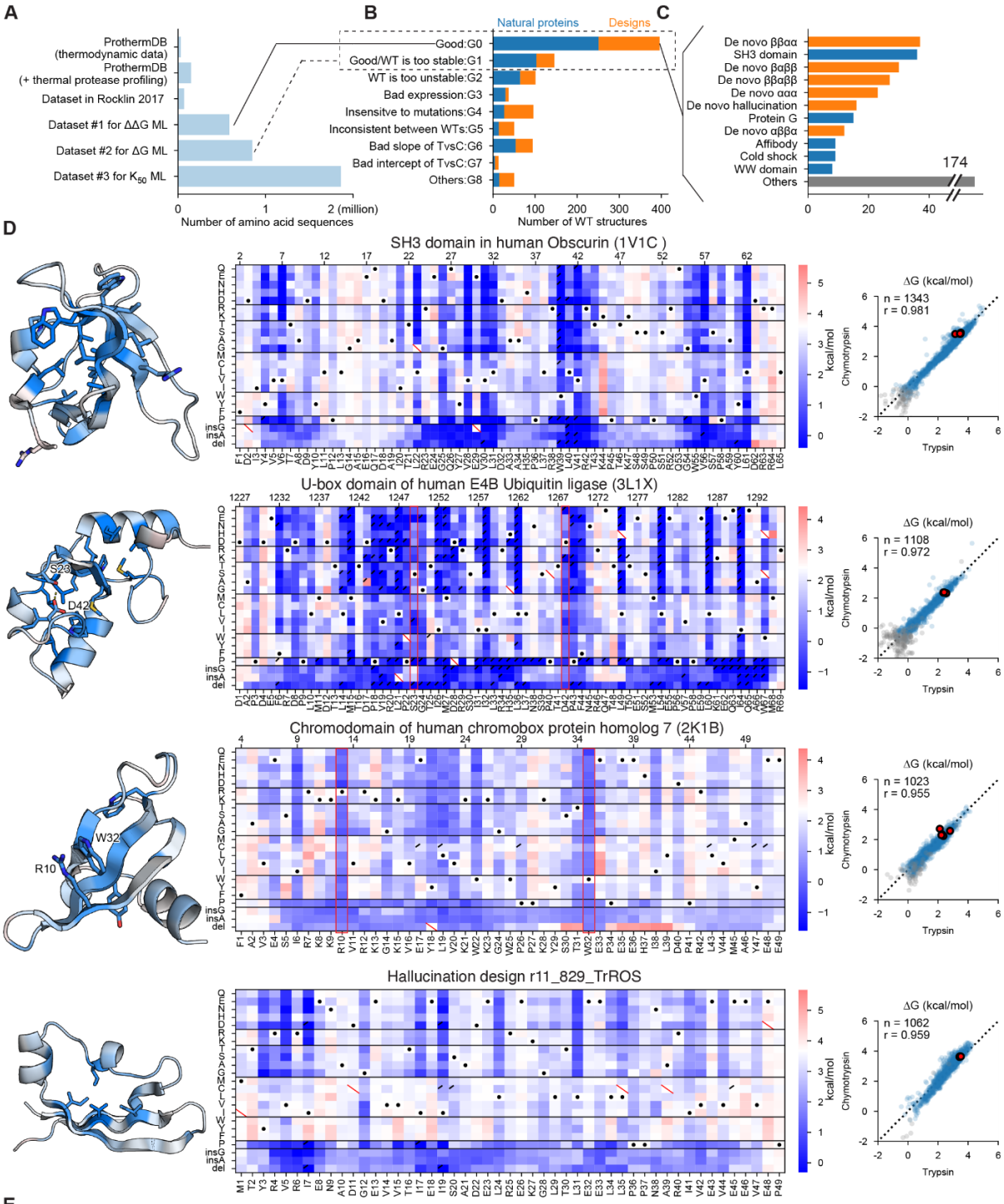

A
ProthermDB (thermodynamic data)
ProthermDB (+ thermal protease profiling)
Dataset in Rocklin 2017
Dataset #1 for ΔΔG ML
Dataset #2 for ΔG ML
Dataset #3 for K50 ML
Number of amino acid sequences
B
Natural proteins
Designs
Good:G0
Good/WT is too stable:G1
WT is too unstable:G2
Bad expression:G3
Insensitve to mutations:G4
Inconsistent between WTs:G5
Bad slope of TvsC:G6
Bad intercept of TvsC:G7
Others:G8
Number of WT structures
C
De novo ββαα
SH3 domain
De novo βαββ
De novo ββαββ
De novo ααα
De novo hallucination
Protein G
De novo αββα
Affibody
Cold shock
WW domain
Others
174
D
SH3 domain in human Obscurin (1V1C)
U-box domain of human E4B Ubiquitin ligase (3L1X)
Chromodomain of human chromobox protein homolog 7 (2K1B)
Hallucination design r11_829_TrROS
ΔG (kcal/mol)
n = 1343
r = 0.981
n = 1108
r = 0.972
n = 1023
r = 0.955
n = 1062
r = 0.959
Chymotrypsin
Trypsin
E
N-terminal domain of phage 434 repressor (1R69)
Anti-CRISPR AcrIF7 (6M3N)
N-terminal domain of E. coli arginine repressor (1AOY)
r6_560_TrROS
r6_831_TrROS

Fig. 2. Comprehensive mutational analysis of stability in designed and natural proteins

(A) Comparison of the size of existing datasets and the datasets from this paper. The data of this paper are divided into three groups: datasets #1, #2, and #3, according to the quality of the data (see Table S1 and Fig. S8 for details). ML: machine learning

(B) Classification of mutational scanning results for each wild-type sequence. The G0 group corresponds to Dataset #1, and G0 and G1 groups combined correspond to Dataset #2 in (A). (G1: Good but WT may be outside the dynamic range)

(C) Wild-type structures classified as G0 in (B) grouped into domain families. The 11 most common domain types are shown; the remaining 174 domains are classified as "Other" (see Fig. S9).

(D) Mutational scanning results for four domains. Left: domain structures colored by the average ΔΔG at each position; darker blue indicates mutants are more destabilizing. The structure of the design r11_829_TrROS is an AlphaFold model. Middle: Heat maps show ΔG for substitutions, deletions, and Gly and Ala insertions at each residue, with the PDB numbering shown at top and our one-indexed numbering at bottom. White represents the wild-type stability and red/blue indicate stabilizing/destabilizing mutations. Black dots indicate the wild-type amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5 kcal/mol), including ΔG estimates near the edges of the dynamic range. Red boxes highlight the S23-D42 hydrogen bond in 3L1X and the R10-W32 cation-π interaction in 2K1B. ΔG values were fit to trypsin and chymotrypsin data together; see Methods. Right: Agreement between mutant ΔG values independently determined using assays with trypsin (x-axis) and chymotrypsin (y-axis). Multiple codon variants of the wild-type sequence are shown in red, reliable ΔG values in blue, and less reliable ΔG estimates (same as above) in gray. The black dashed line represents Y=X. Each plot shows the number of reliable points and the Pearson r-value for the blue (reliable) points.

(E) As in (D) left, structures of five domains are shown colored by the average ΔΔG at each position; darker blue indicates mutants are more destabilizing. The two designed structures are AlphaFold models.

Trends in amino acid fitness at different sites and across domains

We first sought to define the major sources of variation between protein sites that determine the relative stabilities of all 20 amino acids at that site (i.e. the site's stability landscape). To this end, we performed principal component (PC) analysis using 293,697 ΔG measurements at 15,440 sites in 337 domains from Dataset #1 after centering our data to set the average ΔG at each site to zero (Fig. 3A, B). Each principal component expresses specific properties of a site that determine which amino acids are stabilizing or destabilizing. Based on the loadings of the different amino acids onto each principal component (Fig. 3C), we interpreted the first four components to reflect amino acid hydrophobicity (PC1; 31% of the total variance explained by this PC), helical probability (PC2; 15%), aliphatic vs. aromatic favorability (PC3; 12%), and positive vs. negative charge (PC4; 7%). The fifth principal component (6%) was more complex: at one extreme were small amino acids that could be buried in dense environments, along with positively charged amino acids that can "snorkel" their charged moieties to the surface even when partially buried. At the other extreme were negatively charged amino acids that are energetically costly to bury. We interpreted this component to reflect an "ease of burial" that is orthogonal to the hydrophobic property captured by PC1. These interpretations are also consistent with the structural environments at each site, as shown in Fig. 3D. For example, the first principal component reflecting hydrophobicity is high at buried positions and low at exposed positions (Fig. 3D).

These first five principal components collectively form a coarse model of the properties of protein sites, but some sites have unique stability landscapes that cannot be accurately represented by this model. We reconstructed the stability landscapes at all sites using the first five components and examined how different sites and domains deviated from these simplified landscapes (Fig. 3E). On average, stability landscapes reconstructed using five principal components were similarly accurate (in terms of mean absolute error) for both high and low stability domains (Fig. 3F). However, as expected, these coarse reconstructions were less accurate for domains with more varied stability landscapes (domains with a higher standard deviation of ΔG for all substitutions). The coarse model was also more accurate at reconstructing the stability landscapes of de novo designed domains and less accurate at reconstructing the landscapes of natural domains (Fig. 3F). This remained true for any number of principal components and even when designed proteins were excluded from the initial PCA (Fig. S10). This indicates that the de novo design protocols examined here lead to structures with "typical" amino acid environments that can be accurately described by only five principal components, and that these proteins generally lack the more specialized environments found in natural domains. Indeed, wild-type amino acids in natural domains tend to be more stable than the fit from the coarse model (Fig. S11). This suggests the remaining components capture additional biophysical effects that contribute to the compatibility between wild-type amino acids and their environments.

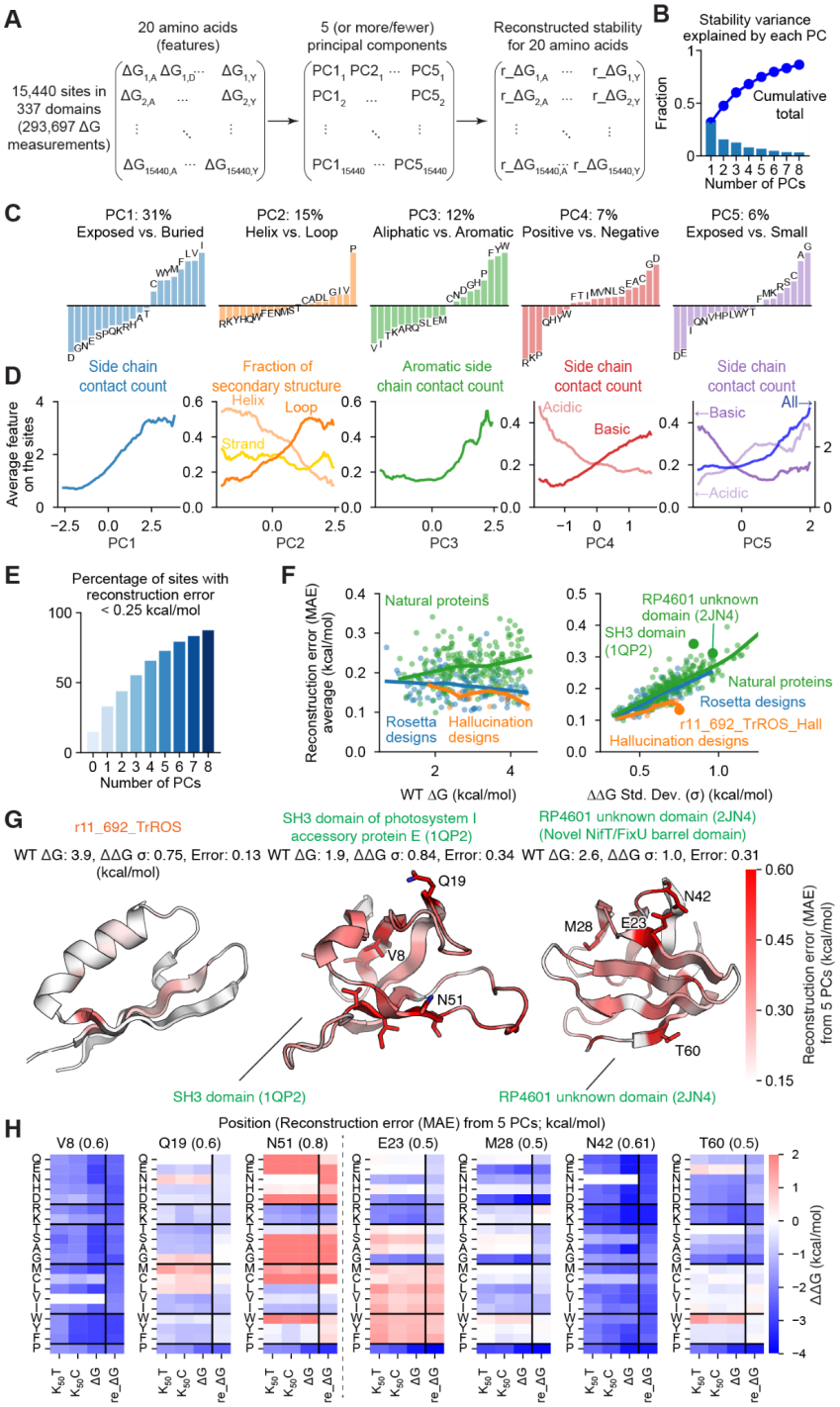


Fig. 3. Environmental factors that determine amino acid stabilities at a position.

(A) Principal component (PC) analysis on a matrix consisting of 15,440 observations (sites in proteins) x 20 amino acids (features) to determine the factors influencing stabilities of different amino acids.

(B) Fraction of the total variance explained by each PC (bars) and the cumulative total (upper line).

(C) Principal components of the stability data indicating the dominant trends for which amino acids would be stable or unstable at a site. We label each component with a biophysical interpretation and show the percent of the total variance explained by that component.

(D) Relationships between the PC values (x-axis) for all 15,440 positions and environmental properties of the position from the three-dimensional structure (y-axis). Colored lines show each environmental feature averaged over a window of 0.25 in the units of each PC.

(E) Fraction of sites (observations) whose stability landscapes can be reconstructed with MAE < 0.25 kcal/mol using the first 1-8 principal components.

(F) Relationship between reconstruction error using five PCs (MAE, y-axis) and wild-type stability (left, x-axis) or variance in the ΔΔG data (right, x-axis). Colors represent protein structures grouped into natural proteins (green), Rosetta designs (blue), and hallucination designs (orange). Three example proteins shown in (G) are shown as large dots.
(G) Structures of three example proteins with each position colored by the error (MAE) between the reconstructed ΔΔG values (using 5 PCs) and the observed ΔΔG values. The left protein was designed by hallucination and each position is accurately reconstructed using only 5 PCs, whereas the middle and right natural proteins have positions with more unusual ΔΔG patterns and larger MAEs. The r11_692_TrROS structure is an AlphaFold model.
(H) For seven positions with large MAE in the center (1QP2) and right domains (2JN4) from (G), we show the experimental trypsin and chymotrypsin K50 values, the ΔG values, and the reconstructed ΔG values based on the top five PCs.

Three example proteins shown in Fig. 3G illustrate how the coarse five-component model captures (or fails to capture) protein stability landscapes. At one extreme, the stability landscape of the designed protein r11_692_TrROS (from trRosetta hallucination) is accurately approximated by the coarse model (average per-residue MAE 0.13 kcal/mol). In contrast, the two natural domains (an SH3 domain (1QP2) and a unique NifT/FixU barrel domain (2JN4); Fig. S12) contain many sites with unique properties that are not accurately represented by the model (average per-residue MAE of 0.34 kcal/mol and 0.31 kcal/mol for the SH3 domain and β-barrel domains respectively). Seven of these sites are highlighted in Fig. 3H. Each stability landscape contains sharp differences between closely related amino acids that are not captured by the coarse model, such as V versus L at V8 and Q19 in 1QP2, and Q versus E at Q19 in 1QP2, M28 in 2JN4, and T60 in 2JN4. These unusual patterns are unlikely to be experimental artifacts because the patterns are consistent between independent experiments with trypsin and chymotrypsin and the same patterns are seen in both our K_{50} and ΔG analysis (Fig. 3H). Our massive dataset enabled us to identify the global trends in stability landscapes as well as specific cases that depart from these trends. These unusual cases with large reconstruction errors may provide the opportunity to study how protein flexibility and/or rare side chain interactions contribute to folding stability. These unusual sites will also serve as stringent test cases for models of protein stability.

Quantifying thermodynamic coupling for hundreds of amino acid pairs

Next, we examined how side chain interaction between amino acid pairs affects folding stability. We constructed comprehensive substitutions (20 x 20 amino acids) of 595 amino acid pairs from 208 natural domains and designs in our ΔG dataset (Dataset #2) and measured stability for all sequences by cDNA display proteolysis. We selected pairs that were suggested to form energetically important hydrogen bonds in our mutational scanning data as well as other pairs forming close contacts (Fig. 4A; Methods). To quantify the interactions between side chains, we constructed an additive model for each amino acid pair with 40 coefficients that capture the independent stability contributions of each amino acid in each position. The deviations from these models quantify the "thermodynamic coupling" between specific amino acids. Among our curated set of wild-type pairs, thermodynamic couplings were typically 0.5-1.0 kcal/mol in magnitude, with the largest couplings stronger than 2 kcal/mol (Fig. 4B). Among all sequences tested (wild-type or mutant pairs), pairs with opposite charges and cysteine pairs tended to have positive (favorable) couplings, whereas pairs with the same charge and acidic-aromatic/aliphatic amino acid pairs tended to have negative couplings (Fig. 4C). These couplings are lower than our observed wild-type couplings because the side chain orientations and environment surrounding wild-type pairs will typically be optimized for that pair. Nonetheless, our data recapitulate expected patterns of side chain interactions, provide a wealth of data for training machine learning models, and identify a wide range of noteworthy interactions for further study.

Several notable pairs are highlighted in Fig. 4D to F. In an OB-domain from *Shewanella oneidensis*, we found strong thermodynamic coupling between two unrelated pairs of amino acids: the wild-type Tyr-Glu pair and a mutant Lys-Trp pair that may form a cation-π interaction (thermodynamic couplings of 1.6±0.2 and 1.4±0.2 kcal/mol respectively; mean±std from calculating the coupling using bootstrap resampling of the ~400 amino acid combinations; Fig. 4D, S13A). In the Alpha-spectrin SH3 domain, our comprehensive double mutant scanning of Y10 and Y52 uncovered the highly stable, tightly coupled double mutant Y10H/Y52K (coupling of 2.5±0.4 kcal/mol for His-Lys versus 1.0±0.2 kcal/mol for the wild-type pair) (Fig. 4E, S13B). AlphaFold modeling predicts that this double mutant introduces a new hydrogen bonding network to replace the original Tyr-Tyr interaction. We also identified an unexpected thermodynamic coupling between an amino acid pair lacking a direct side chain interaction. In the SH3 domain of Myo3, mutations at K24 are destabilizing even though the side chain makes no clear interactions. To investigate interactions of K24, we quantified thermodynamic couplings to nearby Y9 (0.0±0.1 kcal/mol) and D10 (1.0±0.2 kcal/mol) (Fig. 4F and Fig. S13C). The surprising K24-D10 coupling - between two side chains that appear not to interact - highlights the difficulty of inferring energetic interactions from structural data alone, and suggests a possible longer-ranged ionic interaction.

We also investigated thermodynamic couplings within 36 different three-residue networks. For each triplet, we comprehensively measured stability for all possible single and double substitutions in both the wild-type background and in the background where the third amino acid was replaced by alanine (400 mutants x 3 pairs x 2 backgrounds = ~2,400 mutants in total for each triplet). As before, we modeled each set of 400 mutants (i.e. one residue pair in one background) using 40 single-amino acid coefficients (we did not globally model all 2,400 mutants together). One notable triplet is found in the J domain of HSJ1a, where R60 and D64 both interact

with the hydroxyl group on Y3 (Fig. 4G left). We observe strong couplings (> 1.5 kcal/mol) between each pair of two out of the three amino acids. However, when any of the three amino acids is mutated to alanine, the coupling between the remaining two amino acids becomes much weaker (< 0.5 kcal/mol, Fig. 4G middle and right, Fig. S13D). These results reveal a strong third-order thermodynamic coupling: the interaction between two amino acids is mediated by a third amino acid.

This strong three-way coupling is especially noteworthy because the interactions do not appear in the deposited NMR structural ensemble (2LGW; Fig S14A and B). The interaction network shown in Fig. 4G comes from the AlphaFold predicted structure for our wild-type sequence taken from the J domain of human HSJ1a. This network reproduces interactions seen in other J-domain crystal structures from *C. elegans* (2OCH) and *P. falciparum* (6RZY). However, in the deposited NMR ensemble for 2LGW, the backbone near Y5 (Y3 in our numbering) always positions that residue away from the helix containing R62 and D66, making the interaction network impossible. The strong couplings we identify support the AlphaFold model and suggest the deposited ensemble is missing conserved interactions that form in HSJ1a and other J domain proteins. This example illustrates how large-scale folding stability measurements can reveal the thermodynamic effects of a critical interaction even when that interaction is missing in the deposited NMR structure. Notably, AlphaFold itself does not always predict this network either: when we include disordered linkers from the NMR construct or used for cDNA proteolysis, AlphaFold also predicts alternative structures lacking the interaction network (Fig. S14D and E).

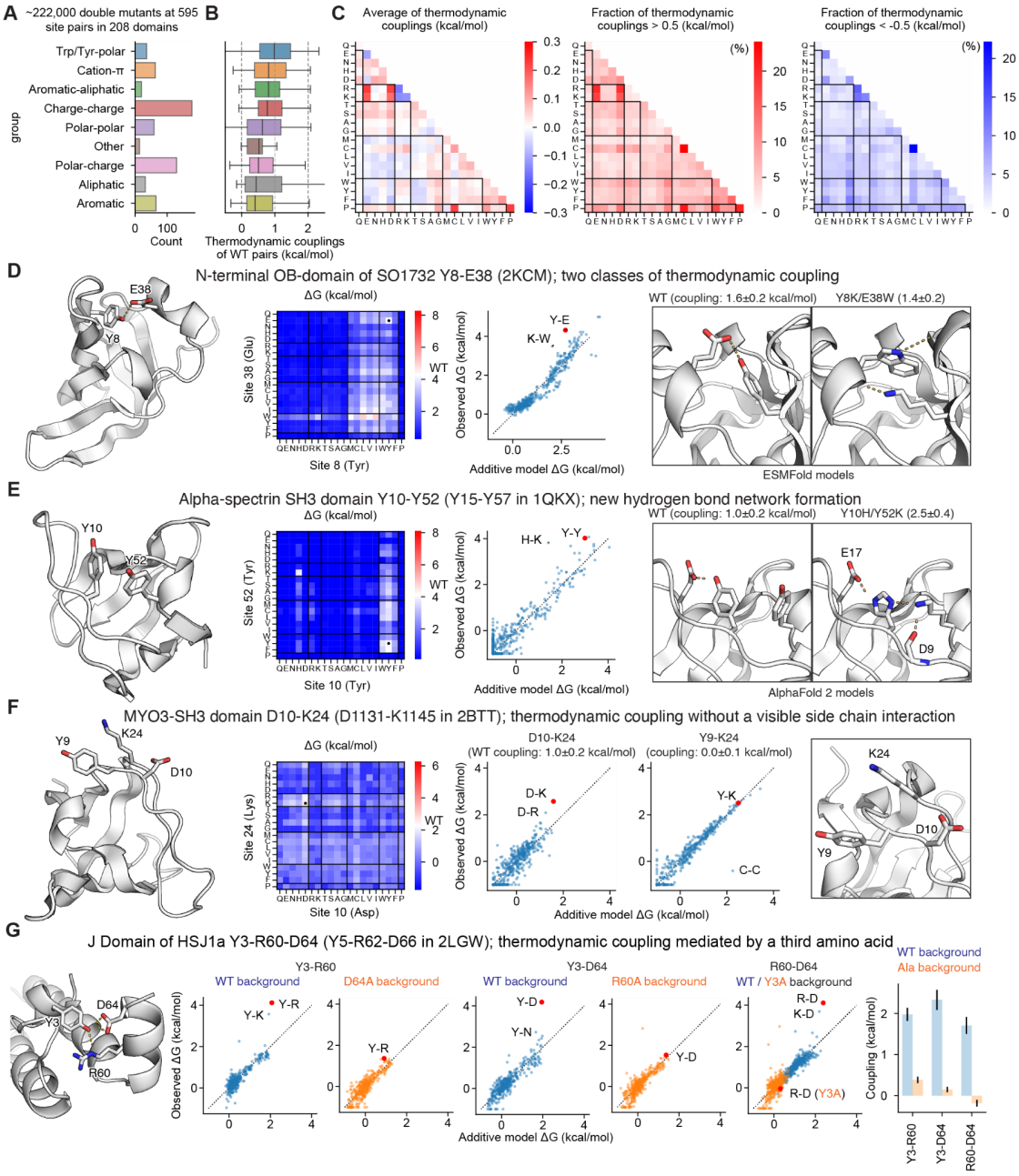


Fig. 4. Quantifying thermodynamic coupling between amino acid pairs

(A) Categorization of 595 pairs of amino acids selected for exhaustive double mutant analysis.

(B) Thermodynamic couplings of the wild-type amino acid pairs according to our additive model broken down by category.

(C) The average thermodynamic couplings (left) and the fraction of amino acid pairs with thermodynamic coupling > 0.5 kcal/mol (middle) and < -0.5 kcal/mol (right) for all amino acid combinations (wild-type and mutant).

(D and E) Analysis of thermodynamic coupling for two notable amino acid pairs. From left to right, we show the structure of each domain and the two positions that were mutated, the stabilities (ΔG) of all pairs of amino acids at those positions, the agreement between the stabilities from the additive model (x-axis) and the observed stabilities (y-axis) with the wild-type pair shown as a red dot, and AlphaFold or ESMFold models of amino acid pairs with strong thermodynamic couplings. Thermodynamic couplings show the observed stability minus the expected stability from the additive model; the uncertainties show the standard deviations from computing the couplings using bootstrapped samples of the 400 double mutants.

(F) Thermodynamic coupling without a visible side chain interaction. From left to right, the structure of the MYO3 SH3 domain and notable residues; the stabilities (ΔG) of all pairs of amino acids at D10 and K24; the stabilities of double mutants in the additive model (x-axis) and experimental data (y-axis); and the zoomed structure for D10, K23, and K24.

(G) Thermodynamic coupling mediated by a third amino acid. Exhaustive amino acid substitutions were performed for each pair of two out of the three amino acids. The same amino acid substitutions were also performed for the mutant background with the third amino acid replaced by Ala. From left to right, the AlphaFold-modeled structure of the J domain of HSJ1a with three interacting amino acids, the stabilities of double mutants in the additive model (x-axis) and experimental data (y-axis) in the wild-type background (blue) and Ala-replaced backgrounds (orange), and the thermodynamic coupling for each pair of wild-type amino acids in the wild-type background (blue) and the Ala-replaced backgrounds (orange). Substituting any of the three amino acids for Ala eliminates the thermodynamic coupling between the other two amino acids. Error bars show standard deviations from bootstrap resampling as before.

The scale of our cDNA display proteolysis experiments makes it straightforward to characterize unique cases like these, and again these cases will serve as stringent tests for models of folding stability. Strong third-order couplings like this example also present a special challenge for computational models that calculate stabilities by summing interaction energies between pairs of residues using a single reference structure. Deep learning models that implicitly represent entire conformational landscapes (Norn et al., 2021) may be more promising, but training these models using large-scale thermodynamic measurements will be essential to achieve their potential.

Natural sequences systematically deviate from their highest stability variants

How does selection for stability influence protein sequence evolution in concert with other evolutionary mechanisms? It is well known that proteins contain specific functional residues that are commonly deleterious to stability (Meiering et al., 1992; Shoichet et al., 1995). However, the challenge of measuring stability has made it difficult to experimentally distinguish selection for stability from other selective pressures on a global level (Cagiada et al., 2022; Høie et al., 2022; Tokuriki and Tawfik, 2009). To examine the strength of selection for stability, we created a simple classification model to predict the wild-type amino acid at any site in a natural protein based on the folding stabilities of all substitution variants at that site (excluding Cys) (Fig. 5A). The model contains two parts: (1) a shared weight function that converts absolute stabilities of protein variants into relative probabilities of those sequences, and (2) amino-acid specific offsets that shift amino acid probabilities by a constant amount at all sites. We fit the shared weight function parameters (a flexible monotonically increasing function) and the offsets together using absolute stability data for wild-type sequences and substitution variants at 4,718 sites in 80 non-redundant natural proteins (85,004 ΔG measurements in all, Fig. 5A). Our simple model fits the data well by three criteria: (1) it correctly produces the overall frequencies of the 19 (non-Cys) amino acids (Fig. 5B), (2) the output amino acid probabilities are correctly calibrated across the full range of probability (Fig. S15), and (3) the model performs similarly well on the training set and on a held-out testing set consisting of 621 sites in 11 domains with no similarity to the training set (Fig. 5E).

The model parameters reveal the strength of selection for stability across this heterogeneous set of domains from many organisms. Within the main range of our data (folding stabilities from 1.5 to 4 kcal/mol), amino acid probabilities increase approximately linearly with increased stability, with a 1 kcal/mol stability difference between protein variants indicating a ~5.7-fold difference in sequence likelihood (Fig. 5C). The global offsets to each amino acid's probability (Fig. 5D) are different from the empirical amino acid frequencies (Fig. 5B) and indicate the probability of each amino acid under conditions where all sequence variants are equally stable. The offsets span a 23-fold range: the most likely amino acid (Glu) is 23-fold more likely to occur ($2^{1.5}/2^{-3.0}$) than the least likely amino acid (Trp) under the conditions that sequence variants containing these amino acids at the same site are equally stable (Fig. 5D). This probability difference corresponds to a stability difference of ~1.8 kcal/mol (Fig. 5C); i.e. Trp and Glu would be equally likely at a site if the Trp variant were 1.8 kcal/mol more stable than the Glu variant. Overall, the most likely amino acids are the charged amino acids Glu, Asp, and Lys, suggesting selection for solubility, whereas the least likely amino acids are the nonpolar aromatic amino acids Trp, Phe, and Tyr, along with Met. These offsets provide a quantitative "favorability" metric incorporating all non-stability evolutionary influences on amino acid composition, including selection for amino acid synthesis cost (Akashi and Gojobori, 2002; Chen and Nielsen, 2022), codon usage (Cope and Gilchrist, 2022; Shah and Gilchrist, 2011), avoiding oxidation-prone amino acid(s), net charge, and function. These offsets also highlight that biophysical models and protein design methods trained to reproduce native protein sequences

will not consistently optimize folding stability; Fig. 5D quantifies how much specific amino acids are over- or underrepresented in small, naturally occurring domains compared to their effects on stability. Notably, these offsets are similar to findings from an independent analysis of global discrepancies between variant effect data and sequence likelihood modeling (Riesselman et al., 2018)

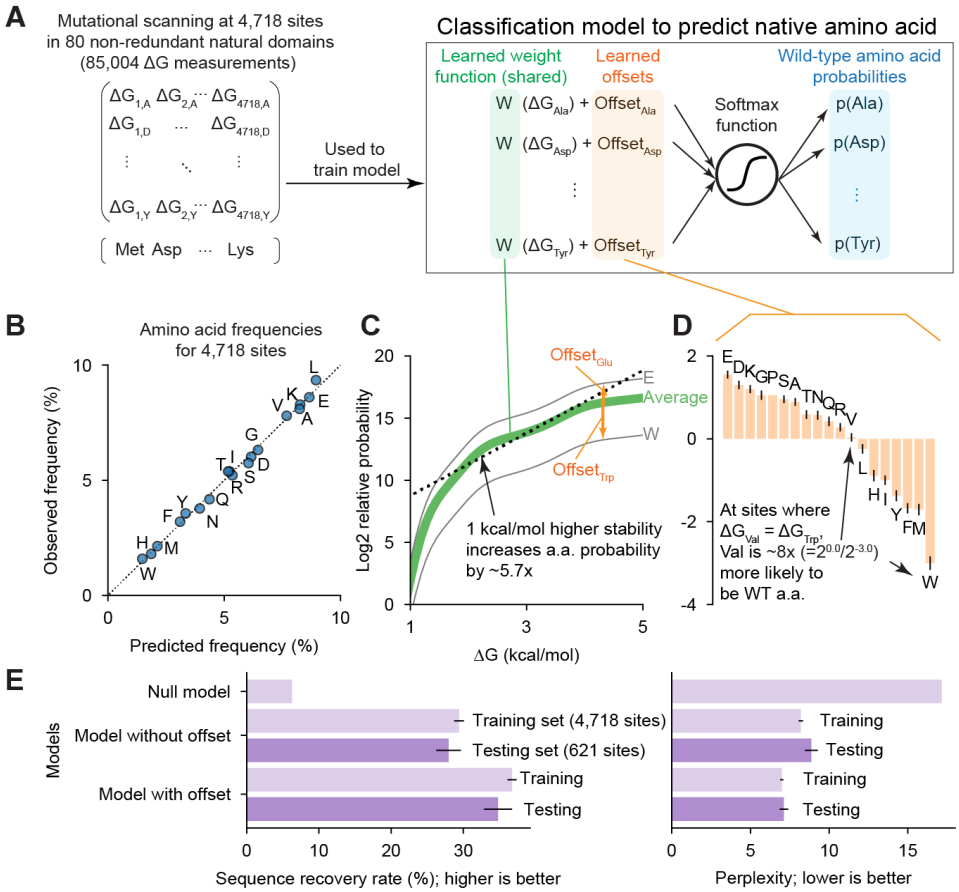


Fig. 5. Amino acid usage in natural proteins systematically deviates from maximizing stability.
(A) Classifier model for predicting wild-type amino acids based on the folding stabilities (ΔG) of each possible protein variant. A shared weighting function converts the stabilities of protein variants containing each amino acid into relative probabilities of those amino acids (green). The relative probability of each amino acid is further modified by a constant offset that is unique for each amino acid (orange).
(B) Predicted and observed amino acid frequencies according to the classifier model after fitting.
(C) The weighting function from the classifier model after fitting (green). Gray lines show the weight function after amino acid-specific offsets for Glu and Trp. In the region between 1.5 and 4 kcal/mol, the function has an approximately constant slope where a 1 kcal/mol increase in stability leads to a 5.7-fold increase in amino acid probability.
(D) Relative offsets for 19 amino acids from the classifier model after fitting. Error bars show the standard deviation of the model posterior.
(E) The sequence recovery rate (left) and perplexity (right) for predicting the wild-type amino acid using several models: an null model that ignores stability and always predicts amino acids at their observed frequencies, our classifier model without amino acid-specific offsets, and our full classifier model. Similar performance of the classifier model on a training set of 4,718 positions (light purple) and a testing set of 621 positions (dark purple) indicates that the model is not overfit. Error bars show standard deviations from bootstrap resampling of the sites in the training set and the testing set.

Properties of functional residues across diverse domains

Selection for function also causes protein sequences to diverge from the highest stability sequence variants. Previous studies (Cagiada et al., 2022; Høie et al., 2022) have applied this strategy to identify functional sites based on the difference between evolutionary conservation and predicted effects on stability. We expanded this strategy to employ experimental stability measurements and examined the properties of functional sites on a large scale. We identified functional sites in 92 diverse protein domains by comparing each site's average ΔΔG of substitutions with its normalized GEMME score, a measure of sequence conservation (Fig. 6A). Sites where wild-type amino acids are critical for stability (higher average ΔΔG, rightward) tend to be more evolutionarily conserved (upward) and vice versa. We defined all sites in the upper

left region (where the wild-type amino acid is conserved yet unimportant for stability, 9.3% in total) to be "functional" sites. This classification correctly identifies key binding residues in the chromodomain of HP1 and the SH3 domain of BBC1 (Fig. 6C, see Fig S16 for mutational scanning and conservation data on these examples). We found that Gly, Asp, and the bulky amino acids Trp Arg, and Tyr were frequently classified as functional (Fig. 6D). However, like previous studies, our classification method has the notable weakness that any site that is important for folding stability will not be considered functional.

Across all 92 domains, the fraction of functional sites ranged from 0 to ~25% (Fig. 6E). The domains with the highest fraction of functional sites (the Sso7d protein (1JIC) and Ribosomal protein S19 (1QKH)) are both nucleic acid binding proteins, with the functional sites located on the surface primarily at the binding interface (Fig. 6F). To identify buried functional sites, we compared each site's conservation (normalized GEMME score) to the average ΔΔG of nonpolar substitutions (Fig. 6G), a more permissive metric. With this approach, most functional sites are still located at the protein surface, but a small fraction are located in the core (Fig. 6H). One example is A64 in the DUF1471 domain of yahO. A64 is highly conserved and buried in the core of the domain, but substitutions to Tyr or Phe increase folding stability (Fig. 6I). This indicates that A64 modulates the function of the domain even without interacting with external partner molecules, perhaps by maintaining the overall protein shape. Similarly, in the N-terminal domain of FK506-binding protein 3, L55 is buried in the core and highly conserved even though substitutions to Ile, Val, or Phe have no effect on stability (Fig. 6J). This domain binds DNA and the other functional residues are mainly located at the binding interface. Although L55 does not directly interact with DNA, substitutions to other hydrophobic amino acids may change the orientations of the surface side chains and prevent proper DNA binding. Notably, chemical shift perturbations in this domain indicate which residues change their magnetic environment in response to DNA binding (Fig. 6J) (Prakash et al., 2016). Chemical shift perturbations are found mainly in the functional residues on the protein surface, but L55 experiences a chemical shift perturbation as well, indicating allosteric communication between the functional surface residues and L55. These results highlight unusual cases where buried sites are conserved due to specific functional requirements rather than to maintain overall stability.

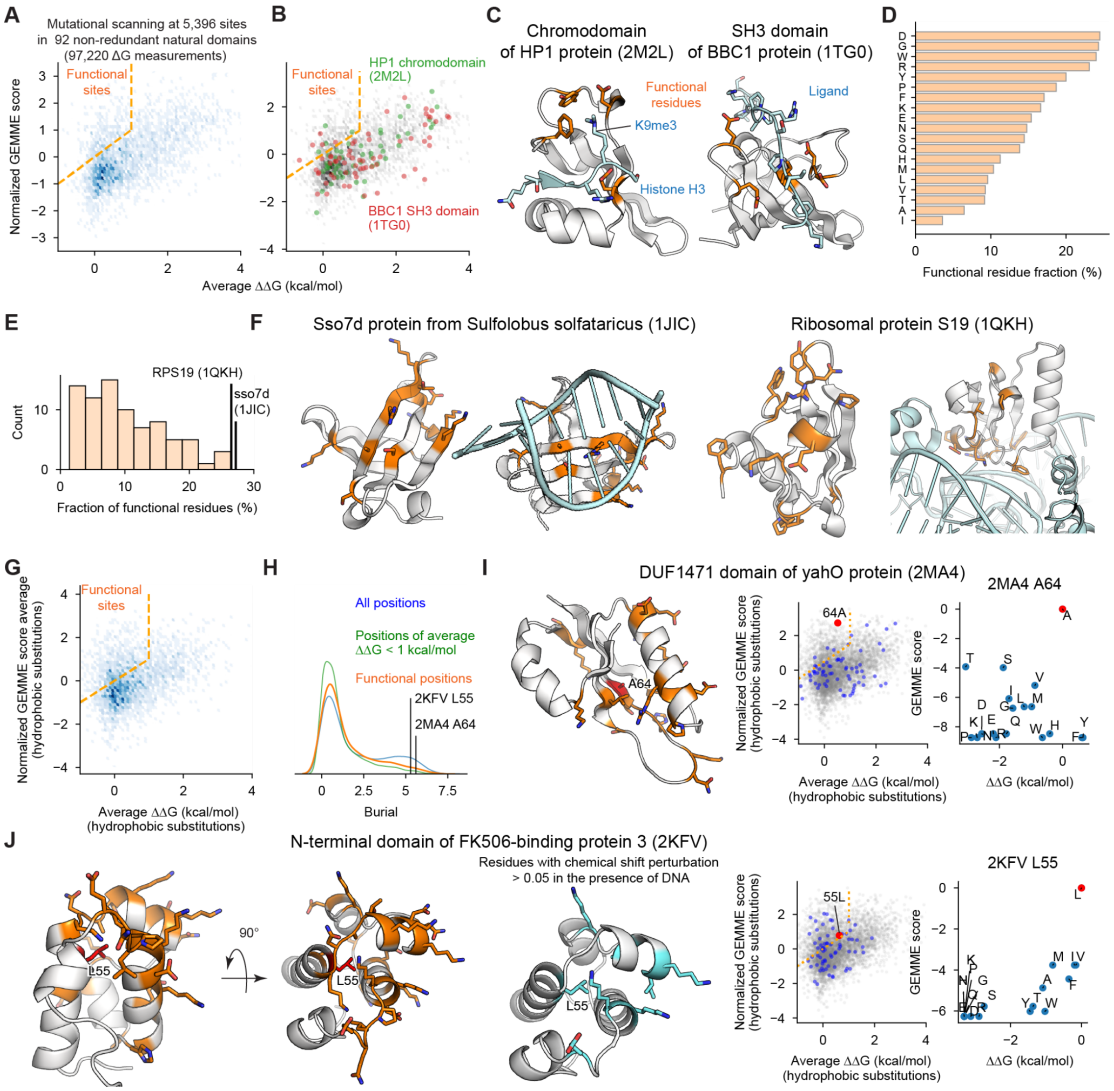


Fig. 6. Properties of functional sites across diverse domains

(A) The relationship between wild-type stability (average ΔΔG for substitutions) and evolutionary conservation (normalized GEMME score). All sites above the orange dashed line are highly conserved but unimportant for stability; we define these as "functional sites".

(B) As in (A), highlighting positions in the HP1 chromo domain (2M2L; green) and the BBC1 SH3 domain (1TG0; red).

(C) Structures of HP1 chromo domain and BBC1 SH3 domain (gray) and their ligands (light blue). Functional sites are shown in orange. Ligand positions were modeled based on PDB structures 1KNA (for HP1) and 2LCS (for the SH3 domain).

(D) Amino acids are ranked by the percentage of positions where that wild-type amino acid is classified as functional, for 5,396 positions in 92 non-redundant natural domains.

(E) The percentage of functional residues in each of the 92 non-redundant domains.

(F) Structures of the two domains with the highest percentages of functional residues. Nucleic acids interacting with each of the structures are shown in light blue and functional residues are shown in orange. The Sso7d-DNA complex is the crystal structure 1BNZ; the S19-RNA complex is modeled based on the 4V5Y structure.

(G) As in (A), except only considering nonpolar substitutions for calculating ΔΔG.

(H) The distributions of burial (side chain contacts) for all sites (blue), sites where the wild-type amino acid is unimportant for stability (average ΔΔG < 1 kcal/mol) (green), and functional sites (orange). Functional sites are generally located on the surface of the protein. Two unusual buried functional residues are highlighted.

(I) Structure of the DUF1471 domain of yahO (2MA4) with functional sites in orange and the unusual buried functional site A64 in red. Ala64 is highly conserved yet the domain is stabilized by substitutions to Tyr or Phe (positive ΔΔG, x-axis). However, Tyr and Phe are rarely found in evolution (low GEMME score, y-axis).

(J) Left: Structure of the N-terminal domain of FK506-binding protein 3 (2KFV) with functional sites in orange and the unusual buried functional site L55 (L78 in PDB numbering) in red. Middle: Residues with chemical shift perturbations in response to DNA binding (Prakash et al., 2016); L55 shows a perturbation despite not contacting DNA. Right: L55 is conserved (high GEMME score, y-axis) but relatively unimportant for stability (low average ΔΔG, x-axis). Substitution to Phe, Val, or Ile is thermodynamically neutral (ΔΔG near zero) but these amino acids are rarely found in evolution (low GEMME score).

Large-scale stability analysis to characterize unique designs, identify stabilizing mutations, and evaluate design methods

The unique scale of cDNA display proteolysis creates new opportunities for improving protein design. Here, we examine three applications of our method and massive dataset: (1) characterizing the stability determinants of rare, highly polar proteins, (2) identifying stabilizing mutations, and (3) benchmarking the protein design tool PROSS (Goldenzweig et al., 2016). The hydrophobic effect is considered the dominant force in protein folding (Dill, 1990), and measuring stability for thousands of our previously-designed domains (Rocklin et al., 2017) by cDNA display proteolysis revealed a general trend of increasing stability with increasing hydrophobicity (Fig. 7A). However, increased hydrophobicity can promote protein aggregation, non-specific interactions, and low expression yield. To study the properties of high stability, low hydrophobicity proteins, we examined hundreds of designed proteins by deep mutational scanning across a wide range of hydrophobicity and stability. Although the mutational scanning patterns for low hydrophobicity proteins were not obviously different from other designs, we identified several designs that possessed exceptionally strong polar interactions (large dots in Fig. 7A). In Fig. 7B, we highlight stabilizing polar networks and a cation-π interaction in these unusual designs (see Fig. S17 for full mutational scanning results). The average ΔΔG for substitutions at these polar sites ranges from -0.20 to -1.33 kcal/mol, corresponding to the top 63 to 1.5%ile for all 3,694 polar sites in 145 designs. Our unusually massive dataset made it possible to identify these rare highly stabilizing interactions. Notably, the second hydrogen bond network in EHEE_rd2_0152 is also found in two other more hydrophobic designs. However, the network is less sensitive to substitution in those designs, highlighting how the overall protein environment mediates the effects of substitutions even on the protein surface (Fig. S18).

We next examined how our approach could be used to identify stabilizing mutations. Predicting and designing stabilizing mutations is a major goal of protein modeling, but prediction accuracy remains low (Broom et al., 2020). In part, this is because stabilizing mutants are rare in current databases (Nikam et al., 2021; Xavier et al., 2021) (outside of reverting a destabilizing mutant), limiting the data available for improving modeling. In contrast, our large-scale approach revealed 2,600 stabilizing mutations, defined as mutations that increase folding stability by at least 1 kcal/mol. The overall fraction of stabilizing mutations was 0.06% to 0.6% for different protein types (Fig. 7C). Stabilizing mutations were enriched at functional sites (23% of the stabilizing mutations from 7.5% of sites classified as functional), but these were still a small fraction of the total. Notably, our set includes 112 examples of stabilizing insertions and deletions which are nearly absent from current databases. In Fig. 7D, we show three examples of different classes of stabilizing mutations found in our dataset with effects ranging from +1.2 to +3.1 kcal/mol (Fig. S19).

Finally, we applied our method to evaluate PROSS (Goldenzweig et al., 2016), an automated method for enhancing folding stability within sequence constraints inferred from a multiple sequence alignment. We tested 1,156 PROSS designs for 266 protein domains (a 10-100x increase over previous benchmarking study (Peleg et al., 2021)). Unlike previous studies, our mutational scanning data for all 266 wild-type domains enabled us to examine the isolated effect of every individual substitution in every PROSS

design. The average increase in stability from PROSS was 0.6±1.0 (mean±std) kcal/mol, and 40% of 727 domains (with wild-type ΔG < 4 kcal/mol) had at least one design with a 1 kcal/mol increase in stability (Fig. 7E). As expected, PROSS avoided mutations at functional positions: only 1.9% of PROSS-designed mutations were found at functional positions compared to 8.7% of sites classified as functional (defined in Fig. 6A). Three examples of domains successfully stabilized by PROSS are shown in Fig. 7F. Although the median number of designed mutations was only 4, more mutations typically led to a larger increase in stability (Fig. S20A), as theorized previously (Broom et al., 2020). Based on our mutational scanning data, the average effect of an individual PROSS mutation was 0.22±0.47 kcal/mol (Green line Fig. S20B). On average, the added stabilization from PROSS is comparable in size to the effect of the best single mutant designed by PROSS, and smaller than the additive effect of the two best designed mutations (Fig. S20C). Evaluating individual mutations recommended by PROSS (or other design tools) by direct comparisons to mutational scanning data provides a novel approach for systematically improving these design methods.

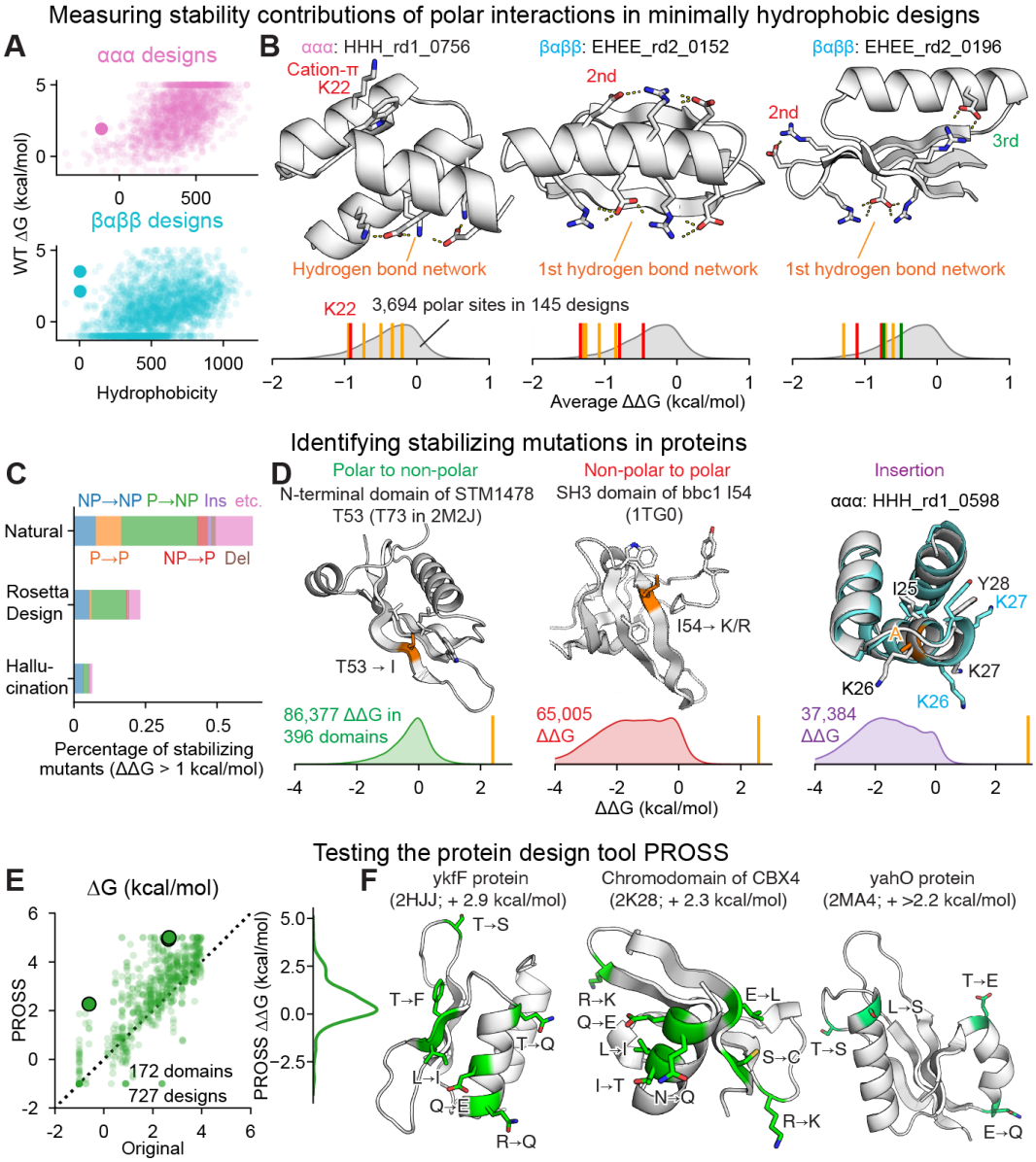


Fig. 7. Application of screening data to protein design

(A) Relationship between hydrophobicity (calculated based on the previous report (Monera et al., 1995)) and folding stability (ΔG) for designed proteins (Rocklin et al., 2017). Examples from (B) are shown as large dots.

(B) For three proteins with high folding stability and low hydrophobicity, we highlight critical hydrophilic interactions stabilizing these proteins. The gray density plots represent the average ΔΔG of substitutions at 3,694 polar sites in 145 designed domains. The colored vertical bars indicate the values for the highlighted positions. These three proteins feature polar amino acids where the average ΔΔG of substitutions is unusually destabilizing

(> top 5%ile). For HHH_rd1_0756, K22 is shown as a red line; the interacting W32 is considered nonpolar and not shown. Full mutational scanning results are shown in Fig. S17. All three structures are design models reported previously (Rocklin et al., 2017), not experimental structures.

(C) Fraction of stabilizing mutations (ΔΔG > 1 kcal/mol) found in natural domains, Rosetta designs, and hallucination designs, broken down by mutation type. NP: non-polar, P: polar, Ins: insertion, Del: deletion.

(D) Three examples of stabilizing mutations identified by our assay, along with the distribution of ΔΔG values for these three mutation types. The highlighted mutations are indicated by vertical bars on the density plots. Full mutational scanning results are shown in Fig. S19. The structure of HHH_rd1_0598 is a design model reported previously (Rocklin et al., 2017), not an experimental structure.

(E) Left: Testing the protein design tool PROSS (Goldenzweig et al., 2016). Each point shows the stability of one domain before (x-axis) and after (y-axis) redesign by PROSS. The dashed black line represents Y=X. Examples from (F) are shown as large dots. Right: Distribution of ΔG change from PROSS redesign. 40% of domains are stabilized by > 1 kcal/mol by the tool. Note that we only show 727 designs with wild-type ΔG < 4 kcal/mol.

(F) Examples of domains stabilized by PROSS. Amino acids mutated by PROSS are shown in green on the Alphafold-generated structural model.

Discussion

The cDNA display proteolysis method massively expands the scale of folding stability experiments. Still, the method currently has notable limitations. Because we digest proteins under native conditions, our inferred thermodynamic stabilities are only accurate when (1) folding is fully cooperative (no unfolded segments get cleaved without global unfolding (Park and Marqusee, 2004)), (2) folding is at equilibrium during the assay (no kinetic stability or spurious stability due to aggregation), (3) $K_{50,U}$ is accurately inferred (Fig. 1C), and (4) cleavage rates fall within the measurable range of the assay, which currently limits the dynamic range to ~5 kcal/mol (Fig. 1C). Many domains - particularly larger protein structures - will not satisfy these conditions, and issues such as non-cooperativity, kinetic stability, or aggregation are invisible in a single measurement. Combining cDNA display proteolysis with chemical denaturation (pulse proteolysis, (Park and Marqusee, 2005)) may overcome these obstacles and enable mega-scale analysis of less cooperative and/or higher stability proteins, while also avoiding the need to infer $K_{50,U}$. Advances in DNA synthesis (including methods like DropSynth (Plesa et al., 2018; Sidore et al., 2020)) will also make it possible to expand cDNA display proteolysis to analyze diverse libraries of larger domains. Lastly, multiplexed measurements and automated data processing always have the potential to introduce inaccuracies, although we worked to exclude unreliable data. For notable individual results, examining the raw data can be helpful, and we included all data and code to regenerate all fits.

Despite these limitations, the unique scale of cDNA display proteolysis opens completely new possibilities for studying protein stability. By comprehensively measuring single mutants across nearly all small structures in the Protein Data Bank, we quantified several global trends: trends in amino acid fitness at different sites, trends in the effects of single and double mutants, and trends in how stability influences sequence evolution. Along with these global trends, our large-scale analysis also uncovered hundreds of exceptional cases that would be challenging to identify by smaller-scale methods. These include mutations with extreme effects, sites with unusual stability landscapes, and pair interactions with unusually strong thermodynamic couplings. The strong thermodynamic couplings we identified in the J domain of human HSJ1a (Fig. 4G) - missing in the deposited NMR structure - highlight how large-scale stability assays can complement other methods for revealing structural details in solution. The 2,400 double mutants examined in that domain made up only 0.3% of the experimental library. Beyond studying the origins of stability, cDNA display proteolysis will have a range of other applications, including assaying designed proteins on a massive scale to systematically improve design methods (Basanta et al., 2020; Kim et al., 2022; Rocklin et al., 2017), identifying folded domains in metagenomic sequences, and dissecting the relationships between folding stability and function (Dou et al., 2018).

Achieving an accurate, quantitative understanding of protein stability and its sequence dependence has been a central goal in biophysics for decades. We envision millions of cDNA display proteolysis measurements forming the foundation for a new generation of deep learning models predicting absolute folding stabilities and effects of mutations. Breakthroughs in deep learning-powered structure prediction have proven the power of these models in protein science, but collecting sufficient thermodynamic data has always been a major obstacle. Due to the scale and efficiency of cDNA display proteolysis, the main limit to measuring stability for millions of small domains is the cost of DNA synthesis (Kosuri and Church, 2014; Kuiper et al., 2022; Sun et al., 2016) and sequencing (Foox et al., 2021; Goodwin et al., 2016) - both of which are rapidly decreasing (Hughes and Ellington, 2017; Levy and Boone, 2019; Pervez et al., 2022; Song et al., 2021). With the flexibility of DNA oligo synthesis, cDNA display proteolysis can assay massive mutational libraries (as shown here) as well as massive libraries of unrelated sequences and structures, which will add essential diversity in training datasets. The size and diversity of protein sequence space creates enormous challenges for biology and protein design. The cDNA display proteolysis method offers a powerful approach to map folding stability across this space on an unprecedented scale.

References and Notes

Agozzino, L., and Dill, K.A. (2018). Protein evolution speed depends on its stability and abundance and on chaperone concentrations. Proc. Natl. Acad. Sci. U. S. A. *115*, 9092–9097. https://doi.org/10.1073/pnas.1810194115.

Akashi, H., and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc. Natl. Acad. Sci. U. S. A. *99*, 3695–3700. https://doi.org/10.1073/pnas.062526999.

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods *16*, 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.

Anishchenko, I., Pellock, S.J., Chidyausiku, T.M., Ramelot, T.A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A.K., et al. (2021). De novo protein design by deep network hallucination. Nature *600*, 547–552. https://doi.org/10.1038/s41586-021-04184-w.

Arai, H., Kumachi, S., and Nemoto, N. (2020). cDNA Display: A Stable and Simple Genotype-Phenotype Coupling Using a Cell-Free Translation System. Methods Mol. Biol. *2070*, 43–56. https://doi.org/10.1007/978-1-4939-9853-1_3.

Araya, C.L., Fowler, D.M., Chen, W., Muniez, I., Kelly, J.W., and Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc. Natl. Acad. Sci. U. S. A. *109*, 16858–16863. https://doi.org/10.1073/pnas.1209751109.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.

Basanta, B., Bick, M.J., Bera, A.K., Norn, C., Chow, C.M., Carter, L.P., Goreshnik, I., Dimaio, F., and Baker, D. (2020). An enumerative algorithm for de novo design of proteins with diverse pocket structures. Proc. Natl. Acad. Sci. U. S. A. *117*, 22135–22145. https://doi.org/10.1073/pnas.2005412117.

Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M., and Church, G.M. (2021). Low-N protein engineering with data-efficient deep learning. Nat. Methods *18*, 389–396. https://doi.org/10.1038/s41592-021-01100-y.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. U. S. A. *103*, 5869–5874. https://doi.org/10.1073/pnas.0510098103.

Broom, A., Trainor, K., Jacobi, Z., and Meiering, E.M. (2020). Computational Modeling of Protein Stability: Quantitative Analysis Reveals Solutions to Pervasive Problems. Structure https://doi.org/10.1016/j.str.2020.04.003.

Cagiada, M., Bottaro, S., Lindemose, S., Schenstrøm, S.M., Stein, A., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2022). Discovering functionally important sites in proteins.

Chen, Y., and Nielsen, J. (2022). Yeast has evolved to minimize protein resource cost for synthesizing amino acids. Proc. Natl. Acad. Sci. U. S. A. *119*. https://doi.org/10.1073/pnas.2114622119.

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G.M., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. Nat. Biotechnol. https://doi.org/10.1038/s41587-022-01432-w.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

Coey, C.T., and Drohat, A.C. (2017). Kinetic Methods for Studying DNA Glycosylases Functioning in Base Excision Repair. Methods Enzymol. *592*, 357–376. https://doi.org/10.1016/bs.mie.2017.03.016.

Cope, A.L., and Gilchrist, M.A. (2022). Quantifying shifts in natural selection on codon usage between protein regions: a population genetics approach. BMC Genomics *23*, 408. https://doi.org/10.1186/s12864-022-08635-0.

Dill, K.A. (1990). Dominant forces in protein folding. Biochemistry *29*, 7133–7155. https://doi.org/10.1021/bi00483a001.

Dodson, C.A., and Arbely, E. (2015). Protein folding of the SAP domain, a naturally occurring two-helix bundle. FEBS Lett. *589*, 1740–1747. https://doi.org/10.1016/j.febslet.2015.06.002.

Dou, J., Vorobieva, A.A., Sheffler, W., Doyle, L.A., Park, H., Bick, M.J., Mao, B., Foight, G.W., Lee, M.Y., Gagnon, L.A., et al. (2018). De novo design of a fluorescence-activating β-barrel. Nature *561*, 485–491. https://doi.org/10.1038/s41586-018-0509-0.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352. https://doi.org/10.1016/j.cell.2008.05.042.

Foox, J., Tighe, S.W., Nicolet, C.M., Zook, J.M., Byrska-Bishop, M., Clarke, W.E., Khayat, M.M., Mahmoud, M., Laaguiby, P.K., Herbert, Z.T., et al. (2021). Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. Nat. Biotechnol. *39*, 1129–1140. https://doi.org/10.1038/s41587-021-01049-5.

Goldenzweig, A., and Fleishman, S.J. (2018). Principles of Protein Stability and Their Application in Computational Design. Annu. Rev. Biochem. *87*, 105–129. https://doi.org/10.1146/annurev-biochem-062917-012102.

Goldenzweig, A., Goldsmith, M., Hill, S.E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., et al. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. Mol. Cell *63*, 337–346. https://doi.org/10.1016/j.molcel.2016.06.012.

Golinski, A.W., Holec, P.V., Mischler, K.M., and Hackel, B.J. (2019). Biophysical Characterization Platform Informs Protein Scaffold Evolvability. ACS Comb. Sci. *21*, 323–335.

https://doi.org/10.1021/acscombsci.8b00182.

Golinski, A.W., Mischler, K.M., Laxminarayan, S., Neurock, N.L., Fossing, M., Pichman, H., Martiniani, S., and Hackel, B.J. (2021). High-throughput developability assays enable library-scale identification of producible protein scaffold variants. Proc. Natl. Acad. Sci. U. S. A. *118*. https://doi.org/10.1073/pnas.2026658118.

Gong, L.I., Suchard, M.A., and Bloom, J.D. (2013). Stability-mediated epistasis constrains the evolution of an influenza protein. Elife *2*, e00631. https://doi.org/10.7554/eLife.00631.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351. https://doi.org/10.1038/nrg.2016.49.

Hamelryck, T., and Manderick, B. (2003). PDB file parser and structure class implemented in Python. Bioinformatics *19*, 2308–2310. https://doi.org/10.1093/bioinformatics/btg299.

Høie, M.H., Cagiada, M., Beck Frederiksen, A.H., Stein, A., and Lindorff-Larsen, K. (2022). Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. Cell Rep. *38*, 110207. https://doi.org/10.1016/j.celrep.2021.110207.

Hoover, D.M., and Lubkowski, J. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res. *30*, e43. https://doi.org/10.1093/nar/30.10.e43.

Huang, P.-S., Ban, Y.-E.A., Richter, F., Andre, I., Vernon, R., Schief, W.R., and Baker, D. (2011). RosettaRemodel: a generalized framework for flexible backbone protein design. PLoS One *6*, e24109. https://doi.org/10.1371/journal.pone.0024109.

Hughes, R.A., and Ellington, A.D. (2017). Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. Cold Spring Harb. Perspect. Biol. *9*. https://doi.org/10.1101/cshperspect.a023812.

Ingraham, Garg, Barzilay, and Jaakkola Generative models for graph-based protein design. Adv. Neural Inf. Process. Syst.

Jäger, M., Dendle, M., and Kelly, J.W. (2009). Sequence determinants of thermodynamic stability in a WW domain--an all-beta-sheet protein. Protein Sci. *18*, 1806–1813. https://doi.org/10.1002/pro.172.

Jarzab, A., Kurzawa, N., Hopf, T., Moerch, M., Zecha, J., Leijten, N., Bian, Y., Musiol, E., Maschberger, M., Stoehr, G., et al. (2020). Meltome atlas-thermal proteome stability across the tree of life. Nat. Methods *17*, 495–503. https://doi.org/10.1038/s41592-020-0801-4.

Jemth, P., Day, R., Gianni, S., Khan, F., Allen, M., Daggett, V., and Fersht, A.R. (2005). The structure of the major transition state for folding of an FF domain from experiment and simulation. J. Mol. Biol. *350*, 363–378. https://doi.org/10.1016/j.jmb.2005.04.067.

Jiang, X., Kowalski, J., and Kelly, J.W. (2001). Increasing protein stability using a rational approach combining sequence homology and structural alignment: Stabilizing the WW domain. Protein Sci. *10*, 1454–1465. https://doi.org/10.1110/ps.640101.

Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C., and Vriend, G. (2011). A series of PDB related databases for everyday needs. Nucleic Acids Res. *39*, D411–D419. https://doi.org/10.1093/nar/gkq1105.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. https://doi.org/10.1002/bip.360221211.

Kim, T.-E., Tsuboyama, K., Houliston, S., Martell, C.M., Phoumyvong, C.M., Haddox, H.K., Arrowsmith, C.H., and Rocklin, G.J. (2021). Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation.

Kim, T.-E., Tsuboyama, K., Houliston, S., Martell, C.M., Phoumyvong, C.M., Lemak, A., Haddox, H.K., Arrowsmith, C.H., and Rocklin, G.J. (2022). Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation. Proc. Natl. Acad. Sci. U. S. A. *119*, e2122676119. https://doi.org/10.1073/pnas.2122676119.

Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T.B., Montelione, G.T., and Baker, D. (2012). Principles for designing ideal protein structures. Nature *491*, 222–227. https://doi.org/10.1038/nature11600.

Kosuri, S., and Church, G.M. (2014). Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods *11*, 499–507. https://doi.org/10.1038/nmeth.2918.

Kuiper, B.P., Prins, R.C., and Billerbeck, S. (2022). Oligo Pools as an Affordable Source of Synthetic DNA for Cost-Effective Library Construction in Protein- and Metabolic Pathway Engineering. Chembiochem *23*, e202100507. https://doi.org/10.1002/cbic.202100507.

Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. (2015). MAESTRO--multi agent stability prediction upon point mutations. BMC Bioinformatics *16*, 116. https://doi.org/10.1186/s12859-015-0548-6.

Levy, S.E., and Boone, B.E. (2019). Next-Generation Sequencing Strategies. Cold Spring Harb. Perspect. Med. *9*. https://doi.org/10.1101/cshperspect.a025791.

Maguire, J.B., Haddox, H.K., Strickland, D., Halabiya, S.F., Coventry, B., Griffin, J.R., Pulavarti, S.V.S.R.K., Cummins, M., Thieker, D.F., Klavins, E., et al. (2021). Perturbing the energy landscape for improved packing during computational protein design. Proteins *89*, 436–449. https://doi.org/10.1002/prot.26030.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

Maxwell, K.L., and Davidson, A.R. (1998). Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects. Biochemistry *37*, 16172–16182. https://doi.org/10.1021/bi981788p.

Meiering, E.M., Serrano, L., and Fersht, A.R. (1992). Effect of active site residues in barnase on activity and stability. J. Mol. Biol. *225*,

585–589. https://doi.org/10.1016/0022-2836(92)90387-y.

Monera, O.D., Sereda, T.J., Zhou, N.E., Kay, C.M., and Hodges, R.S. (1995). Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. J. Pept. Sci. *1*, 319–329. https://doi.org/10.1002/psc.310010507.

Nemoto, N., Miyamoto-Sato, E., Husimi, Y., and Yanagawa, H. (1997). In vitro virus: bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. FEBS Lett. *414*, 405–408. https://doi.org/10.1016/s0014-5793(97)01026-0.

Neuweiler, H., Sharpe, T.D., Rutherford, T.J., Johnson, C.M., Allen, M.D., Ferguson, N., and Fersht, A.R. (2009). The folding mechanism of BBL: Plasticity of transition-state structure observed within an ultrafast folding protein family. J. Mol. Biol. *390*, 1060–1073. https://doi.org/10.1016/j.jmb.2009.05.011.

Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D., and Gromiha, M.M. (2021). ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res. *49*, D420–D424. https://doi.org/10.1093/nar/gkaa1035.

Nisthal, A., Wang, C.Y., Ary, M.L., and Mayo, S.L. (2019). Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. Proc. Natl. Acad. Sci. U. S. A. *116*, 16367–16377. https://doi.org/10.1073/pnas.1903888116.

Norn, C., Wicky, B.I.M., Juergens, D., Liu, S., Kim, D., Tischer, D., Koepnick, B., Anishchenko, I., Foldit Players, Baker, D., et al. (2021). Protein sequence design by conformational landscape optimization. Proc. Natl. Acad. Sci. U. S. A. *118*. https://doi.org/10.1073/pnas.2017228118.

Northey, J.G.B., Maxwell, K.L., and Davidson, A.R. (2002). Protein folding kinetics beyond the phi value: using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state. J. Mol. Biol. *320*, 389–402. https://doi.org/10.1016/S0022-2836(02)00445-X.

Pak, M.A., Markhieva, K.A., Novikova, M.S., Petrov, D.S., Vorobyev, I.S., Maksimova, E.S., Kondrashov, F.A., and Ivankov, D.N. (2021). Using AlphaFold to predict the impact of single mutations on protein stability and function.

Pan, Y., Cheng, K., Mao, J., Liu, F., Liu, J., Ye, M., and Zou, H. (2014). Quantitative proteomics reveals the kinetics of trypsin-catalyzed protein digestion. Anal. Bioanal. Chem. *406*, 6247–6256. https://doi.org/10.1007/s00216-014-8071-6.

Park, C., and Marqusee, S. (2004). Probing the high energy states in proteins by proteolysis. J. Mol. Biol. *343*, 1467–1476. https://doi.org/10.1016/j.jmb.2004.08.085.

Park, C., and Marqusee, S. (2005). Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding. Nat. Methods *2*, 207–212. https://doi.org/10.1038/nmeth740.

Park, C., Zhou, S., Gilmore, J., and Marqusee, S. (2007). Energetics-based protein profiling on a proteomic scale: identification of proteins resistant to proteolysis. J. Mol. Biol. *368*, 1426–1437. https://doi.org/10.1016/j.jmb.2007.02.091.

Peleg, Y., Vincentelli, R., Collins, B.M., Chen, K.-E., Livingstone, E.K., Weeratunga, S., Leneva, N., Guo, Q., Remans, K., Perez, K., et al. (2021). Community-Wide Experimental Evaluation of the PROSS Stability-Design Method. J. Mol. Biol. *433*, 166964. https://doi.org/10.1016/j.jmb.2021.166964.

Pervez, M.T., Hasnain, M.J.U., Abbas, S.H., Moustafa, M.F., Aslam, N., and Shah, S.S.M. (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. Biomed Res. Int. *2022*, 3457806. https://doi.org/10.1155/2022/3457806.

Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro.

Plesa, C., Sidore, A.M., Lubock, N.B., Zhang, D., and Kosuri, S. (2018). Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. Science *359*, 343–347. https://doi.org/10.1126/science.aao5167.

Prakash, A., Shin, J., Rajan, S., and Yoon, H.S. (2016). Structural basis of nucleic acid recognition by FK506-binding protein 25 (FKBP25), a nuclear immunophilin. Nucleic Acids Res. *44*, 2909–2925. https://doi.org/10.1093/nar/gkw001.

Pucci, F., Schwersensky, M., and Rooman, M. (2022). Artificial intelligence challenges for predicting the impact of mutations on protein stability. Curr. Opin. Struct. Biol. *72*, 161–168. https://doi.org/10.1016/j.sbi.2021.11.001.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating Protein Transfer Learning with TAPE. In Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, eds. (Curran Associates, Inc.), pp. 9689–9701.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods *15*, 816–822. https://doi.org/10.1038/s41592-018-0138-4.

de los Rios, M.A., Daneshi, M., and Plaxco, K.W. (2005). Experimental investigation of the frequency and substitution dependence of negative phi-values in two-state proteins. Biochemistry *44*, 12160–12167. https://doi.org/10.1021/bi0505621.

Roberts, R.W., and Szostak, J.W. (1997). RNA-peptide fusions for the *in vitro* selection of peptides and proteins. Proc. Natl. Acad. Sci. U. S. A. *94*, 12297–12302. https://doi.org/10.1073/pnas.94.23.12297.

Rocklin, G.J., Chidyausiku, T.M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V.K., Chevalier, A., et al. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. Science *357*, 168–175. https://doi.org/10.1126/science.aan0693.

Rodríguez-Rodríguez, E.R., Ledezma-Candanoza, L.M., Contreras-Ferrat, L.G., Olamendi-Portugal, T., Possani, L.D., Becerril, B., and Riaño-Umbarila, L. (2012). A single mutation in framework 2 of the heavy variable domain improves the properties of a diabody and a related single-chain antibody. J. Mol. Biol. *423*, 337–350. https://doi.org/10.1016/j.jmb.2012.07.007.

Sato, S., Cho, J.-H., Peran, I., Soydaner-Azeloglu, R.G., and Raleigh,

D.P. (2017). The N-Terminal Domain of Ribosomal Protein L9 Folds via a Diffuse and Delocalized Transition State. Biophys. J. *112*, 1797–1806. https://doi.org/10.1016/j.bpj.2017.01.034.

Savitski, M.M., Reinhard, F.B.M., Franken, H., Werner, T., Savitski, M.F., Eberhard, D., Martinez Molina, D., Jafari, R., Dovega, R.B., Klaeger, S., et al. (2014). Tracking cancer drugs in living cells by thermal profiling of the proteome. Science *346*, 1255784. https://doi.org/10.1126/science.1255784.

Schellenberger, V., Braune, K., Hofmann, H.J., and Jakubke, H.D. (1991). The specificity of chymotrypsin. A statistical analysis of hydrolysis data. Eur. J. Biochem. *199*, 623–636. https://doi.org/10.1111/j.1432-1033.1991.tb16163.x.

Schellenberger, V., Turck, C.W., Hedstrom, L., and Rutter, W.J. (1993). Mapping the S' subsites of serine proteases using acyl transfer to mixtures of peptide nucleophiles. Biochemistry *32*, 4349–4353. https://doi.org/10.1021/bi00067a026.

Schellenberger, V., Turck, C.W., and Rutter, W.J. (1994). Role of the S' subsites in serine protease catalysis. Active-site mapping of rat chymotrypsin, rat trypsin, alpha-lytic protease, and cercarial protease from Schistosoma mansoni. Biochemistry *33*, 4251–4257. https://doi.org/10.1021/bi00180a020.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382–W388. https://doi.org/10.1093/nar/gki387.

Shah, P., and Gilchrist, M.A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc. Natl. Acad. Sci. U. S. A. *108*, 10231–10236. https://doi.org/10.1073/pnas.1016719108.

Shoichet, B.K., Baase, W.A., Kuroki, R., and Matthews, B.W. (1995). A relationship between protein stability and protein function. Proc. Natl. Acad. Sci. U. S. A. *92*, 452–456. https://doi.org/10.1073/pnas.92.2.452.

Sidore, A.M., Plesa, C., Samson, J.A., Lubock, N.B., and Kosuri, S. (2020). DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. Nucleic Acids Res. *48*, e95. https://doi.org/10.1093/nar/gkaa600.

Sieber, V., Plückthun, A., and Schmid, F.X. (1998). Selecting proteins with improved stability by a phage-based method. Nat. Biotechnol. *16*, 955–960. https://doi.org/10.1038/nbt1098-955.

Singer, J.M., Novotney, S., Strickland, D., Haddox, H.K., Leiby, N., Rocklin, G.J., Chow, C.M., Roy, A., Bera, A.K., Motta, F.C., et al. (2022). Large-scale design and refinement of stable proteins using sequence-only models. PLoS One *17*, e0265020. https://doi.org/10.1371/journal.pone.0265020.

Song, L.-F., Deng, Z.-H., Gong, Z.-Y., Li, L.-L., and Li, B.-Z. (2021). Large-Scale de novo Oligonucleotide Synthesis for Whole-Genome Synthesis and Data Storage: Challenges and Opportunities. Front Bioeng Biotechnol *9*, 689797. https://doi.org/10.3389/fbioe.2021.689797.

Stein, A., Fowler, D.M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and Mechanistic Models for Disease-Causing Protein Variants. Trends Biochem. Sci. *44*, 575–588. https://doi.org/10.1016/j.tibs.2019.01.003.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P.M. (2020). Fast and flexible protein design using deep graph neural networks. Cell Syst. *11*, 402–411.e4. https://doi.org/10.1016/j.cels.2020.08.016.

Strokach, A., Lu, T.Y., and Kim, P.M. (2021). ELASPIC2 (EL2): Combining Contextualized Language Models and Graph Neural Networks to Predict Effects of Mutations. J. Mol. Biol. *433*, 166810. https://doi.org/10.1016/j.jmb.2021.166810.

Stutz, C., and Blein, S. (2020). A single mutation increases heavy-chain heterodimer assembly of bispecific antibodies by inducing structural disorder in one homodimer species. J. Biol. Chem. *295*, 9392–9408. https://doi.org/10.1074/jbc.RA119.012335.

Sun, M.G.F., Seo, M.-H., Nim, S., Corbi-Verge, C., and Kim, P.M. (2016). Protein engineering by highly parallel screening of computationally designed variants. Sci Adv *2*, e1600692. https://doi.org/10.1126/sciadv.1600692.

Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. *19*, 596–604. https://doi.org/10.1016/j.sbi.2009.08.003.

Van Vranken, J.G., Li, J., Mitchell, D.C., Navarrete-Perea, J., and Gygi, S.P. (2021). Assessing target engagement using proteome-wide solvent shift assays. Elife *10*. https://doi.org/10.7554/eLife.70784.

Villegas, V., Martínez, J.C., Avilés, F.X., and Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. J. Mol. Biol. *283*, 1027–1036. https://doi.org/10.1006/jmbi.1998.2158.

Walker, E.J., Bettinger, J.Q., Welle, K.A., Hryhorenko, J.R., and Ghaemmaghami, S. (2019). Global analysis of methionine oxidation provides a census of folding stabilities for the human proteome. Proc. Natl. Acad. Sci. U. S. A. *116*, 6081–6090. https://doi.org/10.1073/pnas.1819851116.

Wang, Z., and Moult, J. (2001). SNPs, protein structure, and disease. Hum. Mutat. *17*, 263–270. https://doi.org/10.1002/humu.22.

Wang, B., Gallolu Kankanamalage, S., Dong, J., and Liu, Y. (2021). Optimization of therapeutic antibodies. Antib Ther *4*, 45–54. https://doi.org/10.1093/abt/tbab003.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. (2022). High-resolution de novo structure prediction from primary sequence.

Xavier, J.S., Nguyen, T.-B., Karmarkar, M., Portelli, S., Rezende, P.M., Velloso, J.P.L., Ascher, D.B., and Pires, D.E.V. (2021). ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Res. *49*, D475–D479. https://doi.org/10.1093/nar/gkaa925.

Xiao, S., Bi, Y., Shan, B., and Raleigh, D.P. (2009). Analysis of core packing in a cooperatively folded miniature protein: the ultrafast folding villin headpiece helical subdomain. Biochemistry *48*, 4607–4616. https://doi.org/10.1021/bi8021763.

Xiao, S., Patsalo, V., Shan, B., Bi, Y., Green, D.F., and Raleigh, D.P. (2013). Rational modification of protein stability by targeting surface

sites leads to complicated results. Proc. Natl. Acad. Sci. U. S. A. *110*, 11337–11342. https://doi.org/10.1073/pnas.1222245110.

Yamaguchi, J., Naimuddin, M., Biyani, M., Sasaki, T., Machida, M., Kubo, T., Funatsu, T., Husimi, Y., and Nemoto, N. (2009). cDNA display: a novel screening method for functional disulfide-rich peptides by solid-phase synthesis and stabilization of mRNA-protein fusions. Nucleic Acids Res. *37*, e108. https://doi.org/10.1093/nar/gkp514.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A. *117*, 1496–1503. https://doi.org/10.1073/pnas.1914677117.

Yourik, P., Fuchs, R.T., Mabuchi, M., Curcuru, J.L., and Robb, G.B. (2019). Staphylococcus aureus Cas9 is a multiple-turnover enzyme. RNA *25*, 35–44. https://doi.org/10.1261/rna.067355.118.

Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. J. Mol. Biol. *353*, 459–473. https://doi.org/10.1016/j.jmb.2005.08.020.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620. https://doi.org/10.1093/bioinformatics/btt593.

Zheng, F., and Grigoryan, G. (2017). Sequence statistics of tertiary structural motifs reflect protein stability. PLoS One *12*, e0178272. https://doi.org/10.1371/journal.pone.0178272.

Zheng, F., Zhang, J., and Grigoryan, G. (2015). Tertiary structural propensities reveal fundamental sequence/structure relationships. Structure *23*, 961–971. https://doi.org/10.1016/j.str.2015.03.015.

Zhou, J., Panaitiu, A.E., and Grigoryan, G. (2020). A general-purpose protein design framework based on mining sequence-structure relationships in known protein structures. Proc. Natl. Acad. Sci. U. S. A. *117*, 1059–1068. https://doi.org/10.1073/pnas.1908723117.

Acknowledgments: We thank Epsilon Molecular Engineering (EME) Corp for providing us with cnvK linker for cDNA display, Rush University and Genome Research Core at University of Illinois Chicago for performing next-generation sequencing, David Minh, Timothy Whitehead, Kresten Lindorff-Larsen, David M. McCandlish, Jack Maguire, John Chodera, and the members of the Rocklin lab for discussions and comments on this manuscript.

Funding: Northwestern University Startup Funding (GJR), JSPS KAKENHI 19J30003 (KT), Human Frontier Science Program Long-Term Fellowship (KT), and JST PRESTO Grant JPMJPR21E9 (KT)

Author contributions: KT designed and performed all experiments, and analyzed the data with help from GJR. JD designed and analyzed stabilities of hallucination-based proteins with help from SO. JC designed ββαα proteins using Rosetta with help from GJR. NMM provided assistance with mathematical derivation and review of enzyme kinetic interpretation. GJR and KT conceived the project. GJR supervised the project and acquired funding. KT and GJR wrote and revised the manuscript, with input from all authors.

Competing interests: Authors declare that they have no competing interests

Data and materials availability: All data and codes are available in the main text, the supplementary materials, or available for download at https://zenodo.org/deposit/7401275 or https://github.com/Rocklin-Lab/cdna-display-proteolysis-pipeline

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S20

Tables. S1 to S4

Supplementary Materials

Supplementary Materials for

Mega-scale experimental analysis of protein folding stability in biology and protein design

Authors: Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Niall M. Mangan, Sergey Ovchinnikov, Gabriel J. Rocklin

Correspondence to: grocklin@gmail.com

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S20
Tables. S1 to S4
Captions for Supplementary Materials

Other Supplementary Materials for this manuscript include the following:

Raw_NGS_count_tables.zip
K50_dG_tables.zip
Processed_K50_dG_datasets.zip
Data_tables_for_figs.zip
Pipeline_qPCR_data.zip
Pipeline_K50_dG.zip
Pipeline_figure_model.zip
AlphaFold_model_PDBs.zip
Blueprints_for_EEHH.zip

Materials and Methods

DNA oligo library construction

All sequences were reverse-translated and codon-optimized using DNAworks2.0 (Hoover and Lubkowski, 2002). Sequences were optimized using *E. coli* codon frequencies because we used an *in vitro* translation kit derived from *E. coli*. Oligo libraries encoding amino acid sequences of Library 1 were purchased from Agilent Technologies. Oligo libraries for Libraries 2-4 were purchased from Twist Bioscience.

Library 1: We selected ~250 designed proteins and ~50 natural proteins that are shorter than 45 amino acids. Then, we created amino acid sequences for deep mutational scanning followed by padding by Gly, Ala, Ser amino acids so that all sequences have 44 amino acids. The total number of sequences is ~244,000 sequences.

Library 2: We selected ~350 natural proteins that have PDB structures that are in a monomer state and have 72 or less amino acids after removing N and C-terminal linkers. Then, we created amino acid sequences for deep mutational scanning followed by padding by Gly, Ala, Ser amino acids so that all sequences have 72 amino acids. The total number of sequences is ~650,000 sequences. This library also includes scramble sequences to construct unfolded state model.

Library 3: We selected ~150 designed proteins and created amino acid sequences for deep mutational scanning of the proteins. We also included comprehensive deletion and Gly/Ala insertion of all wild-type proteins inlcuded in Library1 and Libary2. Additionally, amino acid sequences for comprehensive double mutant analysis on polar amino acid pairs were also included.

Library 4: Amino acid sequences for exhaustive double mutant analysis on amino acid pairs located in close proximity were included. We also include overlapped sequences to calibrate effective protease concentration and to check consistency between libraries.

EEHH design method

EEHH protein design was performed in three steps: (1) backbone construction, (2) sequence design, (3) selection of designs for deep mutational analysis. Backbone construction (the de novo creation of a compact, three-dimensional backbone with a pre-specified secondary structure) was performed using a blueprint-based approach described previously (Huang et al., 2011; Koga et al., 2012). All blueprints are included as Blueprints_for_EEHH.zip in Supplementary Materials.

Hallucination design method

We used a TrRosetta hallucination protocol described previously in (Anishchenko et al., 2021; Norn et al., 2021) and available at https://github.com/gjoni/trDesign/tree/master/02-GD to unconditionally generate protein backbones and sequences with lengths ranging from 46 to 69 amino acids by maximizing the Kullback–Leibler divergence between the predicted and background distance/angle distributions. Predicted distograms and anglegrams were used to obtain 3D structures of these models as described in the TrRosetta paper (Yang et al., 2020). We selected the best designs according to the predicted distogram and 3D structure match.

DNA and mRNA preparation for cDNA display proteolysis method

Oligo libraries were amplified by PCR using KOD PCR Master Mix (TOYOBO) to add T7 promoter, PA tag to an N-terminal, and His tag to an C-terminal of the proteins. The number of cycles was chosen based on a test qPCR run to avoid overamplification using SsoAdvanced Universal SYBR Green Supermix (BIORAD). The PCR product was gel extracted to isolate the expected length product. Then we used T7-Scribe Standard RNA IVT Kit (CELLSCRIPT) to synthesize mRNA using the DNA fragment as a template.

Preparation of protein-cDNA complex

We basically follow the protocol described in the previous literature (Arai et al., 2020; Yamaguchi et al., 2009) with some modifications.

Photocross-linking between mRNA and the puromycin linker: We prepared the photocrosslinking reaction solution including 200 mM NaCl, 40 mM Tris-HCl (pH 7.5), 20 μM cnvK linker (EME corporation), 20 μM mRNA. The solution was incubated at 95°C for 5 min, then slowly cooled down to 45°C (0.1°C / 1 second) using a thermal cycler. Then the solution including the duplex was irradiated with UV light at 365 nm using a 6W Handheld lamp (Thermofisher).

In vitro translation and reverse transcription: We prepared PURExfrex 2.0 (GeneFrontier) translation system with mRNA-cnvK linker duplex and RiboLock RNase Inhibitor (Thermofisher) and incubate the sample at 37°C for 2 hrs. After the incubation, 100 mM EDTA was added to the sample to dissociate ribosomes. Then, an equal amount of binding/washing buffer (30 mM Tris pH 7.5, 500 mM NaCl, 0.05% Tween 20) was added. The solution was added to Dynabeads MyOne Streptavidin C1 (Thermofisher) to pull down the protein-mRNA complex and incubated at room temperature for 20 min. Then, the beads were washed by binding/washing buffer once and rinsed twice by TBS (10 mM Tris-HCl pH7.5, 100 mM NaCl), and we added reverse transcription solution (PrimeScript RT Reagent Kit; Takara) onto the beads with protein mRNA complex, and incubated the beads at 37°C for 30 mins.

Purification of protein-cDNA complex: After the reverse transcription, the protein-cDNA complex was eluted by binding/washing buffer with RNase T1 (Thermofisher). The eluent was added His Mag Sepharose Ni (Cytiva) and incubated at room temperature for 30 min. Then the complex was eluted by binding/washing buffer with 400 mM imidazole then the eluent was buffer-exchaged by Zeba Spin Desalting Column (Thermofisher). Then the complex was snap-frozen by liquid nitrogen and stored at -80°C until the following protease assay.

Protease assay on protein-cDNA complex: We prepared 40 μL of 11 protease three-fold dilution series from 25 μM for replicate1 and 43.3 (= 25 x $3^{0.5}$) μM for replicate2, then added them to 12 of 20 μL the protein-cDNA complex. After 5 min protease digestion in room temperature, we added 200 μL chilled 2% BSA in PBS to quench the reaction, then the solution was added to 10 μL Dynabeads Protein G (Thermofisher) with anti-PA tag antibody (Wako; Clone number: NZ-1; 1μg antibody per 30 μL beads), and incubated at 4°C for 1 hr. Then the beads were washed by washing buffer (PBS including 800 mM NaCl and 1% Triton) three times and rinsed by PBS three times, then the complex was eluted with 50 μL PBS including 250 μg/mL PA peptide (Wako) and 200 μg/mL BSA (Thermofisher).

qPCR analysis of cDNA display proteolysis results on individual proteins (for Fig. S1)
The cDNA amount for each specific sequence in the eluents was quantified by qPCR using SsoAdvanced Universal SYBR Green Supermix and specific primers for each sequence. The qPCR was performed using CFX96 Touch Real-Time PCR Detection System (BIORAD), and the qPCR cycles were determined by the CFX Maestro Software (BIORAD).

Next-generation sequencing sample preparation
For DNA library analysis, one-half volume (25 μL) of the eluted cDNA of the complex was amplified by PCR using SsoAdvanced Universal SYBR Green Supermix (BioRad) to add P5 and P7 NGS adapter sequence. The number of cycles was chosen based on a test qPCR run using the same PCR reagents to avoid overamplification. The DNA fragment length and concentration were confirmed by 4200 TapeStation System (Agilent), then the samples were analyzed by NovaSeq 6000 System (Illumina).

Processing of next-generation sequencing data
Each library in a sequencing run was identified via a unique 6 or 8 bp barcode. Following sequencing, reads were paired using the PEAR program (Zhang et al., 2014) then the adapter sequences were moved by Cutadapt (Martin, 2011). Reads were considered counts for a sequence if the read perfectly matched the ordered sequences at the nucleotide level.

Overall strategy for inferring K_{50} and ΔG from sequencing data
We used Bayesian inference to infer K_{50} and ΔG values for all sequences in our library. This analysis uses two main models. The first model is called the "K_{50} model" and infers each sequence's K_{50} values based on the sequencing count data. The second model is called the "unfolded state model" and predicts each sequence's unfolded state K_{50} value ($K_{50,U}$) based on its sequence. Both models are implemented in Python 3.9 using the Numpyro package (Phan et al., 2019) version 0.80. Here, we first describe the structure of each model, and then we describe the practical process of fitting the parameters of each model. Our scripts to reproduce the complete fitting process are provided in the Supplementary Materials.

Structure of the K_{50} model to infer K_{50} values from next-generation sequencing data
We modeled our selection results using the single turnover kinetics model described in Fig. 1B. We chose this model because we expect that the total concentration of protein-cDNA complex is low compared to the amount of added enzyme and because the model captures the saturation behavior observed by qPCR at high enzyme concentration (Fig. S1). Instead of attempting to capture the microscopic complexity of our system (millions of different substrates and potential inhibitors), the purpose of the model is to treat each substrate in a consistent, simplified manner and infer reasonable parameters.

Our model makes two main assumptions. First, we assume that each sequence is cleaved independently, with no competition or product inhibition. As described by Fig. 1 eqs. 2 and 3, cleavage is described by four parameters: enzyme concentration (E), time (t), and the kinetic parameters K_{50} and $k_{\max}$. All experiments used a fixed five minute reaction time. Based on qPCR analysis of individual sequences (Fig. S1), we fixed the quantity $k_{max} * t$ at $10^{0.65}$ for all sequences. Each sequence's unique stability is defined by the K_{50} parameter that represents the enzyme concentration producing the half maximal cleavage rate (Fig. 1 eq. 3). Our second main assumption is that we can interpret our K_{50} values as representing the dissociation constants (K_D) between each protein sequence and the enzyme ($K_{50} \approx K_D$, Fig. 1 eq. 6). From this assumption, we can determine the folding stability of each sequence (ΔG) based on the relationship between the observed K_{50} value and theoretical K_{50} values for the fully folded and fully unfolded states ($K_{50,F}$ and $K_{50,U}$, Fig. 1 eqs. 5-7). Although we can directly fit K_{50} values without making any assumptions about the microscopic basis for K_{50} (see Supplementary Text for the detail), assuming that $K_{50} \approx K_D$ aids our interpretation and enables us to directly fit ΔG values to our data using the *Coupled* approach described below.

To fit our model to our sequencing counts data, we first assume that the cDNA display process produces an unknown initial distribution of full-length protein-cDNA complexes (the $cDNA_0$ distribution). The distribution of sequences at enzyme concentration E (the $cDNA_E$ distribution) is the product of the initial sequence distribution $cDNA_0$ and the surviving fraction of each sequence according to Fig. 1 eqs. 2 and 3, after re-normalizing the total surviving fraction of all sequences to 1.

$$cDNA_{E,i} = cDNA_{0,i} * Frac\ ([E], K_{50,i}) / \sum_j (cDNA_{0,j} * Frac\ ([E], K_{50,j}) \quad (8)$$

Finally, we assume that our deep sequencing counts result from n_{sel} independent selections from the $cDNA_E$ distribution, where n_{sel} is the number of sequencing reads that exactly matched our specified DNA sequences.

We apply the K_{50} model in two different ways based on whether K_{50} values for trypsin and chymotrypsin are *Independent* or *Coupled*. The "Independent" procedure is used in Steps 1, 2 and 5 in the section "Procedure for fitting all data". In the independent procedure, the inputs to the model are the sequencing counts data from experiments with one protease, the enzyme concentrations, the reaction time, and the k_{max} constant. We fit the model by sampling two parameters per sequence from normal prior distributions: (1) K_{50}, and (2) the initial fraction of each sequence in the $cDNA_0$ distribution. The "Coupled" procedure is used in Step 5 in the section "Procedure for fitting all data". In the coupled procedure, the inputs to the model are the sequencing counts data from experiments with both proteases, the enzyme concentrations, the reaction time, the k_{max} constant, the $K_{50,F}$ constants representing the universal K_{50} value for sequences in the folded state (one for each protease), and the predicted $K_{50,U}$ values for all sequences for both proteases from the unfolded state model. We then assume that each sequence has a specific ΔG value that is shared across both proteases. We use this shared ΔG value along

with $K_{50,F}$ and $K_{50,U}$ (for each protease) to determine K_{50} for each protease according to Fig. 1 Eqs. 5 and 7. Finally, we fit the coupled model by sampling two parameters per sequence from normal prior distributions: (1) ΔG, and (2) the initial fractions of each sequence in $cDNA_0$.

Full results from both the independent and coupled fitting procedure are provided in K50_dG_Dataset1_Dataset2.csv and K50_Dataset3.csv. For our stability parameters (protease-specific K_{50} in the independent procedure and ΔG in the coupled procedure) we report the median of the posterior distribution as well as the upper and lower limits of the 95% confidence interval (the 2.5%ile and 97.5%ile values of the posterior distribution). We also used the protease-specific K_{50} values from the independent procedure to compute protease-specific ΔG values. We do this using the same $K_{50,F}$ and $K_{50,U}$ values used in the coupled procedure according to Fig. 1 Eqs. 5 and 7. These protease-specific ΔG estimates are also reported in K50_dG_Dataset1_Dataset2.csv and are only used to examine the consistency between different proteases (e.g. Fig. 1F and Fig. 2D). In some cases, the independently fit K_{50} values can lead to impossible values for ΔG. This can occur if K_{50} is higher than $K_{50,F}$ (observed cleavage is slower than our limit for cleavage in the folded state) or if K_{50} is lower than $K_{50,U}$ (observed cleavage is faster than predicted cleavage in the unfolded state). If the median protease-specific K_{50} or the confidence interval limits for a particular sequence lead to impossible ΔG values for that sequence, we report dummy values for the corresponding protease-specific ΔG estimates.

<u>Structure of the unfolded state model to infer unfolded K_{50} ($K_{50,U}$) from scrambled sequence data</u>
Our unfolded state model is similar to the model employed previously (Rocklin et al., 2017) with two notable differences. First, instead of assuming that all scrambled sequences are fully unfolded, we assume that each scrambled sequence has its own unknown folding stability, with a prior distribution biased toward low stability (normal prior centered at ΔG = -1, sigma = 4). Second, instead of fitting an unfolded state model for each protease independently, we assume that each scrambled sequence's stability (ΔG) is common across both proteases, and fit the models for each protease together. As a result, the majority of scrambled sequences are modeled as completely unfolded (Fig. S2C), but some scrambled sequences are modeled as stable when that interpretation is consistent with both the trypsin and chymotrypsin data.

Our unfolded state has three parts: (1) a position specific scoring matrix (PSSM) that describes how the amino acid sequence in a 9-mer window (the P5 to P4' positions in protease nomenclature) determine the cleavage rate at the P1 position, (2) a local response function describing the saturation of the cleavage rate for a single P1 position, (3) a global response function that determines $K_{50,U}$ based on the sum of the cleavage rates at all possible P1 positions in the full sequence.

To fit the PSSM, we assumed an identical normal prior distribution of scores at all positions, with several exceptions. Due to known critical importance of the P1 position, we used a wider prior distribution of scores for all amino acids in the P1 position for both proteases. We also used wider prior distributions at all positions (P5-P4') for the amino acids Asp, Glu, and Pro, due to the established large effects of these amino acids on cutting rates.

For the local response function to saturation of the cleavage rate at P1 site k, we used a logistic function:

$$SS_k = logistic\left(\sum_{site=P5}^{P4'} PSSM(aa_{site}, site)\right) \quad (9)$$

where SS_k (site saturation) is the saturation of the cutting rate at site P1=k, aa_{site} is the amino acid identity at *site*, and *logistic* is the logistic function f(x) = 1 / (1+e^x). We fit the 21 (20 amino acids + 'X' representing empty sites) x 9 =189 elements of the PSSM for each protease.

For the global response function (determining $K_{50,U}$ based on the sum of SS_k across the full protein sequence), we use a sum of logistic functions with 10 different activation thresholds.

$$K_{50,U} = maxK_{50,U} - Scale * \sum_{l=1}^{10} logistic\left(\left(\sum_{k=1}^{length\ of\ seq} SS_k\right) - threshold_l\right) \quad (10)$$

where $maxK_{50,U}$ is the highest possible $K_{50,U}$ value ($K_{50,U}$ assuming no cut sites), *Scale* is the range of possible $K_{50,U}$ values, and $threshold_l$ is the value of the lth activation threshold for the global response function. All K_{50} values (including $maxK_{50,U}$) are in $\log_{10}$ molar units.

The key parameters of the unfolded state model (for a single protease) are the 21 x 9 =189 elements of the *PSSM*, the $maxK_{50,U}$, the *scale*, and the 10 *threshold* values. These parameters determine $K_{50,U}$ for each sequence by Eqs. 9 and 10. In addition to these parameters, we also sample the ΔG values for each scrambled sequence during fitting. These sampled parameters (as well as the universal $K_{50,F}$ value for all sequences) are sufficient to determine a theoretical K_{50} value for each scrambled sequence by re-writing Fig. 1 Eq. 6:

$$1/K_{50} = \frac{([U]/K50_u) + ([F]/K50_f)}{([U]+[F])} = f_U/K_{50,U} + (1 - f_U)/K_{50,F} \quad (11)$$

where f_U is the fraction of unfolded molecules:

$$f_U = 1/(1 + \frac{\Delta G}{RT}) \quad (12)$$

The input data for the model are the observed K_{50} values for all scrambled sequences. The parameters of the model are fit by assuming that all observed K_{50} values should agree (with small, normally distributed errors) with the theoretical K_{50} values determined by the model parameters. After fitting the model, we used the median of the posterior distributions of *PSSM*, $maxK_{50,U}$, *scale*, and the 10 *threshold* parameters as the final model parameters. We used these final model parameters to calculate $K_{50,U}$ for all sequences in our experiments without considering any uncertainty from the model posterior distribution.

Procedure for fitting all data

Step 1: Estimation of 'effective' protease concentrations for each library: We employed four DNA oligonucleotide libraries for this study. Although we tried to minimize the difference between assay conditions, we also fit "effective" protease concentrations to our data in order to minimize batch-to-batch differences. We used the K_{50} model to perform this fitting and fit protease concentrations for trypsin and chymotrypsin entirely independently. The main assumption of this fitting is that each sequence should have the same K_{50} when assayed in different libraries. By enforcing that each sequence had a single K_{50} value regardless of what library it appears in, we calibrated the protease concentrations in each library against each other. Although we did not use universal control sequences in all four libraries, each library contained 1000 to 2000 sequences that overlapped at least one other library in a fully connected graph. Specifically, the library pairs 1+4, 2+4, 3+4, 1+2, and 2+3 each included 1,000 to 2,000 overlapping sequences.

The overall model included 96 experimental conditions (12 protease concentrations per replicate x 2 replicates x 4 libraries; one of the 12 protease concentrations was the fixed "no protease" starting condition). However, each sequence was only present in 48 of the 96 conditions because any individual sequence was only present in two out of the four libraries. The inputs to fit the model were the sequencing counts data, the reaction time (t), and the k_{max} constant. Additionally, to set the overall scale of the protease concentration series, we fixed the effective protease concentrations for Library 4 at the expected protease concentrations (i.e. three-fold serial dilutions of 25 μM protease (Replicate 1) or 43.3 μM protease (Replicate 2)). We also fixed all of the starting samples at zero protease. Using these model inputs, we sampled the K_{50} values (one per sequence), the remaining 66 protease concentrations, and the initial sequence distributions $cDNA_0$ (a separate $cDNA_0$ was used for each of the 8 replicates). Normal priors (with lower/upper boundaries for some parameters) covering the range of experimentally relevant values were used for the model parameters. Sampling was performed using the No U-Turn Sampler (NUTS) in Numpyro with 50 steps of equilibration and 25 steps of production. We used the medians of the protease concentrations from our 25 posterior samples as our final calibrated protease concentrations for all further analysis (discarding the uncertainties).

Step 2: Estimation of K_{50} values of scramble sequences: To train the unfolded state model, we need to determine K_{50} values for our scramble sequences, which were included in Library 2. We used the Independent K_{50} model for this step. The input data were the sequencing counts data from two replicates (i.e. 12 protease concentrations x 2 replicates = 24 data points per sequence), the reaction time (t), the k_{max} constant, and the effective protease concentrations obtained in Step 1. We sampled the initial sequence distribution $cDNA_0$ (a separate $cDNA_0$ for each replicate) and K_{50} for all sequences included in Library 2. Normal priors (with lower/upper boundaries for

some parameters) covering the range of experimentally relevant values were used for the model parameters. Sampling was performed using the No U-Turn Sampler (NUTS) in Numpyro with 100 steps of equilibration and 50 steps of production.

Step 3: Construction of unfolded state model: We trained the unfolded state model for predicting $K_{50,U}$ using K_{50} values obtained in Step 2. The input sequences were scrambled sequences of wild-type domains selected for deep mutational screening. In addition to our set of exactly scrambled sequences (matching the wild-type amino acid composition 100%), we also included scrambled sequences containing 50%, 60%, 70%, 80%, and 90% of the number of hydrophobic amino acids in the original wild-type sequences. These sequences helped ensure the large majority of our scrambled pool was fully unfolded. Additionally, because all sequences in our experiments are padded with G/S/A linkers up to a constant length, we generated scrambled sequences using two different padding procedures. In the first approach, we designed scrambled sequences that matched the original wild-type length and were padded with G/S/A up to 72 amino acids. In the second approach, we designed 72 amino acid-length scrambles approximately matching the composition of an original wild-type domain, regardless of the length of that wild-type. These scrambled sequences required no additional padding. After measuring K_{50} for all scrambles, we only used sequences with a 95% confidence interval smaller than 0.5 $\log_{10}$ molar units for model training for model fitting (64,238 sequences in total, see Fig. S3). In addition to the exact experimental sequences, we also augmented the training dataset with dummy sequences where GS linkers were replaced by the blank 'X' amino acid.

The inputs for the model are amino acid sequences created as described above, and their observed K_{50} for trypsin and chymotrypsin obtained in Step 2. The parameters of the model are fit by assuming that all observed K_{50} values should agree (with small, normally distributed errors) with the theoretical K_{50} values. In this model, we sampled the 21 x 9 =189 elements of the *PSSM*, the *site bias*, the $maxK_{50,U}$, the *scale*, and the 10 *threshold* values. These parameters determine $K_{50,U}$ for each sequence by Eqs. 9 and 10. In addition to these parameters, we also sample the ΔG values for each scrambled sequence during fitting.

Normal priors (with lower/upper boundaries for some parameters) covering the range of experimentally relevant values were used for the model parameters. Using NUTS model, we sampled the parameters described above, then reported the median of the 100 posteriors after removing the initial 400 steps. In Step 4, we used these final model parameters to calculate $K_{50,U}$ for all sequences in our experiments without considering any uncertainty from the model posterior distribution.

Step 4: Prediction of unfolded K_{50} values ($K_{50,U}$) across the full dataset: Using the final model parameters obtained in Step 3, we predicted $K_{50,U}$ values for each amino acid sequence in the libraries without considering any uncertainty. Additionally, since the model was constructed to

predict unfolded K_{50} for sequences with 86 amino acids, we added a Gly linker 'GGG' to both ends, followed by padding by 'X' up to 86 amino acids.

Step 5: Estimation of K_{50} values and calculation of ΔG for trypsin and chymotrypsin: We applied the *Coupled* K_{50} model to each of the four libraries separately. The inputs to the model are the sequencing count data from trypsin and chymotrypsin experiments (i.e. 12 protease concentrations x 2 replicates x 2 proteases = 48 data points per sequence), the effective protease concentrations obtained in Step 1, the reaction time, the k_{max} constant ($t*k_{max} = 10^{0.65}$ based on qPCR analysis; see Fig. S1), the $K_{50,F}$ constants (3 for trypsin, 2 for chymotrypsin; determined based on the dynamic range of proteolysis experiment; see Fig. S5), and the $K_{50,U}$ values predicted by the unfolded model in Step 4. Using the inputs, we sampled ΔG shared between trypsin and chymotrypsin, and initial sequence distribution $cDNA_0$ for each protease for each replicate (although our experiments utilized the same batch of the cDNA-protein complex for two replicates).

Normal priors (with lower/upper boundaries for some parameters) covering the range of experimentally relevant values were used for the model parameters. Using NUTS in Numpyro module, we sampled the posteriors of shared ΔG along with other parameters, then obtained the median of the 50 posterior samples after removing the initial 100 steps. Full results from both the independent and coupled fitting procedure are provided in K50_dG_Dataset1_Dataset2.csv and K50_Dataset3.csv. For our stability parameters (protease-specific K_{50} in the independent procedure and ΔG in the coupled procedure) we report the median of the posterior distribution as well as the upper and lower limits of the 95% confidence interval (the 2.5%ile and 97.5%ile values of the posterior distribution).

We also applied the *Independent* K_{50} model to each of the four libraries separately. The inputs to the model are the sequencing count data (i.e. 12 protease concentrations x 2 replicates = 24 data points per sequence), the effective protease concentrations obtained in Step 1, the reaction time, the k_{max} constant ($t*k_{max} = 10^{0.65}$ based on qPCR analysis; see Fig. S1). Using the inputs, we sampled K_{50} for each protease, and initial sequence distribution $cDNA_0$ for each protease for each replicate (although we utilized the same batch of the cDNA-protein complex for two replicates).

Normal priors (with lower/upper boundaries for some parameters) covering the range of experimentally relevant values were used for the model parameters. Using NUTS in Numpyro module, we sampled the posteriors of K_{50} for trypsin and K_{50} for chymotrypsin along with other parameters, then obtained the median of the 50 posterior samples after removing the initial 100 steps.

Then, we computed protease-specific ΔG values using the protease-specific K_{50} values from the *Independent* model. We do this using the same $K_{50,F}$ and $K_{50,U}$ values used in the coupled

procedure according to Fig. 1 Eqs. 5 and 7. These protease-specific ΔG estimates are also reported in K50_dG_Dataset1_Dataset2.csv and K50_Dataset3.csv, and are only used to examine the consistency between different proteases (e.g. Fig. 1F and Fig. 2D). In some cases, the independently fit K_{50} values can lead to impossible values for ΔG. This can occur if K_{50} is higher than $K_{50,F}$ (observed cleavage is slower than our limit for cleavage in the folded state) or if K_{50} is lower than $K_{50,U}$ (observed cleavage is faster than predicted cleavage in the unfolded state). If the median protease-specific K50 or the confidence interval limits for a particular sequence lead to impossible ΔG values for that sequence, we reported dummy values for the corresponding protease-specific ΔG estimates.

The actual number of sequencing counts, as well as the number of counts predicted for all sequences at all concentrations according to the fitted model parameters, are given in Raw_NGS_count_tables.zip and Pipeline_K50_dG.zip.

Data selection for Fig. 1E and F

We show all data from Library 3 within the range -2 < ΔG < 5 kcal/mol & log10_K50_trypsin < 1.75 & log10_K50_chymotrypsin < 2.25. We then overlaid the wild-type and four mutants of Protein G measured in Library 2.

Replicate analysis of K_{50} (Fig. 1E)

Instead of sampling K_{50} values using 24 samples per protease at one time as described in Step 5 above, we sampled K_{50} values using one experiment set (i.e. 12 samples) and obtained K_{50} for trypsin replicate 1 and 2, and chymotrypsin replicate 1 and 2. Note that we still used the calibrated protease concentrations to improve consistency between replicates. The replicates were conducted on different days using the same preparation of the protein-cDNA complex.

Classification of Datasets #1, #2, and #3 based on the quality of the data (For Fig. 2)

All mutational scanning data was classified into nine groups (0 through 8) according to the protocol in Fig. S8. We determined that a mutational scan was high quality (suitable for Dataset #2) if there was minimal missing data, minimal low confidence data, an appropriate slope, intercept, and correlation between the trypsin and chymotrypsin samples, sufficient wild-type stability, and the mutational scan did not include an unusual fraction of stabilizing mutations suggesting poor folding. For inclusion in the smaller Dataset #1, we additionally required that the wild-type stability was lower than 4.5 kcal/mol so that stabilizing mutations could still fall within the assay's dynamic range. These sequences are considered "Group 0"; the remaining sequences in Dataset 2 are considered "Group 1". Double mutant sequences were included in Datasets 1 and 2 based on whether the original wild-type mutational scan was included in that dataset.

All sequences in Dataset 1 and Dataset 2 are included in K50_dG_Dataset1_Dataset2.csv. All sequences in this file have an inferred ΔG estimate value, but only sequences in Dataset 1 have a

tabulated ΔΔG estimate. Of course, one can calculate ΔΔG for the remaining sequences in Dataset 2, but these ΔΔG values will be biased toward destabilizing mutations because stabilizing mutations would typically be indistinguishable from the wild-type stability. *Note that Datasets 1 and 2 include a small number of sequences with low quality data because these sequences come from mutational scans that are high quality overall.* Although these tables include all K_{50}, Δ G, and ΔΔG data (for Dataset 1), low quality data have been filtered out and replaced by a – symbol in the columns labeled "_ML" (for machine learning).

The remaining groups were defined this way:
Group 2: The wild-type protein is too unstable to see sequence-stability relationships.
Group 3: Poor expression (low counts in next-generation sequencing) for the assay.
Group 4: Very few destabilizing mutations, suggesting aggregation and/or molten globule formation
Group 5: The wild-type is too stable to see consistency between trypsin and chymotrypsin
Group 6 and 7: Low agreement between trypsin and chymotrypsin due to the absence of aromatic amino acids (i.e. chymotrypsin cleavage sites) or the presence of protease recognition sequences in the linker region.
Group 8: Did not fit into groups 2-7, but did not pass the quality metrics for groups 0 and 1.

Dataset #3 includes all data combined (Groups 0-8), even the data from Groups 2-8 that were excluded from Datasets 1 and 2. Although many of the K_{50} values from Groups 2-8 likely reflect factors other than folding stability (e.g. aggregation, low expression, etc.), these data can still be used to train models that directly predict K_{50}. Again, a small fraction (~4%) of the K_{50} values in Dataset #3 are low confidence and have been replaced by a – symbol in the "_ML" columns.

Principal component analysis (related to Fig. 3)
We performed principal component analysis to determine the factors influencing stability of different amino acids. To this end, we utilized 15,440 sites in the 337 domains that are classified as G0 in the above. All folding stability data were clipped between from -1 to 5 (kcal/mol) because the folding stability outside the dynamic range is not reliable, and then the average of the stability for 20 amino acids for each site was subtracted from the data. Using the data, we performed PC analysis using the scikit-learn library implemented in Python 3.

Side chain contacts and burial analysis (Fig. 3D and 6H)
Burial values and contact counts were computed based on AlphaFold models (Jumper et al., 2021) of all sequences using the included script Burial_side_chain_contact_Fig3_Fig6.ipynb based on Bio.PDB (Hamelryck and Manderick, 2003) and BioPython (Cock et al., 2009)). The calculation is based on the Rosetta "sidechain_neighbors" LayerDesign method previously reported (Rocklin et al., 2017). Briefly, to calculate the burial or contacts of residue X, we added up the number of residues in a cone projecting out 9 Å away from the Cβ atom on residue X in

the direction of the residue X Cα-Cβ vector. "Burial" (Fig. 6H) indicates the number of Cα atoms in the cone. Contact counts (Fig. 3D) each count different atoms inside the cone: "Side chain contact count" (Fig. 3D) counts all Cβ atoms; "Aromatic side chain contact count" counts all CE2 atoms of Phe, Tyr, and Trp;, "Acidic side chain contact count" counts all Glu OE1 and Asp OD1 atoms; and "Basic side chain contact count" counts all Lys NZ and Arg NE atoms.

Secondary structure determination (Fig. 3D)

Using the DSSP algorithm (Joosten et al., 2011; Kabsch and Sander, 1983), we obtained secondary structure information based on AlphaFold models.

Selection method of site pairs for double mutational analysis (related to Fig. 4)

Double mutants were selected for analysis in two ways. First, we manually selected polar interactions where either amino acid appeared important for stability in single mutational analysis. These pairs were mainly included in Library 3. Second, we used the program confind (Zheng and Grigoryan, 2017; Zheng et al., 2015) to identify interacting residues. All confind pairs with notable interactions such as polar interactions and cation-π interactions were selected, along with a randomly chosen subset of more common interactions such as hydrophobic interactions. These pairs were included in Library 4.

Thermodynamic coupling analysis (related to Fig. 4)

Thermodynamic coupling refers to the change in folding stability due to the interaction between two amino acids after removing folding stability effects from each amino acid individually. To determine this "nonadditivity", we first modeled our double mutant data using a fully additive model (no thermodynamic coupling). The deviations from this model then reveal the thermodynamic coupling. Our additive model assumes that the absolute stability (ΔG) of each sequence is the sum of an amino acid-dependent term for site one (ΔG_1) and an amino acid-dependent term for site two (ΔG_2)

$$Expected\ \Delta G_{aa1,aa2} = \Delta G_{1,aa1} + \Delta G_{2,aa2} \quad (13)$$

The forty site-specific terms (one ΔG_1 term for each amino acid at site one and one ΔG_2 term for each amino acid at site two) are not experimentally measurable; they are inferred based on minimizing the error of the additive model. We used Bayesian inference to infer the forty ΔG_1 and ΔG_2 terms for each set of mutants. The inputs to fit the model were the observed 400 ΔG values (20 amino acids at site one x 20 amino acids at site two) for a particular site pair. Using NUTS, we sampled ΔG_1 and ΔG_2 by assuming that the 400 observed ΔG values should agree (with small, normally distributed errors) with the expected ΔG values determined by eq. 13. Both expected and observed ΔG values were clipped to the range of -1 to 5 kcal/mol. We used 100 steps of burn-in and used the median of 50 posterior samples as the final values of the ΔG_1 and

ΔG_2 terms. Using these terms, we calculated the expected (additive) ΔG for each sequence, and then the thermodynamic coupling:

$$Thermodynamic\ coupling_{aa1,\,aa2} = Observed\ \Delta G_{aa1,aa2} - Expected\ \Delta G_{aa1,aa2} \quad (14)$$

To calculate the uncertainty in the thermodynamic coupling, we re-fit the additive model 50 times by bootstrap resampling of the 400 observed ΔG values. This ensures the ΔG_1 and ΔG_2 terms are not overly dependent on a single experimental measurement. The model fitting code is provided in Additive_model_Fig4.ipynb.

Wild-type amino acid prediction model (related to Fig. 5)

The classification model in Fig. 5 used a sum of logistic functions with learned amplitudes to define the weighting function. The overall model is defined below:

$$p(aa) = Softmax(\ (\sum_{i=1}^{100} amp_i\ *\ logistic(\ (\Delta G_{aa} - threshold_i)\ *\ steepness))\ +\ offset_{aa}) \quad (15)$$

where *p(aa)* is the probability of amino acid *aa*, *softmax* is the softmax function $\frac{e^{x_{aa}}}{\sum_{aa} e^{x_{aa}}}$, *logistic* is the logistic function $f(x) = 1 / (1+e^x)$, *i* indexes the 100 logistic functions defining the weighting function, *amp* is the learned vector describing the amplitudes of the logistic functions, *threshold* is the vector describing the centers of the logistic functions, *steepness* defines the steepness of the logistic functions, and *offset* the learned vector (length 19 for the 19 non-Cys amino acids) describing the absolute probability offset for each amino acid.

We used Bayesian inference to infer the *amp* vector (length 100) and *offset* vector (length 19 for the 19 non-Cys amino acids). The logistic *threshold* vector was fixed at 100 evenly spaced points between -2 and 7 kcal/mol. The *steepness* term was fixed at 5. The inputs to fit the model were the observed ΔG values and the wild-type amino acid identities for each site within the natural protein domains. Using NUTS, we sampled *amp* and *offset* by assuming that the observed wild-type amino acids were randomly chosen at each site according to the predicted probability distribution for that site, calculated according to eq. 15. We then reported the median and the standard deviation of 100 posterior samples after removing the initial 500 steps. The fitting script is included in Classification_model_Fig5.ipynb.

Supplementary Text

Derivation of eq.3 in Fig. 1B

We modeled the cleavage events, where Protease enzymes (E) and protein substrates (S) form an ES complex to produce cleaved protein products (P). The goal is to get a product formation equation in terms of the total product, initial enzyme and substrate concentrations and kinetic constants.

$$[E] + [S] \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} [ES] \xrightarrow{k_2} [E] + [P] \quad (1)$$

Also, we defined equilibrium constant K_{50}:

$$K_{50} = \frac{[E][S]}{[ES]} \quad (1')$$

Based on the model (1), we can obtain the following dynamic formulas:

$$\frac{d[S]}{dt} = -k_1[E][S] + k_1[ES] \quad (16)$$

$$\frac{d[ES]}{dt} = k_1[E][S] - k_1[ES] - k_2[ES] \quad (17)$$

$$\frac{d[P]}{dt} = k_2[ES] \quad (18)$$

The first two of these are assumed to be at quasi-steady state. The following are additional conservation equations for substrate-product and enzyme:

$[S_0] = [S] + [ES] + [P]$ (19) where $[S_0]$ is initial amount of substrates

$$[S_{total}] = [S] + [ES] \quad (20)$$

Additionally, the reaction conditions in the study were not substrate-excessive but enzyme-excessive:

$[E_{total}] = [E] + [ES] \approx [E]$ (21) (because [E] >> [ES] or [S])

Using eqs. 1',19, and 20, the following can be derived to find an expression for the enzyme-substrate complex in terms of the initial substrate and enzyme concentration:

$$[ES] = 1/K_{50}[E][S] = 1/K_{50}[E]\,(S_0 - [ES] - [P])$$

$$[ES](1 + 1/K_{50}[E]) = 1/K_{50}[E](S_0 - [P]) = 1/K_{50}[E][S_{total}]$$

$$[ES] = \frac{1/K_{50}[E]}{1+1/K_{50}[E]}[S_{total}] = \frac{[E]}{K_{50}+[E]}[S_{total}] \quad (22)$$

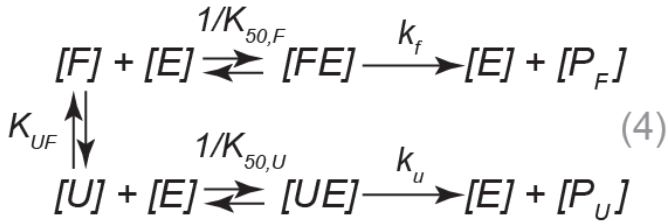
Substituting eq. 22 into eq. 18 and using the approximation $[E_{total}] \approx [E]$, the an expression for the dynamics of the product formation in terms of enzyme concentration and substrate can be found:

$$\frac{d[P]}{dt} = \frac{k_2\,[E_{total}]}{K_{50}+[E_{total}]}[S_{total}]$$

Thus, the observed kinetic rate is $k_{obs} = \frac{k_2[E_{total}]}{K_{50}+[E_{total}]}$ (This eq. 3)

<u>Derivation of eq.6 and eq.7 in Fig1B</u>

We modeled the cleavage events, where Protease enzymes (E) and folded substrates (F) or unfolded substrates (U) form a FE or UE complex to produce cleaved protein products (P_F or P_U). The goal is to get a product formation equation in terms of the total product, initial enzyme and substrate concentrations and kinetic constants. We follow a similar derivation to that above for a single enzyme/substrate:



(4)

where $1/K_{50,F} = \frac{[FE]}{[F][E]}$ (23), $1/K_{50,U} = \frac{[UE]}{[U][E]}$ (24), $K_{UF} = \frac{[U]}{[F]}$ (25) , and k_f and k_u are rate constant for cleavage of the bound folded substrates and unfolded substrates. Assuming binding and unbinding equations and the folding and unfolding transition rates are in a quasi-equilibrium then eq 23, 24, and 25 hold throughout the time-course.

We write an equation for the overall product formation:

$$\frac{d([P_F]+[P_U])}{dt} = k_f[FE] + k_u[UE] \quad (26)$$

Conservation equations for substrate-product and enzyme in this case are:

$[S_0] = [FE] + [UE] + [F] + [U] + [P_F] + [P_U]$ (27) where $[S_0]$ is initial and total concentration of substrate

$[E_0] = [E] + [FE] + [UE]$ (28) where $[E_0]$ is the initial concentration of enzyme.

<u>Step 1: Write product formation eq. 26 in terms of [FE] and constants only, by substituting for [UE] complex.</u>

$[UE] = 1/K_{50,U}[U][E]$ (use eq. 24)

$= 1/K_{50,U} * K_{UF}[F][E]$ (use eq. 25)

$$= \frac{K_{50,F}K_{UF}}{K_{50,U}}[FE] \text{ (29)} \qquad \text{(use eq. 23)}$$

$$\frac{d([P_F]+[P_U])}{dt} = k_f[FE] + k_u[UE] = (k_f + \frac{K_{50,F}K_{UF}}{K_{50,U}}k_u)[FE] \text{ (30, use eq. 26)}$$

<u>Step 2: Replace [FE] dependence with ([S_0] - [P_F] - [P_u]) dependence using conservation laws</u>

$$[S_0] = [FE] + [UE] + [F] + [U] + [P_F] + [P_U] \text{ (27)}$$

$$= [FE] + \frac{K_{50,F}K_{UF}}{K_{50,U}}[FE] + \frac{K_{50,F}}{[E]}[FE] + \frac{K_{UF}K_{50,F}}{[E]}[FE] + [P_F] + [P_U]$$

(use eq. 29, 23, and 23+25)

$$= [FE](1 + \frac{K_{50,F}K_{UF}}{K_{50,U}} + \frac{K_{50,F}(1+K_{UF})}{[E]}) + [P_F] + [P_U]$$

Thus, we get an equation which describes the dependence of [FE] on initial substrates and products, with terms in the denominator that capture sequestration in intermediate bound states.

$$[FE] = ([S_0] - [P_F] - [P_U])/(1 + \frac{K_{50,F}K_{UF}}{K_{50,U}} + \frac{K_{50,F}(1+K_{UF})}{[E]})$$

$$= \frac{(S_0 - [P_F] - [P_U])[E]K_{50,U}}{[E]K_{50,U} + K_{50,F}K_{UF}[E] + K_{50,U}K_{50,F}(1+K_{UF})}$$

$$= \frac{(S_0 - [P_F] - [P_U])[E]K_{50,U}}{[E](1/K_{50,F} + K_{UF}/K_{50,U}) + 1 + K_{UF}} \text{ (31)}$$

Substituting this into the product formation equation:

$$\frac{d([P_F]+[P_U])}{dt} = (k_f + \frac{K_{50,F}K_{UF}}{K_{50,U}}k_u)[FE] \text{ (30)}$$

$$= (k_f + \frac{K_{50,F}K_{UF}}{K_{50,U}}k_u)\frac{(S_0 - [P_F] - [P_U])[E]K_{50,U}}{[E](1/K_{50,F} + K_{UF}/K_{50,U}) + 1 + K_{UF}} \text{ (use eq. 31)}$$

Then, we defined $[P_{total}] = [P_F] + [P_U]$

$$\frac{d([P_{total}])}{dt} = (k_f + \frac{K_{50,F}K_{UF}}{K_{50,U}}k_u)\frac{[E]K_{50,U}}{[E](1/K_{50,F} + K_{UF}/K_{50,U}) + 1 + K_{UF}}([S_0] - [P_{total}])$$

$$= \frac{k_f/K_{50,F} + K_{UF}k_u/K_{50,U}}{1/K_{50,F} + K_{UF}/K_{50,U}} \frac{[E]}{[E] + (1+K_{UF})/(1/K_{50,F} + K_{UF}/K_{50,U})}([S_0] - [P_{total}])$$

Because the reaction conditions in the study were not substrate-excessive but enzyme-excessive (i.e. [E] >> [S] or [ES]), [E]≈[E_0]:

$$\frac{d([P_{total}])}{dt} = \frac{k_f/K_{50,F} + K_{UF}k_u/K_{50,U}}{1/K_{50,F} + K_{UF}/K_{50,U}} \frac{[E_0]}{[E_0] + (1+K_{UF})/(1/K_{50,F} + K_{UF}/K_{50,U})}([S_0] - [P_{total}]) \text{ (32)}$$

Finally, We can rewrite the product formation eq. 3 in terms of initial substrate concentration, total product, and an observed kinetic rate, which is a function of kinetic rates and initial enzyme concentration,:

$$\frac{d([P_{total}])}{dt} = k_{obs} * ([S_0] - [P_{total}]) = \frac{k_{max}[E_0]}{K_{50} + [E_0]} * ([S_0] - [P_{total}]) \quad (3')$$

<u>Step 3, Derivation eq.6 and eq.7 in Fig. 1B</u>

By comparing eq. 32 with eq. 3', we can derive the following equations (including eq. 6 in Fig. 1B):

$$k_{max} = \frac{k_f/K_{50,F} + K_{UF}k_u/K_{50,U}}{1/K_{50,F} + K_{UF}/K_{50,U}} \quad (33)$$

$$K_{50} = \frac{1+K_{UF}}{1/K_{50,F} + K_{UF}/K_{50,U}} \quad (34)$$

$$= \frac{1 + [U]/[F]}{[FE]/[E][F] + ([U]/[F])*([UE]/[E][U])}$$

$$= \frac{1 + [U]/[F]}{[FE]/[E][F] + [UE]/[E][F]}$$

$$= \frac{([F] + [U])[E]}{[FE] + [UE]} \quad \text{(This is eq. 6)}$$

Using eq. 34 to rewriting a formula for K_{UF} in terms of the half-max reaction rates:

$$K_{50}(1/K_{50,F} + K_{UF}/K_{50,U}) = 1 + K_{UF}$$

$$K_{UF}(K_{50}/K_{50,U} - 1) = 1 - K_{50}/K_{50,F}$$

Thus, eq.7 which gives the ratio of unfolded to folded substrate is derived:

$$\frac{[U]}{[F]} = K_{UF} = \frac{1/K_{50} - 1/K_{50,F}}{1/K_{50,U} - 1/K_{50}} \quad \text{(This is eq. 7)}$$

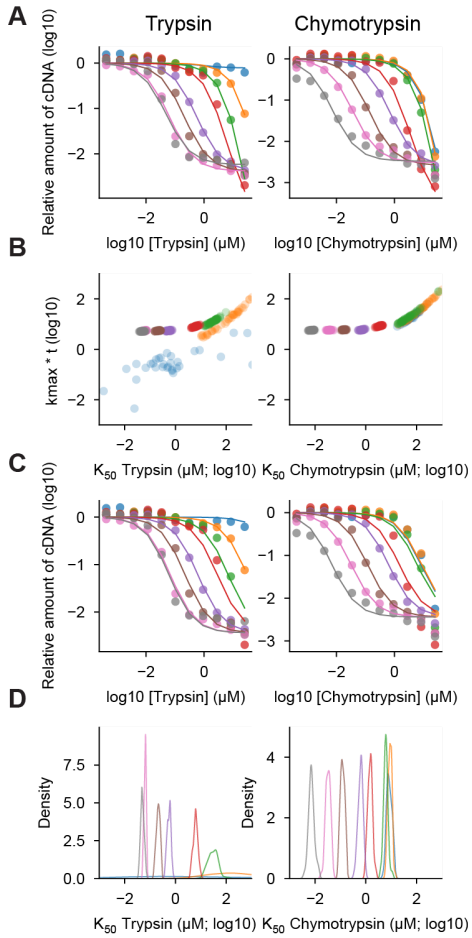


Fig. S1. Single turnover model fitting on qPCR data

(A) To test the single turnover model, we performed cDNA display proteolysis on a mixture of eight mini protein sequences with diverse folding stability and quantified the surviving amount of each cDNA using qPCR. We then each curve one at a time by Bayesian inference using the single turnover kinetics model in Fig. 1B. We sampled k_{max}*t and K_{50} for each sequence. Dots represent the observed cDNA amount quantified by qPCR and lines show the two-parameter fits.

(B) Posterior distributions of k_{max}*t and K_{50} for eight proteins were shown. Whereas K_{50} values vary between different proteins, k_{max}*t values (indicating saturation at high protease concentrations) were either constant or unconstrained by the data.

(C) Based on the analysis (B), we fixed k_{max}*t at $10^{0.65}$ and re-sampled K_{50} for each protein. Dots represent the observed cDNA amount quantified by qPCR (same as in (A)) lines show the one-parameter fits.

(D) Posterior distributions of K_{50}. For trypsin, the K_{50} values for the two most stable proteins (orange and blue) could not be defined because they were too stable and outside of the dynamic range of this proteolysis assay.

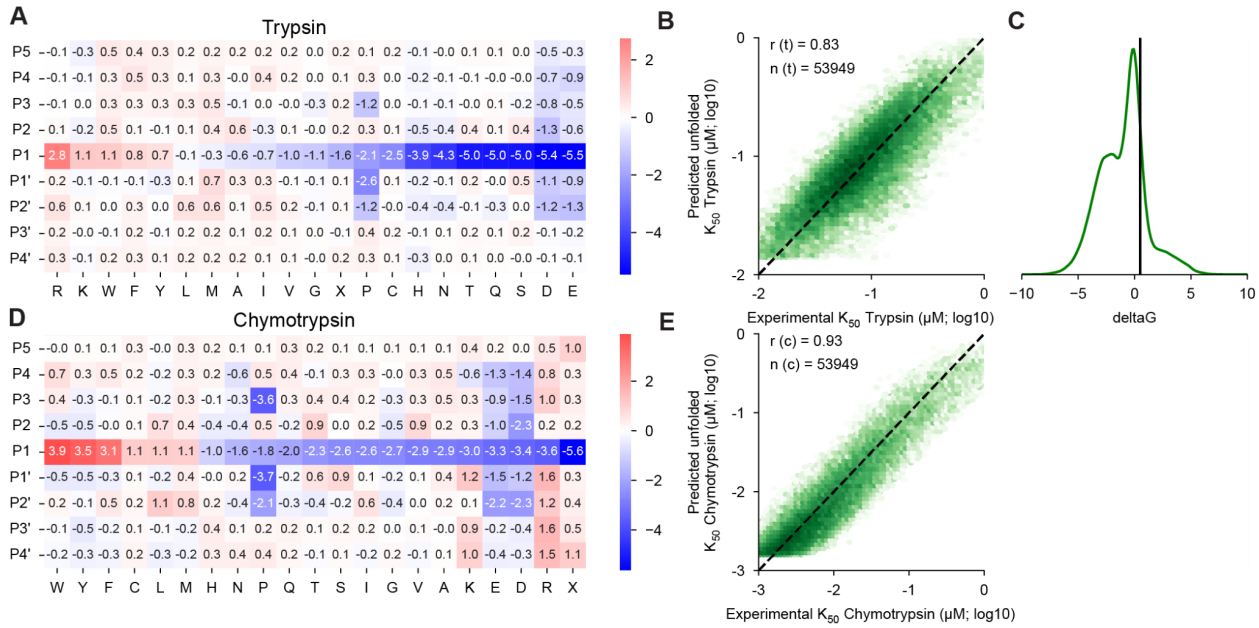


Fig. S2. Unfolded state model parameters and goodness of fit.

(A) Fit parameters for the unfolded state model position-specific scoring matrix (PSSM) for trypsin. The mean of all coefficients (-0.4) was subtracted from the values in the figure to aid visualization. Positive values indicate faster proteolysis and lower predicted $K_{50,U}$ values. By using different prior distribution widths for different rows during fitting, we guided the strongest rate determinants into the center row of each matrix, which we label "P1" (the assay cannot actually identify the specific location of cutting). Overall, the heatmap resembles similar data as previously reported (Rocklin et al., 2017) and is consistent with known trypsin specificity determinants, including the preference for R/K at P1, the inhibitory effect of P, and the unfavorability of D and E (Pan et al., 2014).

(B) 2D-histogram showing the overall agreement between the trypsin model (predicted $K_{50,U}$, y-axis) and the data (experimental K_{50}, x-axis). Only scrambled sequences with inferred $\Delta G <$ 0.5 kcal/mol (where we can assume $K_{50} \approx K_{50,U}$) are shown (53,949 out of 64,238 total sequences used in training). The Pearson r value is shown.

(C) Overall distribution of inferred ΔG of all scramble sequences. The vertical line represents 0.5 kcal/mol, which is a threshold used in (B).

(D, E) As above, for chymotrypsin. As in our previous report (Rocklin et al., 2017), the coefficients resemble established features of chymotrypsin specificity, including the preference for F/Y/W followed by M/L at P1, the inhibitory effect of P at P3, P1', and P2', and the general unfavorability of D and E (Schellenberger et al., 1991, 1993, 1994). The mean of all coefficients (-0.5) was subtracted from the values in the figure to aid visualization.

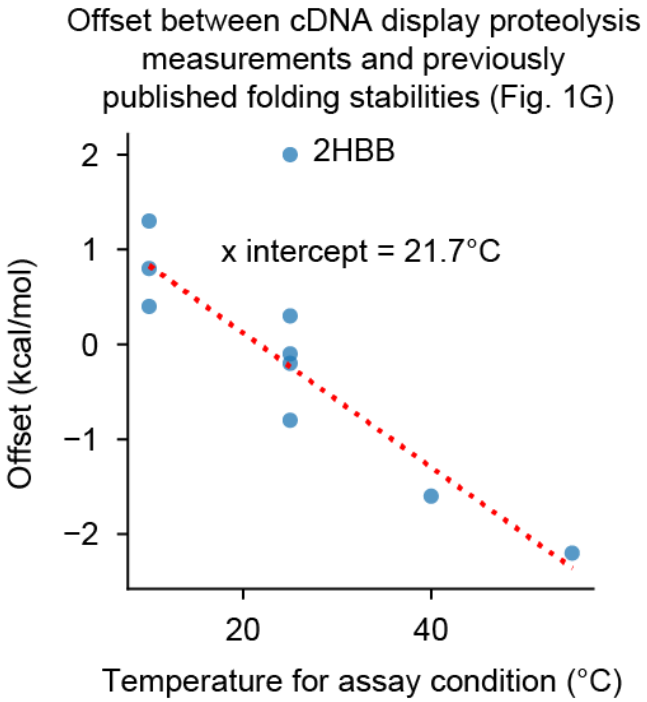


Fig. S3. Relationship between offset in Fig. 1G and assay temperature
Previous studies shown in Fig. 1G used diverse conditions including buffer, pH, ion strength, and temperature (see Table. S2) (Araya et al., 2012; Dodson and Arbely, 2015; Jäger et al., 2009; Jemth et al., 2005; Jiang et al., 2001; Maxwell and Davidson, 1998; Neuweiler et al., 2009; Nisthal et al., 2019; Northey et al., 2002; de los Rios et al., 2005; Sato et al., 2017; Villegas et al., 1998; Xiao et al., 2009, 2013). However, our measurements were all conducted in PBS at room temperature (approximately 22°C). In general, the offsets observed in Fig. 1G are correlated to the temperatures used in the previous studies, suggesting that the assay temperature is the main cause of the offsets. The red line represents a best fit line after removing the 2HBB point. The x-intercept (21.7°C) is close to our assay condition (approximately 22°C). 2HBB (the N-terminal domain of Ribosomal Protein L9) is an outlier and not included in the linear fit; the origin of the offset here is currently unknown.

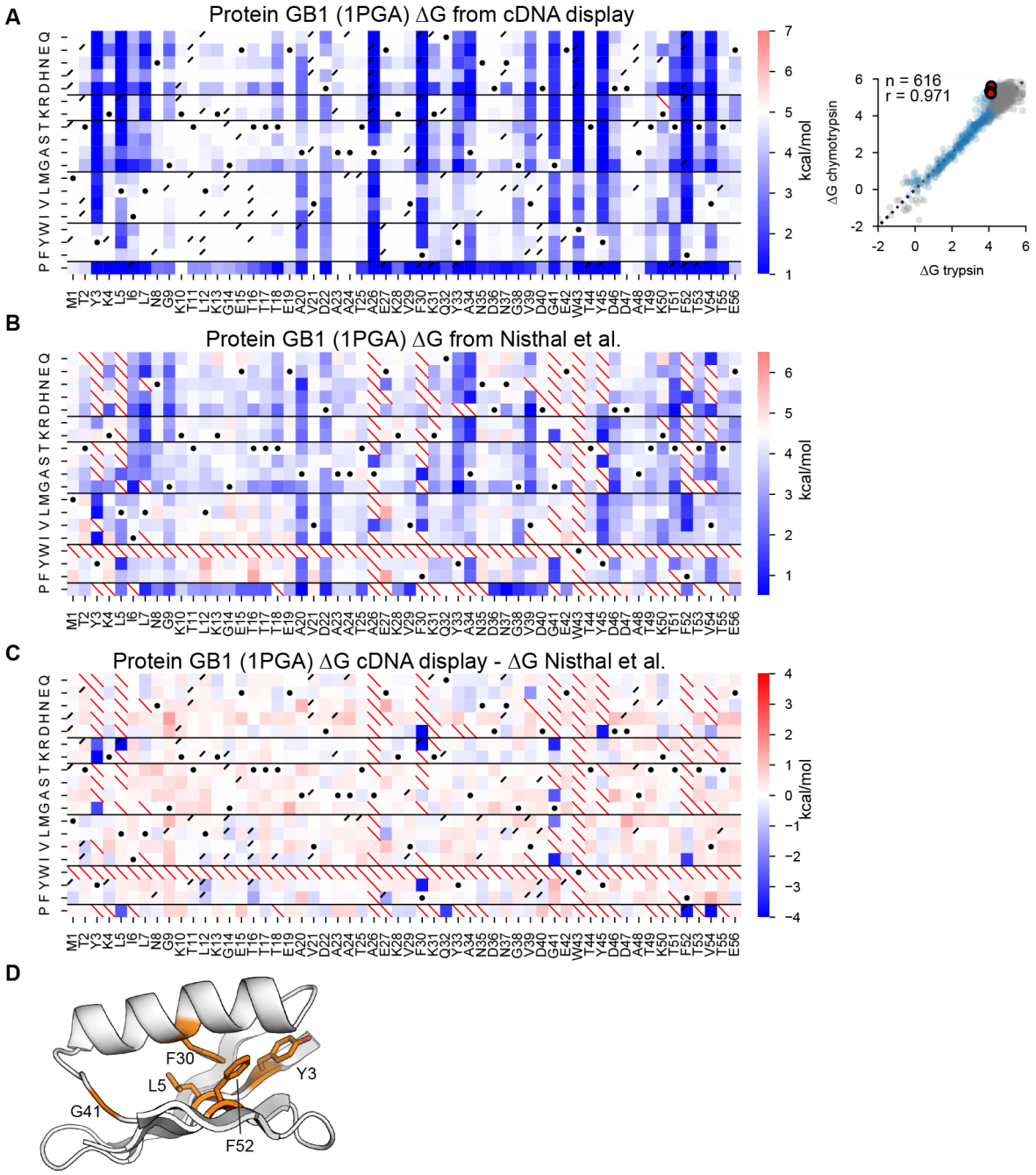


Fig. S4. Folding stability discrepancies between cDNA display proteolysis and previous measurements on Protein GB1

(A) Left: Mutational scanning results from cDNA display proteolysis. As in Fig. 2, white represents the folding stability of wild-type and red/blue indicates stabilizing/destabilizing mutations. Black dots indicate the wild-type amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5

kcal/mol), including ΔG estimates near the edges of the dynamic range. Right: Agreement between variant ΔG values independently determined using assays with trypsin (x-axis) and chymotrypsin (y-axis). Multiple codon variants of the wild-type sequence are shown in red, reliable ΔG values in blue, and less reliable ΔG estimates (same as above) in gray. The black dashed lines represent Y=X. Each plot shows the number of reliable points and the Pearson r-value.

(B) Mutational scanning results from robotics-enabled high-throughput purification and chemical denaturation (Nisthal et al., 2019), colored as in (A).

(C & D) Difference heat-map (C) showing the consistency between cDNA display proteolysis (A) and robotics-enabled high-throughput purification and chemical denaturation (B). Dark blue squares indicate highly inconsistent positions where cDNA display proteolysis (A) observes low stability but robotics-assisted chemical denaturation (B) observes high stability. These positions are mainly located in the protein core (shown in D). We hypothesize that many of the inconsistent variants are actually very unstable (as shown by cDNA display proteolysis, A), leading to poorly expressed protein samples that appeared stable in (B) due to the lack of a clear melting signal in chemical denaturation. This would also explain the inconsistencies between closely related mutants seen in (B). For example, the published chemical denaturation data show that Y3R is poorly expressed (no data) along with many other variants at Y3, yet Y3K is measured as very stable (both Y3R and Y3K are unstable in cDNA display proteolysis). The same pattern is seen at L5 in the core: the published data shows that L5K is poorly expressed (no data), yet L5R is measured as very stable (again both L5K and L5R appear unstable in cDNA display proteolysis). The same biophysically inconsistent patterns appear at F30, Q32 (can Q32P in the middle of the helix really be as stable as wild-type?), G41, Y45, F52, and V54. These biophysical inconsistencies at sites where many variants are poorly expressed in the published data suggest to us that the cDNA display proteolysis measurements are more likely to be correct.

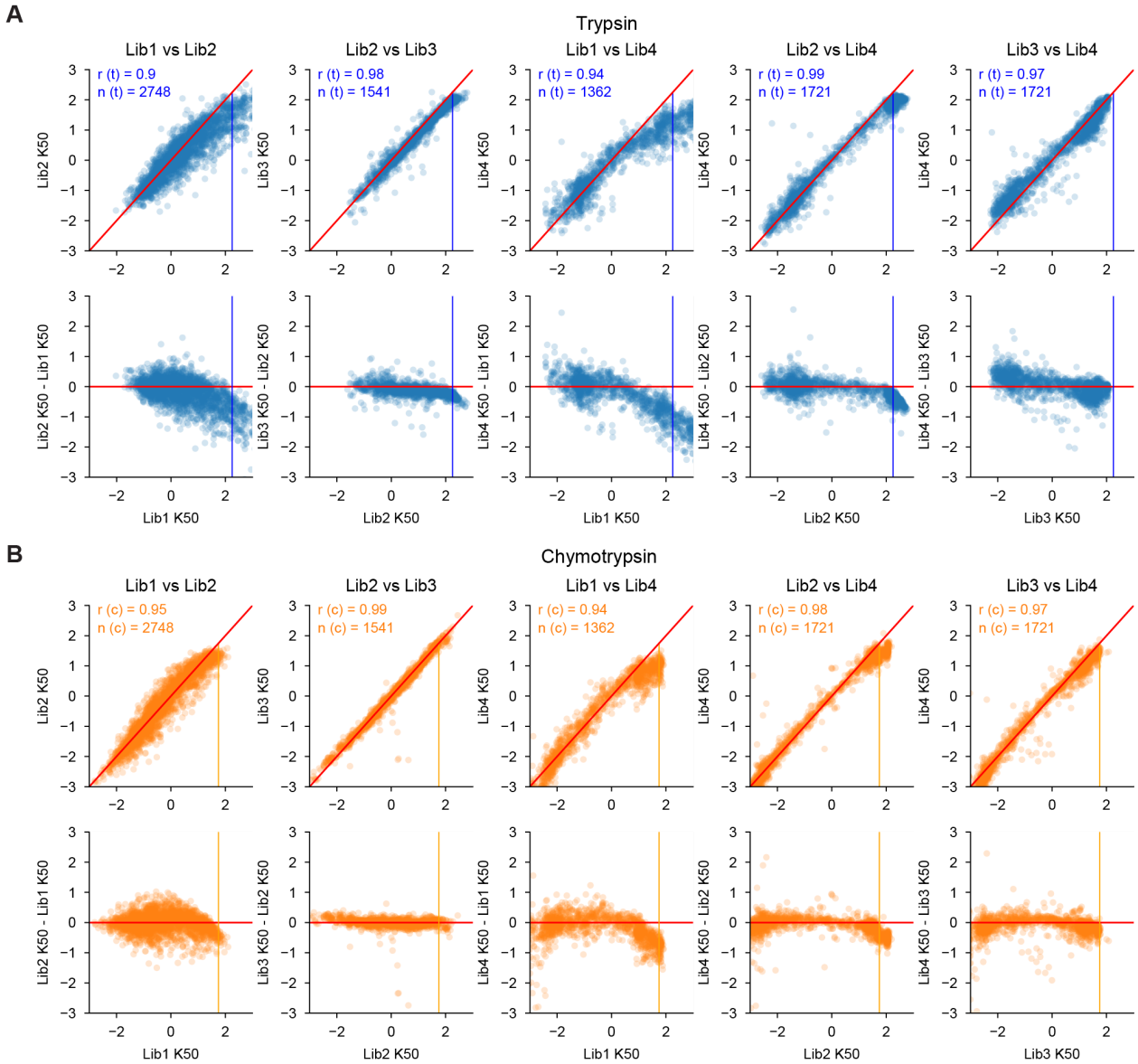


Fig. S5. Consistency of K_{50} measurements across libraries

(A and B) To examine the consistency between K_{50} (μM) values measured in different libraries, we included identical sequences (potentially with different padding at the termini) in multiple libraries. For each pair of libraries with overlapping sequences, we show the K_{50} values for those sequences in both libraries for trypsin (A) and chymotrypsin (B). The top row shows raw K_{50} values for overlapping sequences in each library; the second row shows the difference in K_{50} estimates plotted against the K_{50} in one of the libraries. The red diagonal line shows Y=X in the top row and Y=0 (i.e. identical K_{50} estimates) in the bottom row. Blue/orange vertical lines show $K_{50,F}$; all K_{50} values above $K_{50,F}$ are treated as equivalent. Each plot is annotated at the top-left with the total number of overlapping sequences and Pearson r-value between the libraries.

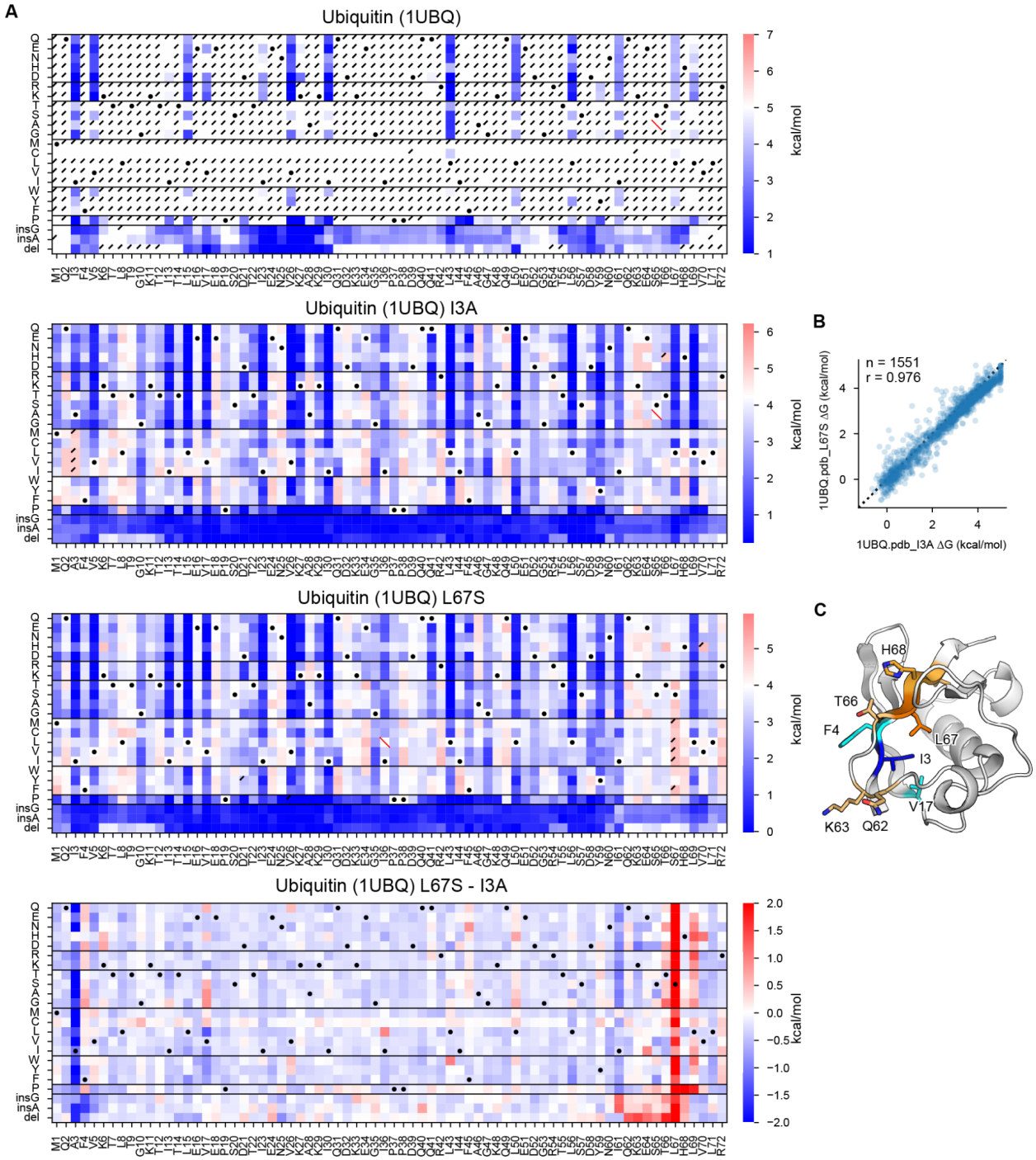


Fig. S6. Heat maps for a stable domain (Ubiquitin; 1UBQ) and its destabilizing mutants
(A) Mutational scanning results for human erythrocytic ubiquitin (1UBQ) and its destabilizing mutant backgrounds (I3A and L67S). Heat maps show the ΔG of wild-type ubiquitin (top), ubiquitin I3A (middle-top), ubiquitin L67S (middle-bottom), and the difference (ΔΔG) between two mutant backgrounds (bottom) for substitutions, deletions, and Gly and Ala insertions at each residue. In the three ΔG heat maps, white represents the folding stability of the wild-type and red/blue indicates stabilizing/destabilizing mutations. Black dots indicate the background

(wild-type or mutant) amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5 kcal/mol), including ΔG estimates near the edges of the dynamic range.

(B) Consistency between mutant stabilities measured in the I3A background (x-axis) and L67S (y-axis) background. The plot is annotated with the number of points and the Pearson r value.

(C) Ubiquitin structure highlighting the mutant points (I3 and L67) and the residues with a different effect on stability between two mutational backgrounds.

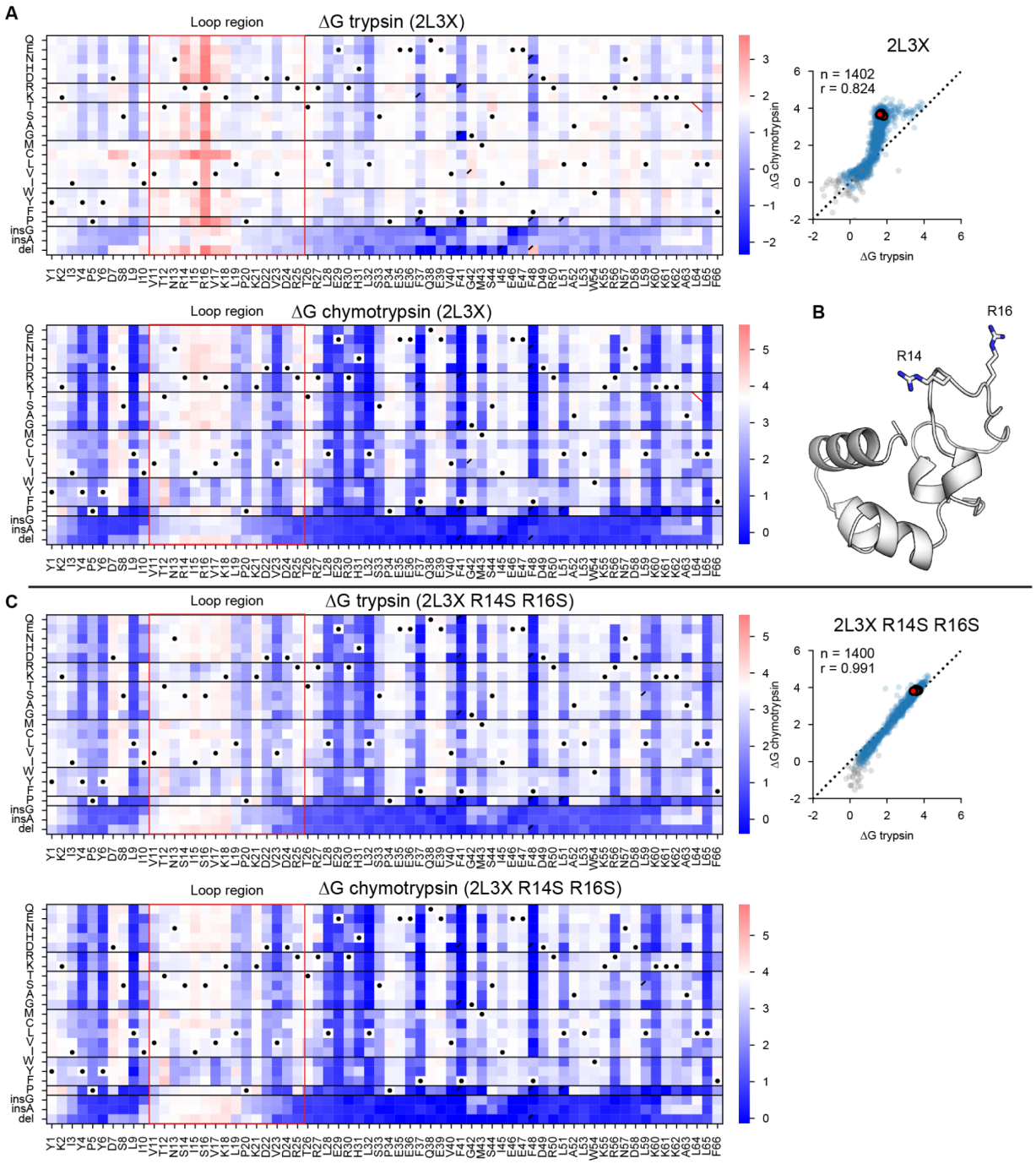


Fig. S7. Heat maps for one domain with trypsin cleavage sites in loop region

(A) Mutational scanning results for 2L3X, which includes trypsin cleavage sites in the loop region. Left: Heat maps show the ΔG trypsin (top) and chymotrypsin challenge (bottom) for substitutions, deletions, and Gly and Ala insertions at each residue, with our one-indexed numbering at the bottom. Black dots indicate the wild-type amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5 kcal/mol), including ΔG estimates near the edges of the dynamic range. The

colored boxes highlight sites in the flexible loop region. Right: Replicates of the wild-type sequence are shown in red, reliable ΔG values in blue, and less reliable ΔG estimates (same as above) in gray. The black dashed lines represent Y=X. Each plot shows the number of reliable points and the Pearson r-value. The dots show a reverse 'L' shape due to the cleavage of the flexible loop region in the trypsin challenge.

(B) 2L3X structure highlighting Args in the loop region (R14 and R16).

(C) Same as (A) for 2L3X with replacement of Arg in the loop (R14 and R16) with Ser. In this deep mutational scanning, we observed higher consistency between trypsin and chymotrypsin challenges because we removed sites that could be cleaved in the folded state.

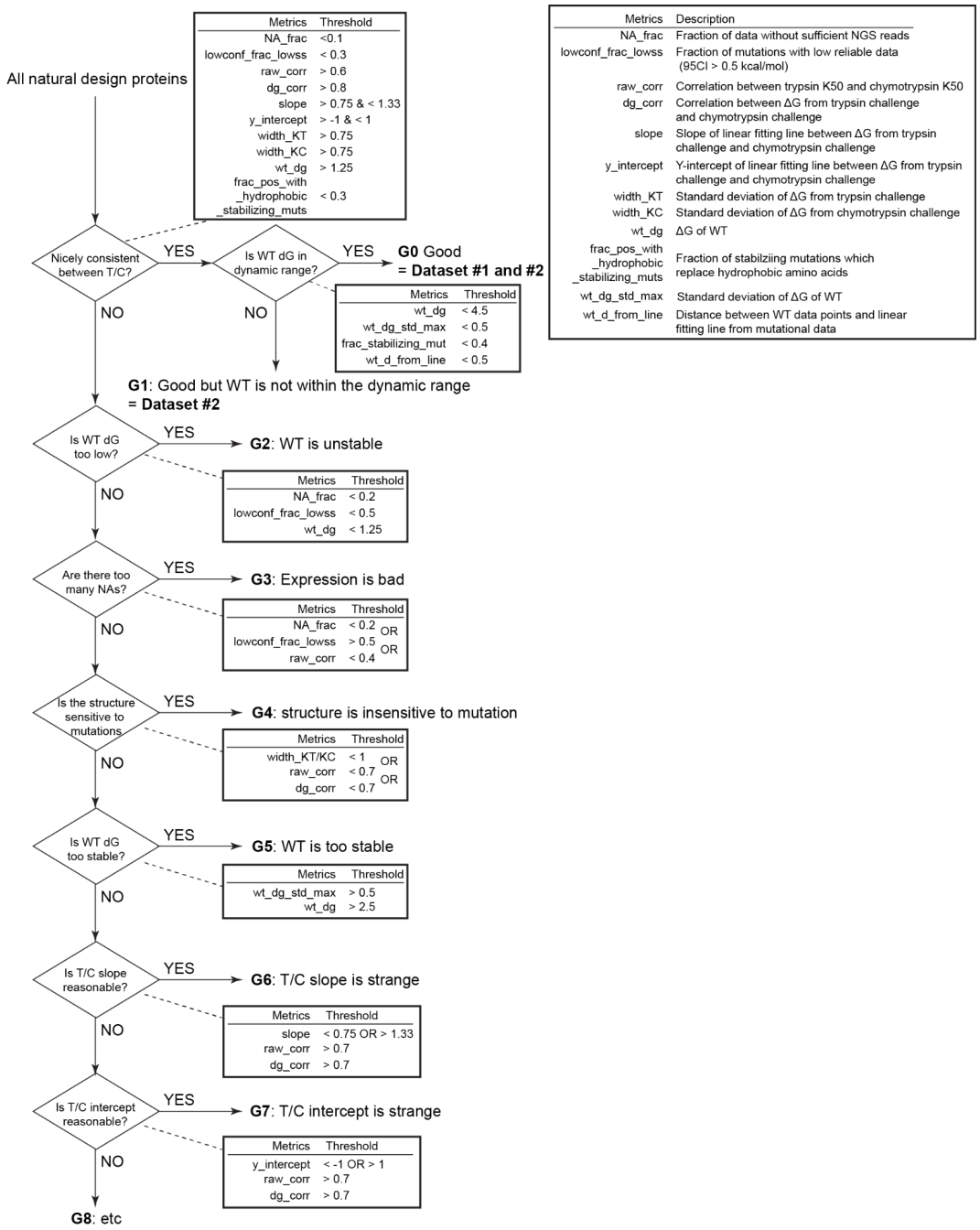


Fig. S8. Classification of deep mutational scanning results

We classified all deep mutational scanning results into nine groups shown in Fig. 2B. Here, we show the classification criteria. The description of all metrics is also included in Table. S3, and the metrics of all domains for the classification are included in Single_DMS_list.csv.

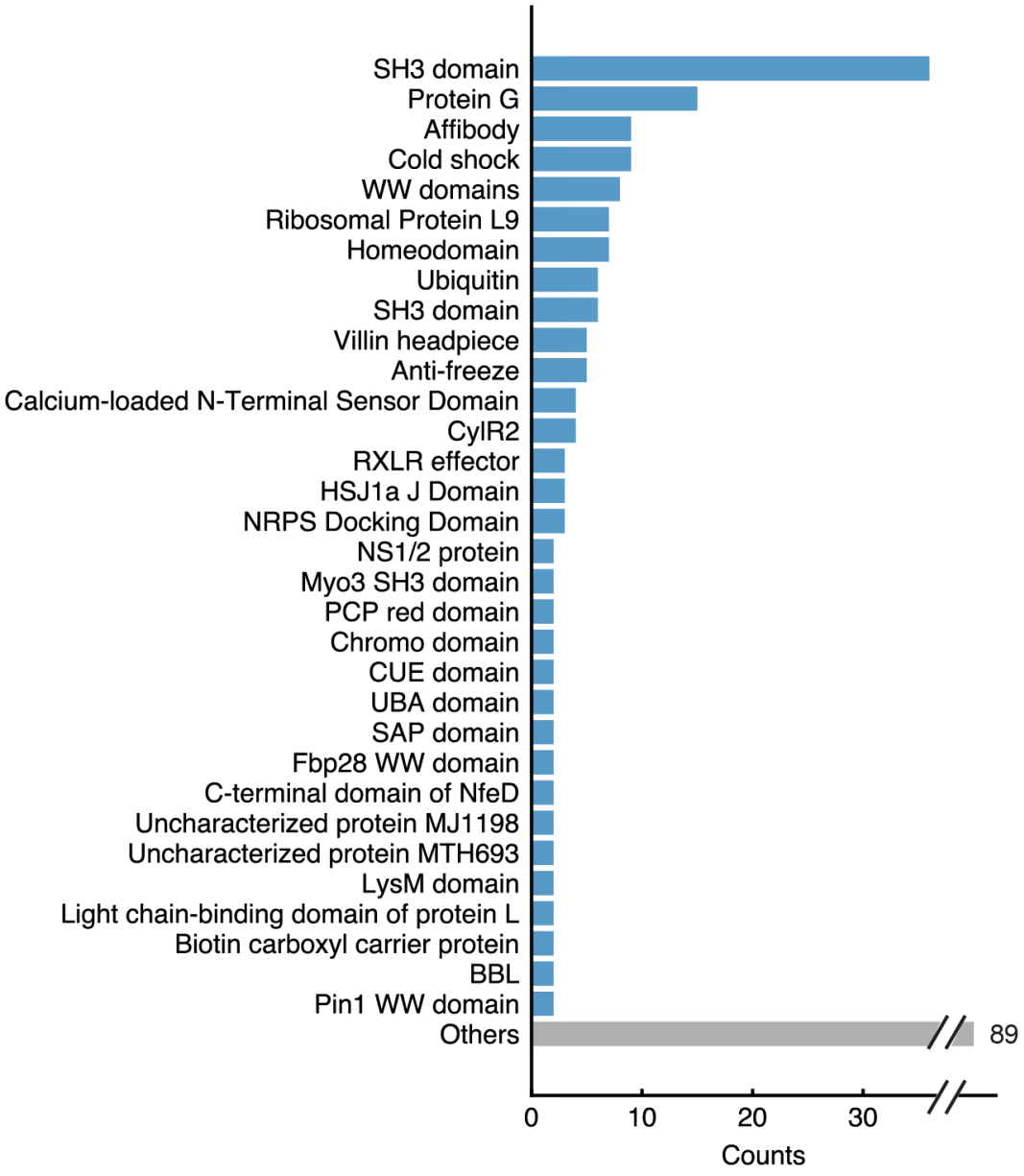


Fig. S9. Classification of the natural protein domains investigated in cDNA display proteolysis

Comprehensive group list of wild-type structures classified as G0 in Fig. 2B grouped into domain families.

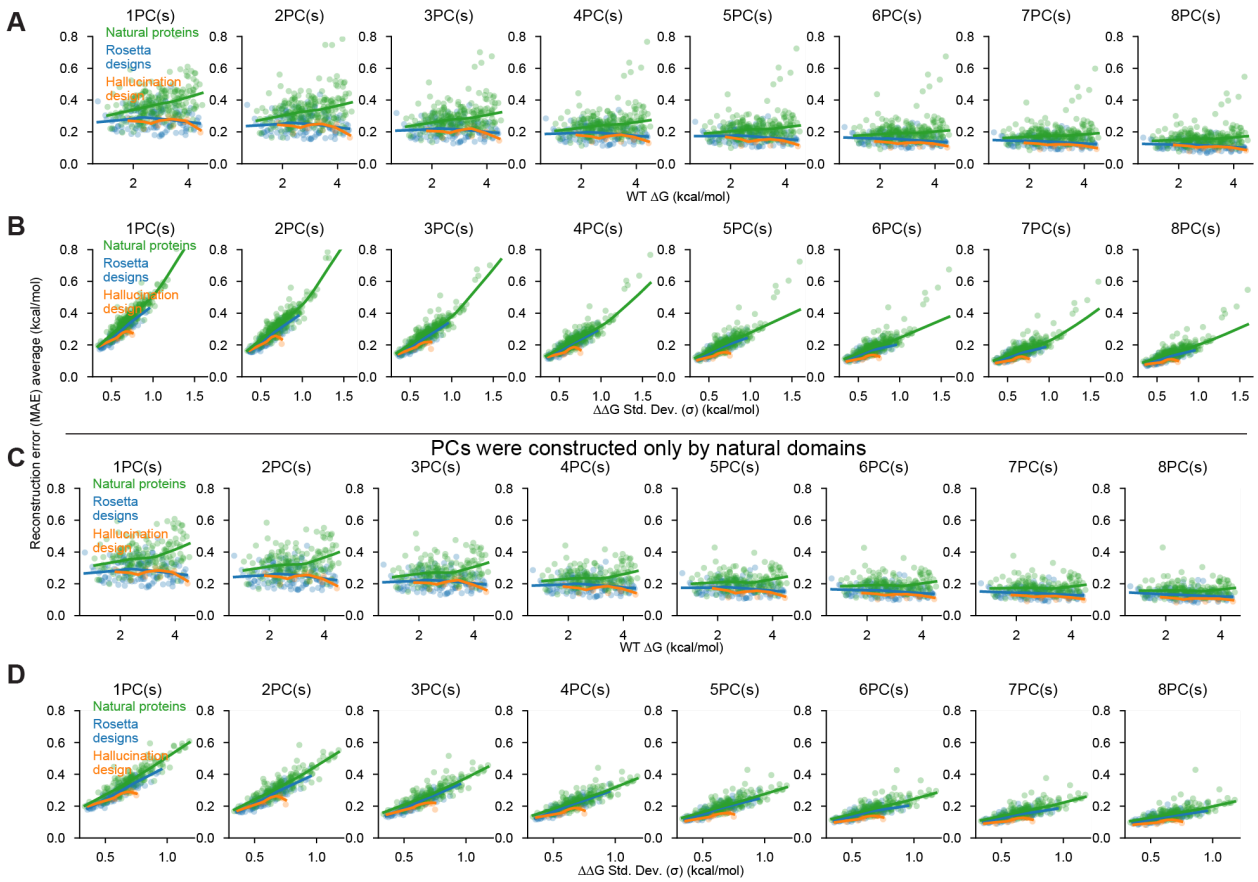


Fig. S10. PC analysis with different numbers of PCs

(A and B) Relationship between reconstruction error (MAE) using 1-8 PCs (MAE, y-axis) and wild-type stability (A, x-axis) or variance in the ΔΔG data (B, x-axis). Colors represent protein structures grouped into natural proteins (green), Rosetta designs (blue), and hallucination designs (orange). Lines show LOESS fits to the data.

(C and D) Same as (A) and (B). PCs were constructed using only natural domains. The overall tendency is the same as (A) and (B).

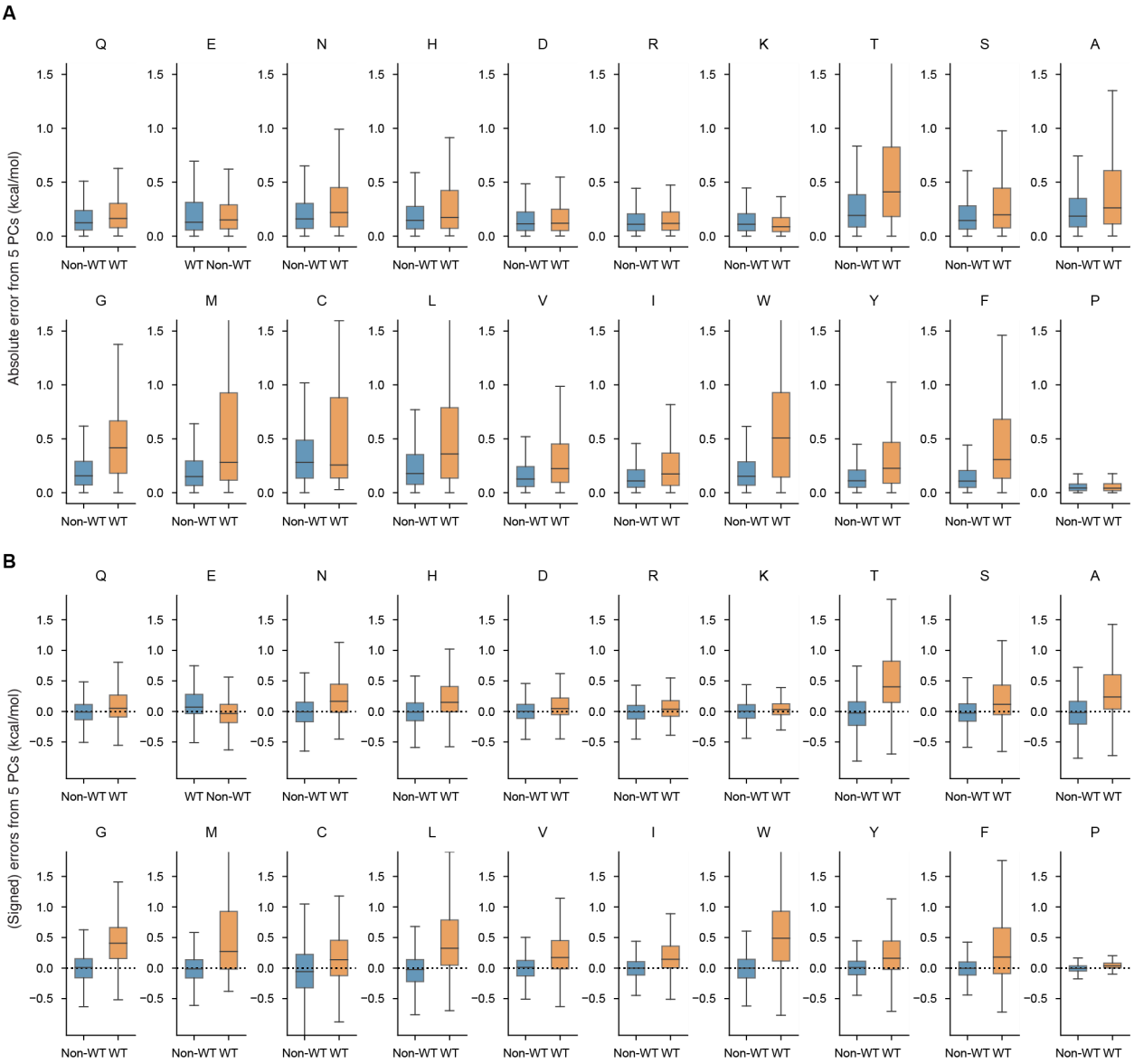


Fig. S11. Errors between observed ΔG and reconstructed ΔG by five PCs for wild-type (WT) or non-WT residues for 20 amino acids.

Absolute (A) and signed (B) errors between observed ΔG and reconstructed ΔG (from five PCs) are shown. Wild-type residues tend to show larger (A), more positive errors (B), meaning that the five-PC model underestimates wild-type stabilities.

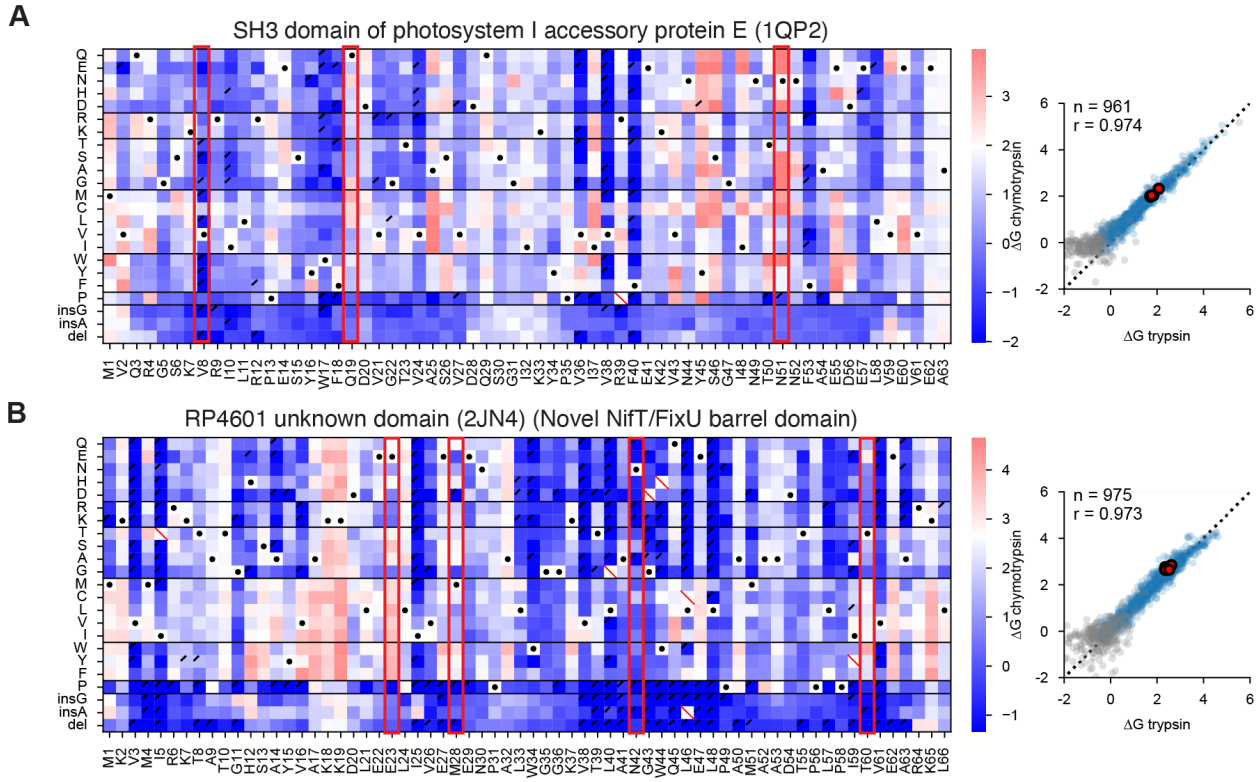
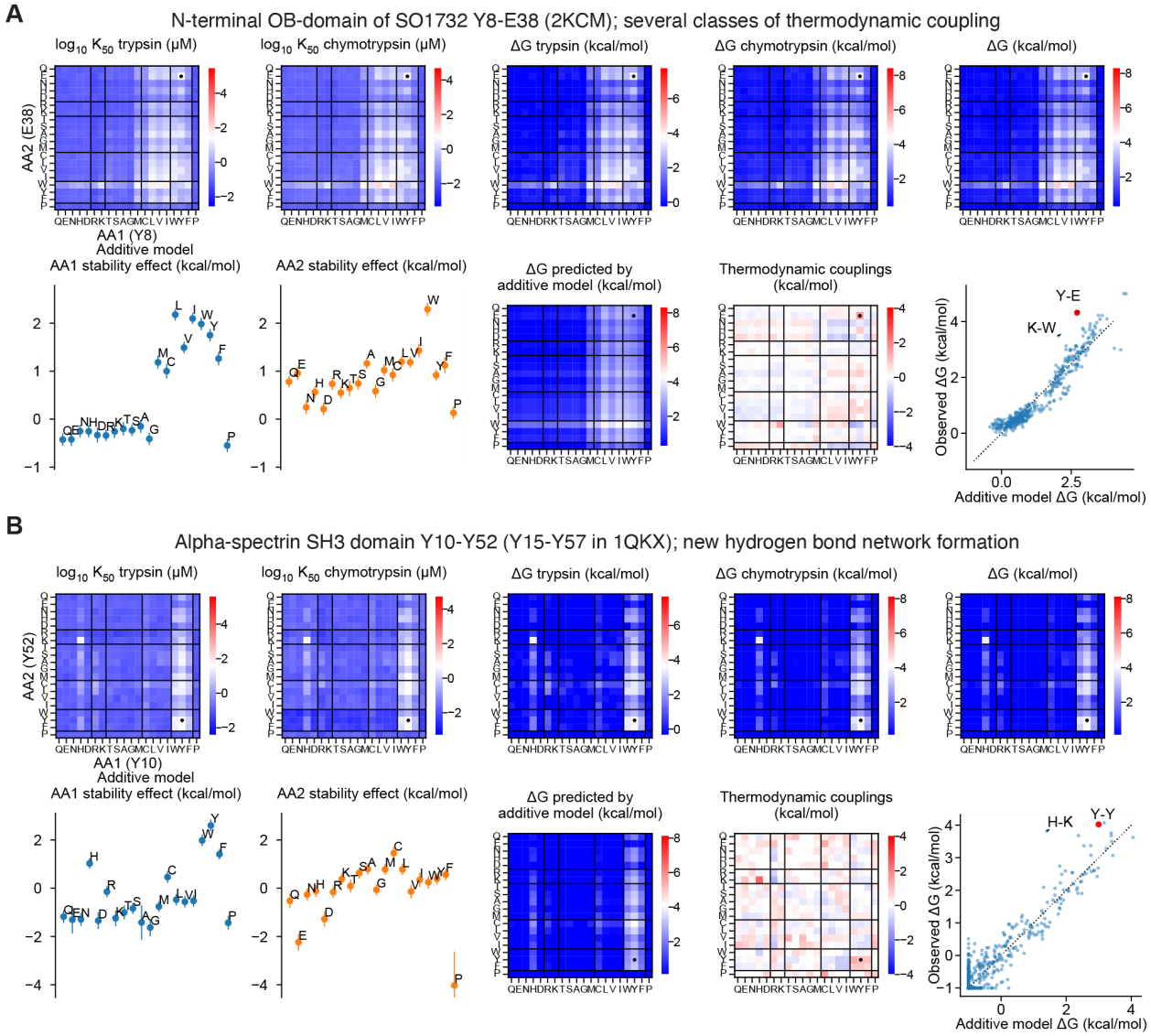
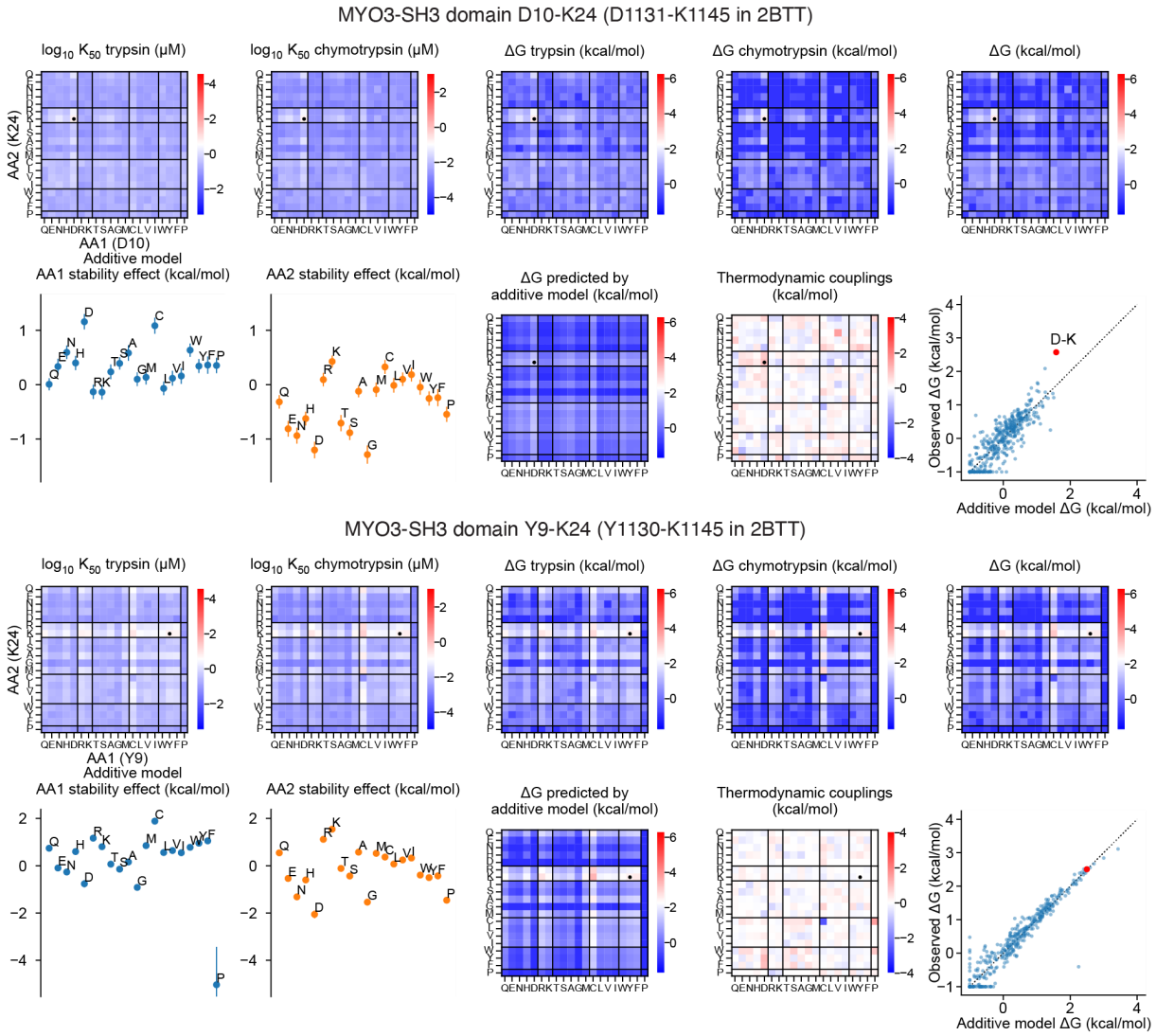


Fig. S12. Heatmaps for the notable domains with large errors between observed ΔG and reconstructed ΔG by five PCs.

Mutational scanning results for the two notable domains described in Fig. 3G. Left: Heat maps show the ΔG for substitutions, deletions, and Gly and Ala insertions at each residue, and our one-indexed numbering at the bottom. Black dots indicate the wild-type amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5 kcal/mol), including ΔG estimates near the edges of the dynamic range. Red boxes highlight the seven notable residues with large MAE described in Fig. 3H. Right: Agreement between variant ΔG values independently determined using assays with trypsin (x-axis) and chymotrypsin (y-axis). Multiple codon variants of the wild-type sequence are shown in red, reliable ΔG values in blue, and less reliable ΔG estimates (same as above) in gray. The black dashed lines represent Y=X. Each plot shows the number of reliable points and the Pearson r-value.


A
N-terminal OB-domain of SO1732 Y8-E38 (2KCM); several classes of thermodynamic coupling
log10 K50 trypsin (μM)
log10 K50 chymotrypsin (μM)
ΔG trypsin (kcal/mol)
ΔG chymotrypsin (kcal/mol)
ΔG (kcal/mol)
AA2 (E38)
AA1 (Y8)
Additive model
AA1 stability effect (kcal/mol)
AA2 stability effect (kcal/mol)
ΔG predicted by additive model (kcal/mol)
Thermodynamic couplings (kcal/mol)
Y-E
K-W
Observed ΔG (kcal/mol)
Additive model ΔG (kcal/mol)
B
Alpha-spectrin SH3 domain Y10-Y52 (Y15-Y57 in 1QKX); new hydrogen bond network formation
log10 K50 trypsin (μM)
log10 K50 chymotrypsin (μM)
ΔG trypsin (kcal/mol)
ΔG chymotrypsin (kcal/mol)
ΔG (kcal/mol)
AA2 (Y52)
AA1 (Y10)
Additive model
AA1 stability effect (kcal/mol)
AA2 stability effect (kcal/mol)
ΔG predicted by additive model (kcal/mol)
Thermodynamic couplings (kcal/mol)
H-K
Y-Y
Observed ΔG (kcal/mol)
Additive model ΔG (kcal/mol)

C


MYO3-SH3 domain D10-K24 (D1131-K1145 in 2BTT)
log10 K50 trypsin (μM)
log10 K50 chymotrypsin (μM)
ΔG trypsin (kcal/mol)
ΔG chymotrypsin (kcal/mol)
ΔG (kcal/mol)
AA2 (K24)
AA1 (D10)
Additive model
AA1 stability effect (kcal/mol)
AA2 stability effect (kcal/mol)
ΔG predicted by additive model (kcal/mol)
Thermodynamic couplings (kcal/mol)
D-K
Observed ΔG (kcal/mol)
Additive model ΔG (kcal/mol)
MYO3-SH3 domain Y9-K24 (Y1130-K1145 in 2BTT)
AA1 (Y9)

D

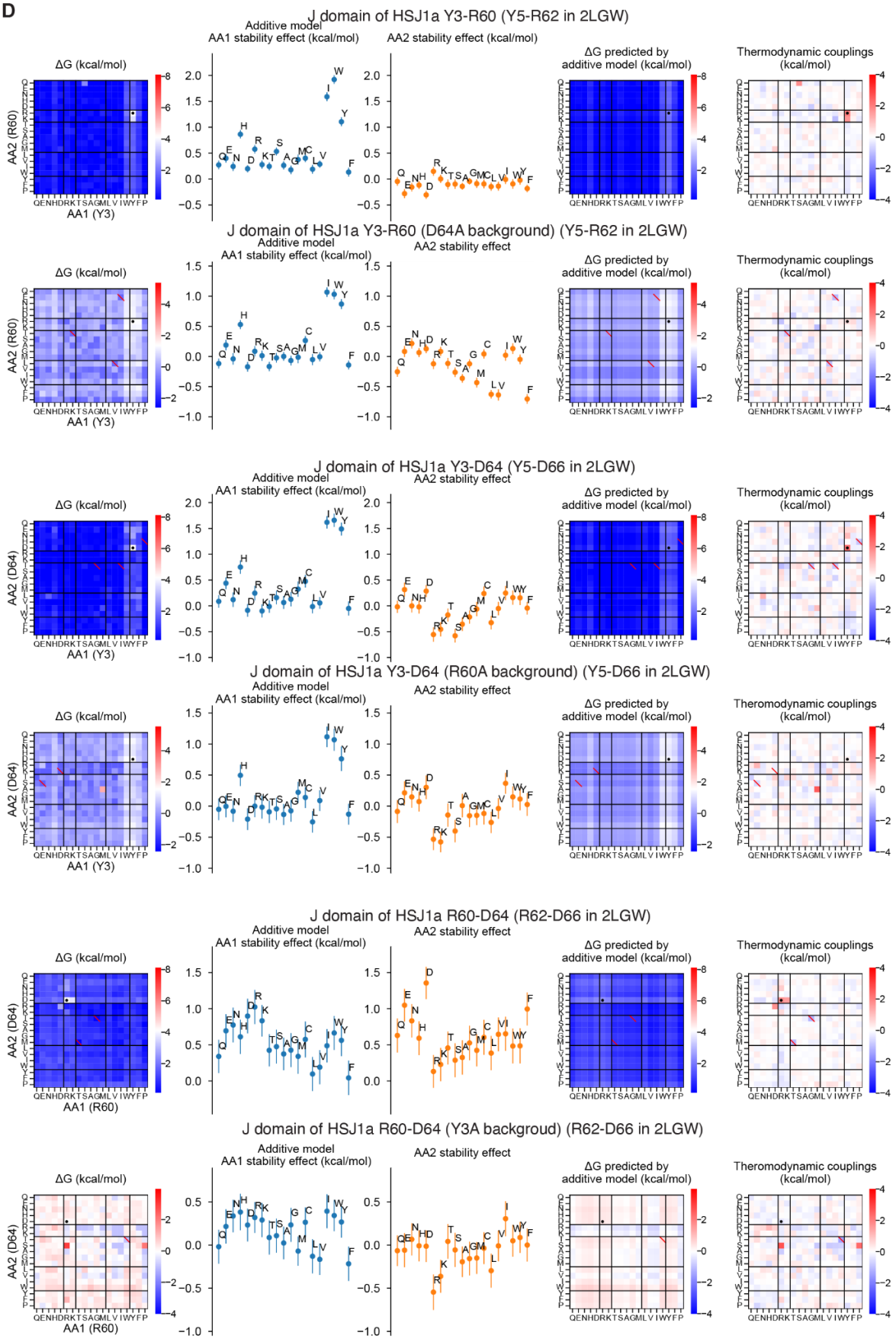

J domain of HSJ1a Y3-R60 (Y5-R62 in 2LGW)
Additive model
AA1 stability effect (kcal/mol)
AA2 stability effect (kcal/mol)
ΔG (kcal/mol)
ΔG predicted by additive model (kcal/mol)
Thermodynamic couplings (kcal/mol)
AA2 (R60)
AA1 (Y3)
J domain of HSJ1a Y3-R60 (D64A background) (Y5-R62 in 2LGW)
AA2 stability effect
J domain of HSJ1a Y3-D64 (Y5-D66 in 2LGW)
AA2 (D64)
J domain of HSJ1a Y3-D64 (R60A background) (Y5-D66 in 2LGW)
Theromodynamic couplings (kcal/mol)
J domain of HSJ1a R60-D64 (R62-D66 in 2LGW)
AA1 (R60)
J domain of HSJ1a R60-D64 (Y3A backgroud) (R62-D66 in 2LGW)

Fig. S13. Comprehensive double mutational data for the notable amino acid pairs

(A and B) Analysis of thermodynamic coupling for two notable amino acid pairs. In the first row, we show stabilities for all 20x20 double mutants according to five different experimental metrics. From left to right, we show trypsin K_{50}, chymotrypsin K_{50}, ΔG inferred from trypsin experiments, ΔG inferred from chymotrypsin experiments, and ΔG inferred from both sets of experiments together. In the second row, we show the results of the additive model. From left to right, the first two plots show the inferred single amino acid terms for all 20 amino acids in the first and second sites of the amino acid pair. Error bars represent the standard deviation of the posterior distributions. The middle heatmap shows stability (ΔG) for all amino acid pairs according to the additive model (the sum of the two single amino acid terms). The fourth plot shows the observed thermodynamic coupling; e.g. the experimental ΔG (rightmost plot in the first row) minus the prediction from the additive model (middle plot of the second row). The final scatter plot shows experimental stabilities for all double mutants (y-axis) plotted against the results from the additive model (x-axis).

(C) Same analysis as (A) and (B) for two site pairs in MYO3-SH3 domain (2BTT).

(D) Analysis of thermodynamic coupling for all amino acid pairs from a notable amino acid triple. The same amino acid substitutions were also performed for the mutant background with the third amino acid replaced by Ala. From left to right, we show the stabilities (ΔG) of all pairs of amino acids, the single amino acid terms in the additive model (error bars show the standard deviation of the posterior distribution), the stabilities for all pairs according to the additive model, and the thermodynamic coupling for all pairs of amino acids.

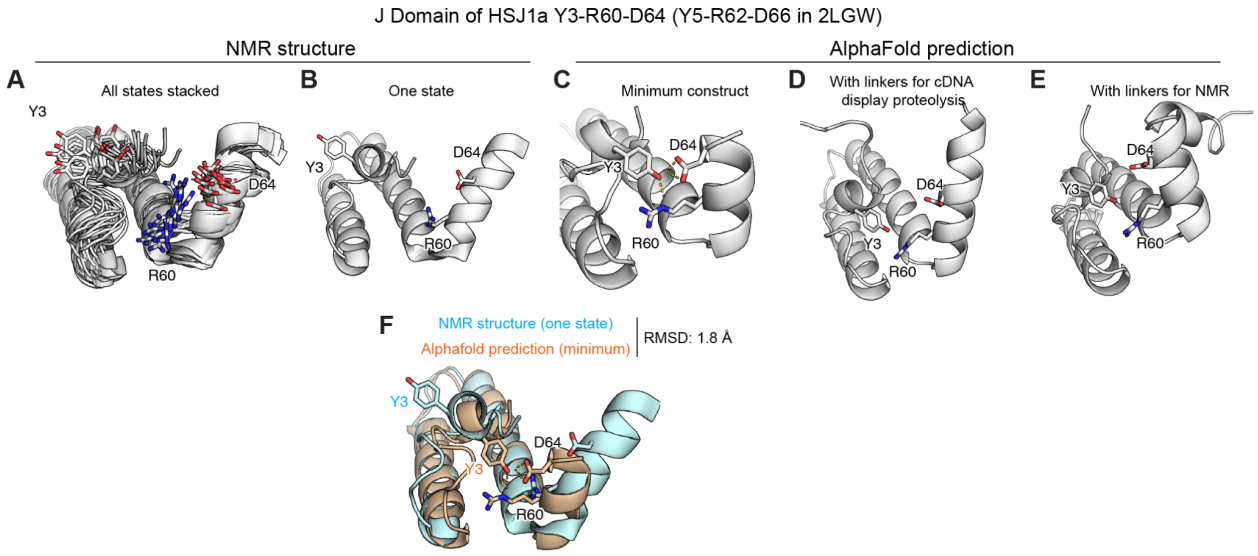


Fig. S14. Comparison of AlphaFold model and NMR structure for J domain of HSJ1a

Structure of J domain in HSJ1a (2LGW). We show NMR structure of all states stacked (A) and the first state (B), and AlphaFold predicted structures for the minimum construct (the variable segment in cDNA display) (C), the construct with linkers for cDNA display proteolysis (D), and the exact sequence used for NMR (E). In (F), we overlay the first state of the NMR ensemble (cyan) with the AlphaFold structure (orange) of the minimal construct.

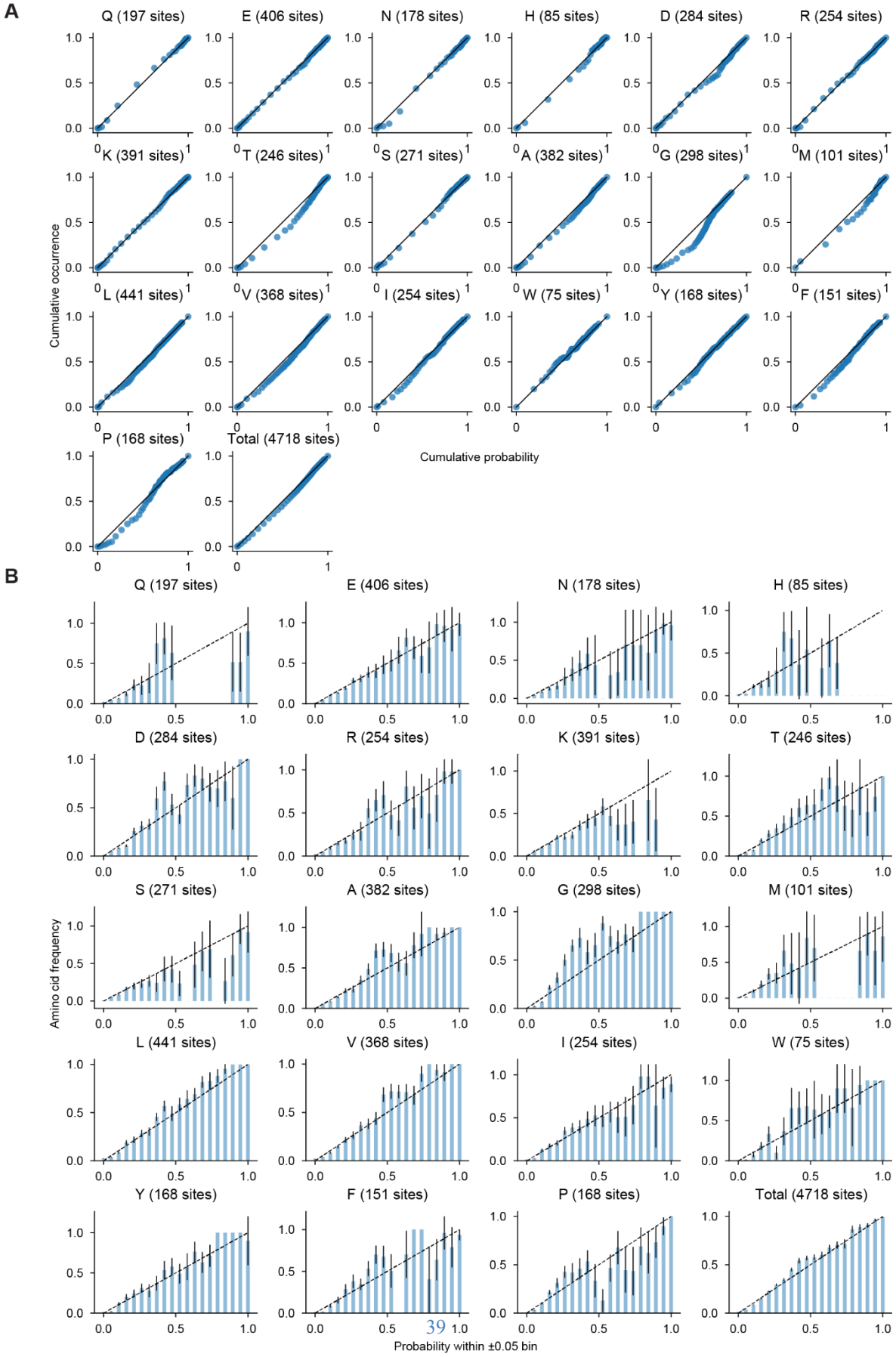
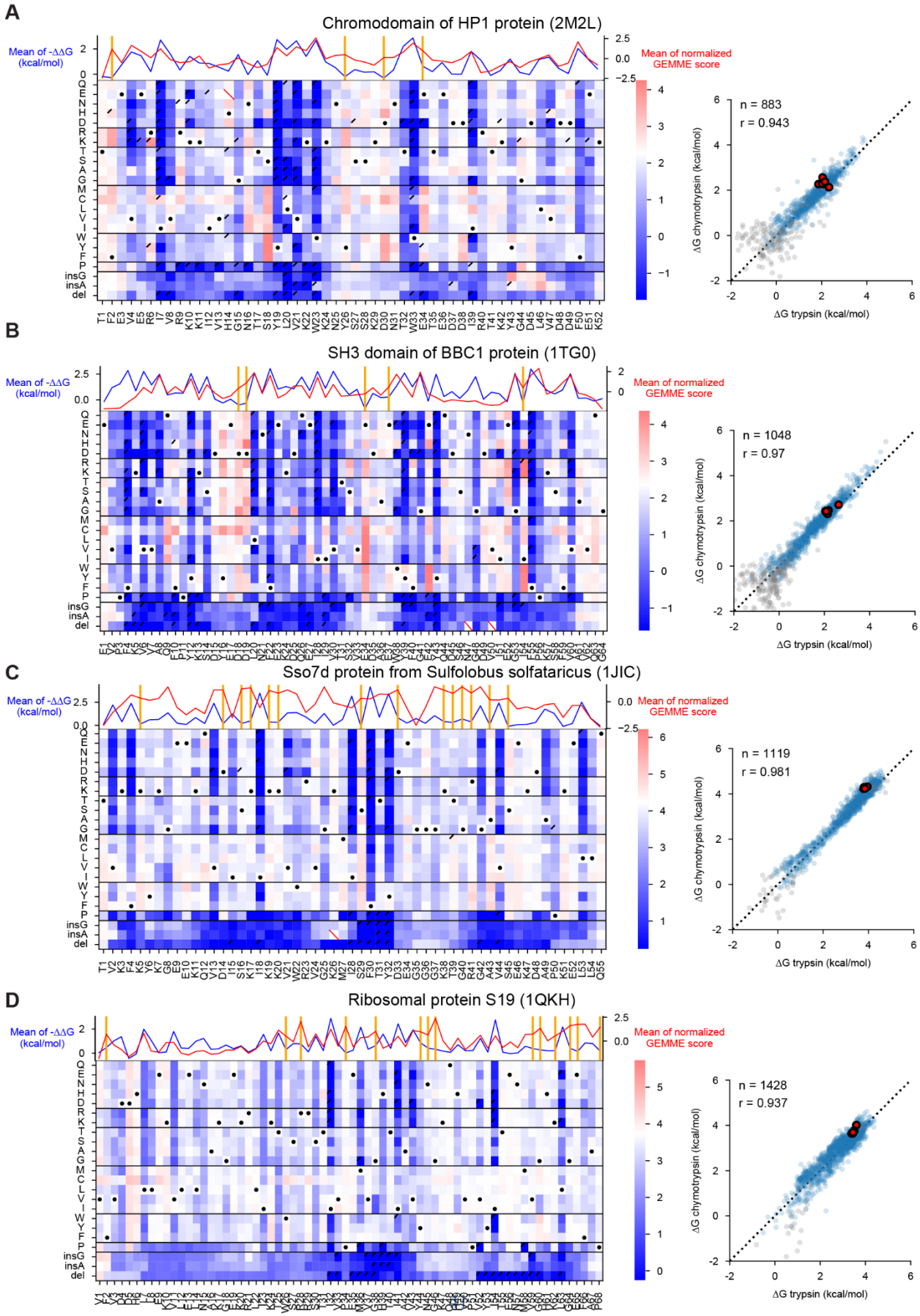

A
Q (197 sites)
E (406 sites)
N (178 sites)
H (85 sites)
D (284 sites)
R (254 sites)
K (391 sites)
T (246 sites)
S (271 sites)
A (382 sites)
G (298 sites)
M (101 sites)
L (441 sites)
V (368 sites)
I (254 sites)
W (75 sites)
Y (168 sites)
F (151 sites)
P (168 sites)
Total (4718 sites)
Cumulative occurrence
Cumulative probability
B
Amino cid frequency
Probability within ±0.05 bin

Fig. S15. Testing calibration of classification model for predicting wild-type amino acids

(A) Relationship between predicted cumulative probability and observed cumulative occurrence for each of 19 amino acids and total data. For each of the 19 amino acids (excluding Cys), we order all 4,718 sites from lowest to highest probability for that amino acid, then step through the sites in that order while plotting the fraction of the total cumulative probability (x-axis) and the fraction of all occurrences of that amino acid (y-axis). For the "Total" plot, we order all 89,642 (4,718*19) amino acid possibilities at all sites from lowest probability to highest probability, then step through all amino acid possibilities in that order while plotting the fraction of the total cumulative probability (x-axis) and the fraction of all actual amino acid occurrences (y-axis). The black diagonal lines show Y=X.

(B) Relationship between modeled amino acid probabilities and actual amino acid frequencies. For each of the 19 amino acids (excluding Cys), we binned all 4,718 sites into 20 bins according to the probability of that amino acid. Bins are spaced every 0.05 probability units and each bin has a width of 0.1, so sites can appear in two neighboring bins. For each bin (x-axis), the bar shows the true frequency of that amino acid in that bin (y-axis); error bars indicate the standard deviation of the true frequency from bootstrap resampling of all the sites. The black diagonal lines show Y=X (e.g. the predicted probability matches the true frequency). For the "Total" plot, we binned all 89,642 (4,718*19) amino acid possibilities at all sites as before, then counted the fraction of matching amino acids in each bin. Error bars represent the standard deviation of the frequencies from bootstrap resampling of all sites.


A
Chromodomain of HP1 protein (2M2L)
Mean of -ΔΔG (kcal/mol)
Mean of normalized GEMME score
n = 883
r = 0.943
ΔG chymotrypsin (kcal/mol)
ΔG trypsin (kcal/mol)
B
SH3 domain of BBC1 protein (1TG0)
Mean of -ΔΔG (kcal/mol)
Mean of normalized GEMME score
n = 1048
r = 0.97
ΔG chymotrypsin (kcal/mol)
ΔG trypsin (kcal/mol)
C
Sso7d protein from Sulfolobus solfataricus (1JIC)
Mean of -ΔΔG (kcal/mol)
Mean of normalized GEMME score
n = 1119
r = 0.981
ΔG chymotrypsin (kcal/mol)
ΔG trypsin (kcal/mol)
D
Ribosomal protein S19 (1QKH)
Mean of -ΔΔG (kcal/mol)
Mean of normalized GEMME score
n = 1428
r = 0.937
ΔG chymotrypsin (kcal/mol)
ΔG trypsin (kcal/mol)

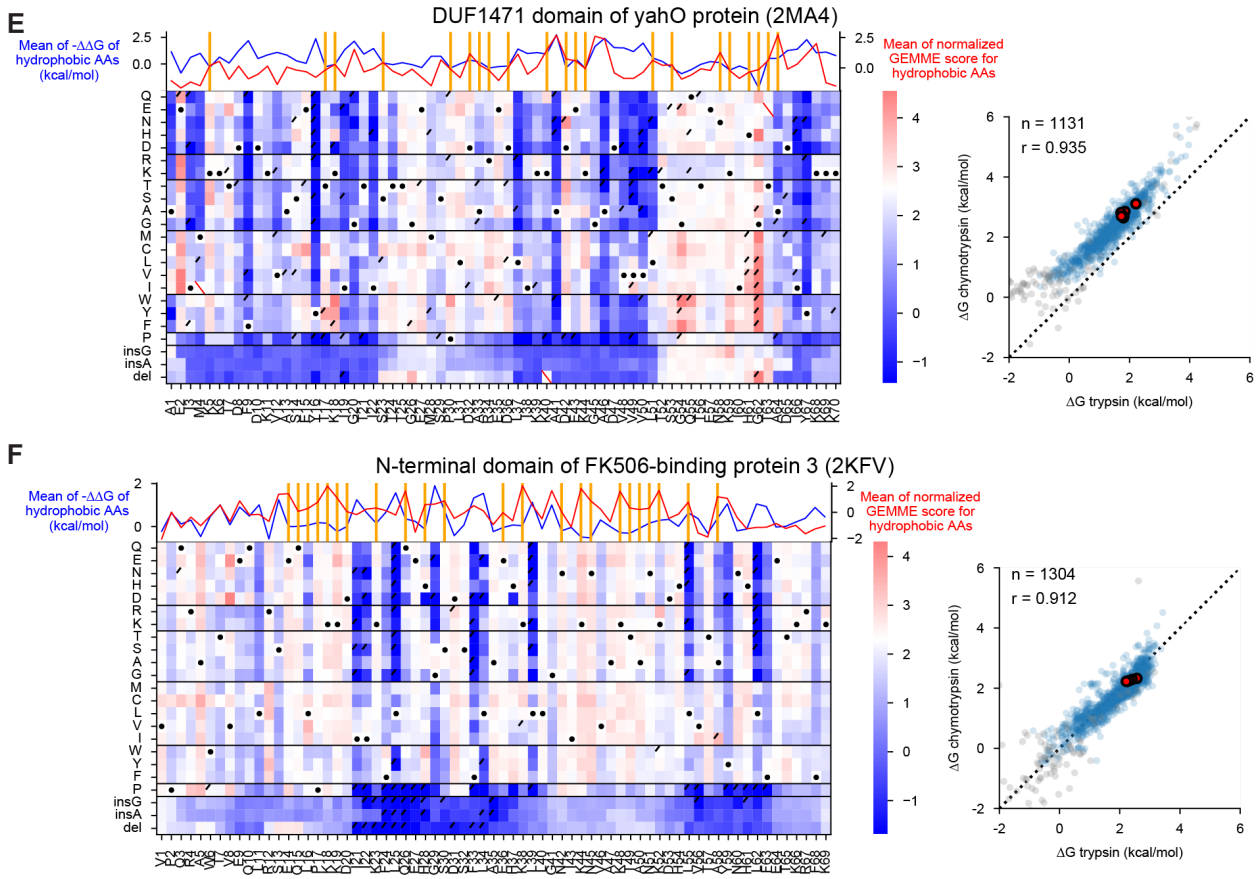


Fig. S16. Heat maps for notable domains with functional residues

(A- D) Mutational scanning results for four domains. Left: Heat maps show ΔG for substitutions, deletions, and Gly and Ala insertions at each residue. White indicates the wild-type stability and red/blue indicates stabilizing/destabilizing. Black dots indicate the wild-type amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5 kcal/mol), including ΔG estimates near the edges of the dynamic range. At top, lines show the mean ΔΔG (blue) and the mean normalized GEMME score (red), with functional sites (classified according to Fig. 6A) marked with vertical orange lines. Right: Agreement between variant ΔG values independently determined using assays with trypsin (x-axis) and chymotrypsin (y-axis). Multiple codon variants of the wild-type sequence are shown in red, reliable ΔG values in blue, and less reliable ΔG estimates (same as above) in gray. The black dashed line represents Y=X. Each plot shows the number of reliable points and the Pearson r-value for the blue (reliable) points.

(E- F) Same as (A-D), but top lines indicate the mean of ΔΔG for hydrophobic amino acid substitutions (blue) and mean normalized GEMME score of hydrophobic amino acids (red). Functional sites are classified according to Fig. 6G.

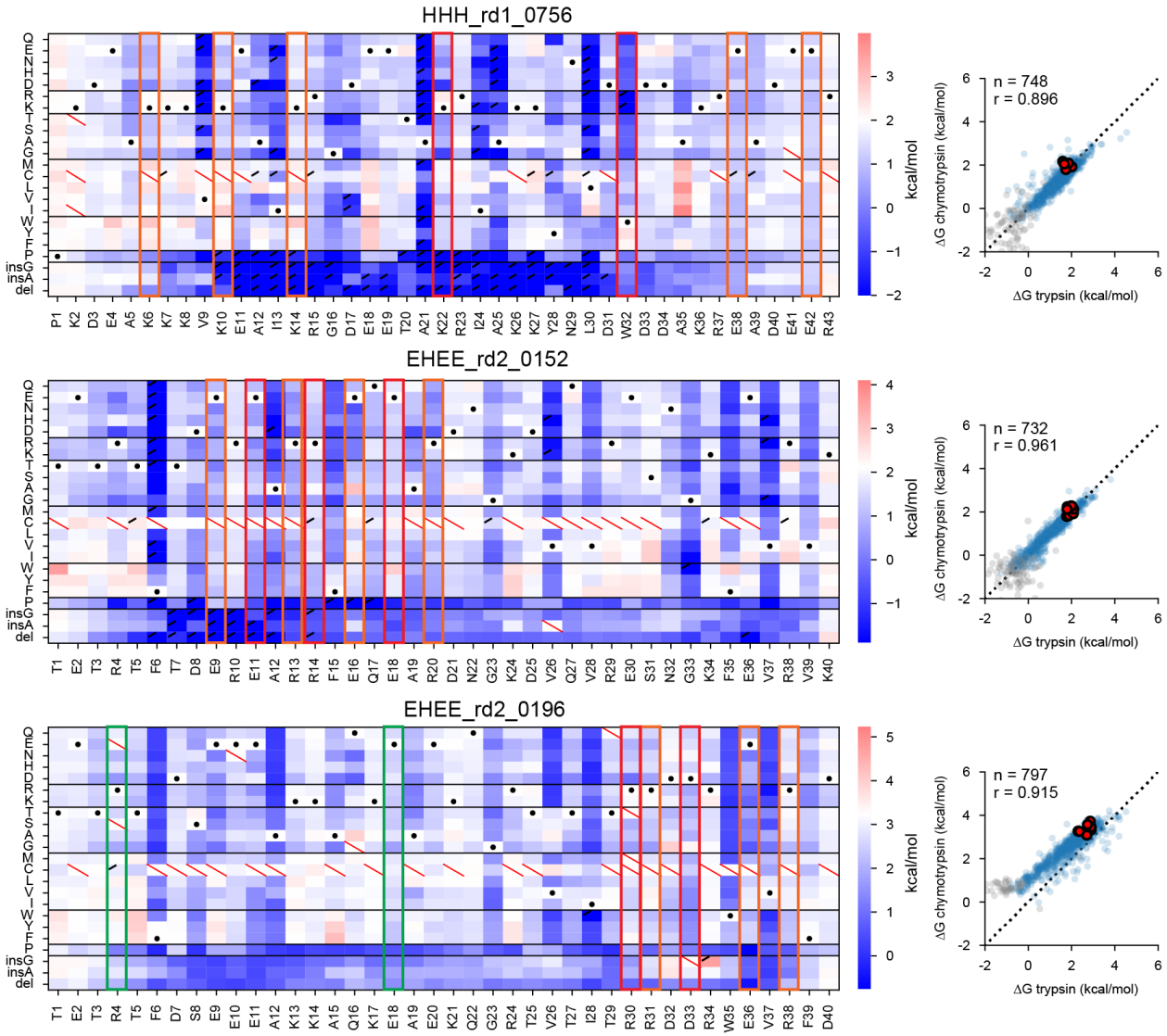


Fig. S17. Heat maps for three designed domains with notable polar interactions.

Mutational scanning results for three domains with notable polar interactions. Left: Heat maps show ΔG for substitutions, deletions, and Gly and Ala insertions at each residue. White indicates the wild-type stability and red/blue indicates stabilizing/destabilizing. Black dots indicate the wild-type amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5 kcal/mol), including ΔG estimates near the edges of the dynamic range. The polar networks shown in Fig. 7B are highlighted in orange, red, and green. Right: Agreement between variant ΔG values independently determined using assays with trypsin (x-axis) and chymotrypsin (y-axis). Multiple codon variants of the wild-type sequence are shown in red, reliable ΔG values in blue, and less reliable ΔG estimates (same as above) in gray. The black dashed line represents Y=X. Each plot shows the number of reliable points and the Pearson r-value for the blue (reliable) points.

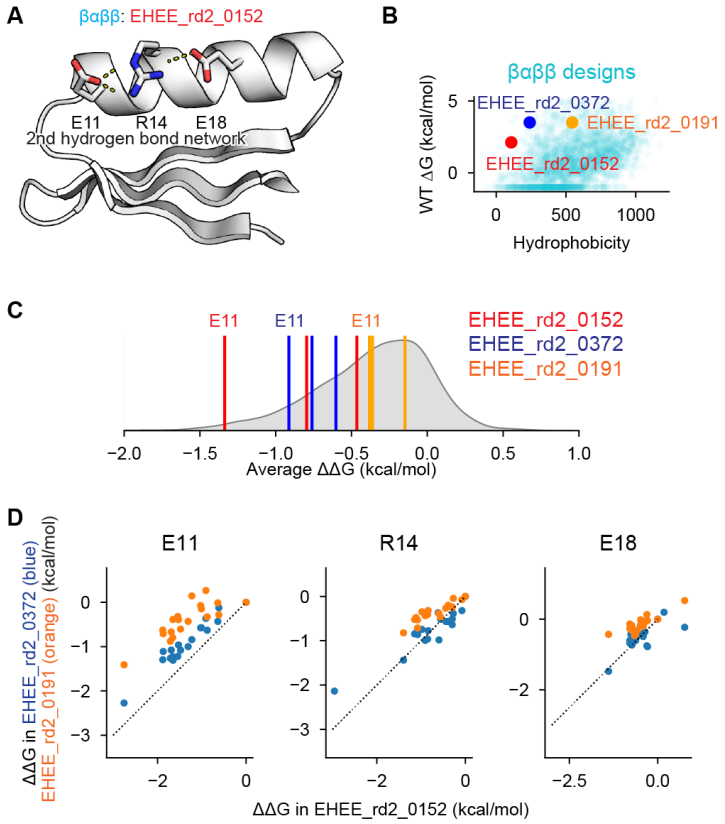


Fig. S18. Comparison of the notable hydrogen bond networks in three designs

(A) Designed structure of EHEE_rd2_0152 from (Rocklin et al., 2017) highlighting residues in the "2nd hydrogen bond network" defined in Fig. 7B middle.

(B) Relationship between hydrophobicity (calculated based on (Monera et al., 1995)) and folding stability (ΔG) for designed βαββ proteins. The three dots in the plot represent three designs with the same hydrogen bond network.

(C) The gray density plots represent the average ΔΔG of substitutions at 3,715 polar sites in 144 designed domains. The colored vertical bars indicate the values for the sites related to the 2nd hydrogen bond network.

(D) Relationship between ΔΔG in EHEE_rd2_0152 and in the other designs EHEE_rd2_0372 or EHEE_rd2_0191 for E11, R14, and E18. At E11, substitutions to the 19 other amino acids have smaller effects in EHEE_rd2_0372 (blue) and EHEE_rd2_0191 (orange) compared to in EHEE_rd2_0152 (e.g. all points are above the dashed Y=X line). However, the points are ordered similarly; i.e. the rank ordering of the 19 other amino acid variants in stability is similar between the three designs. For R14 and E18, substitutions in EHEE_rd2_372 (blue) have similar effect sizes to EHEE_rd2_0152, but substitutions in EHEE_rd2_0191 (orange) have smaller effects. Again, the rank ordering of the amino acid variants by stability is similar across the three designs.

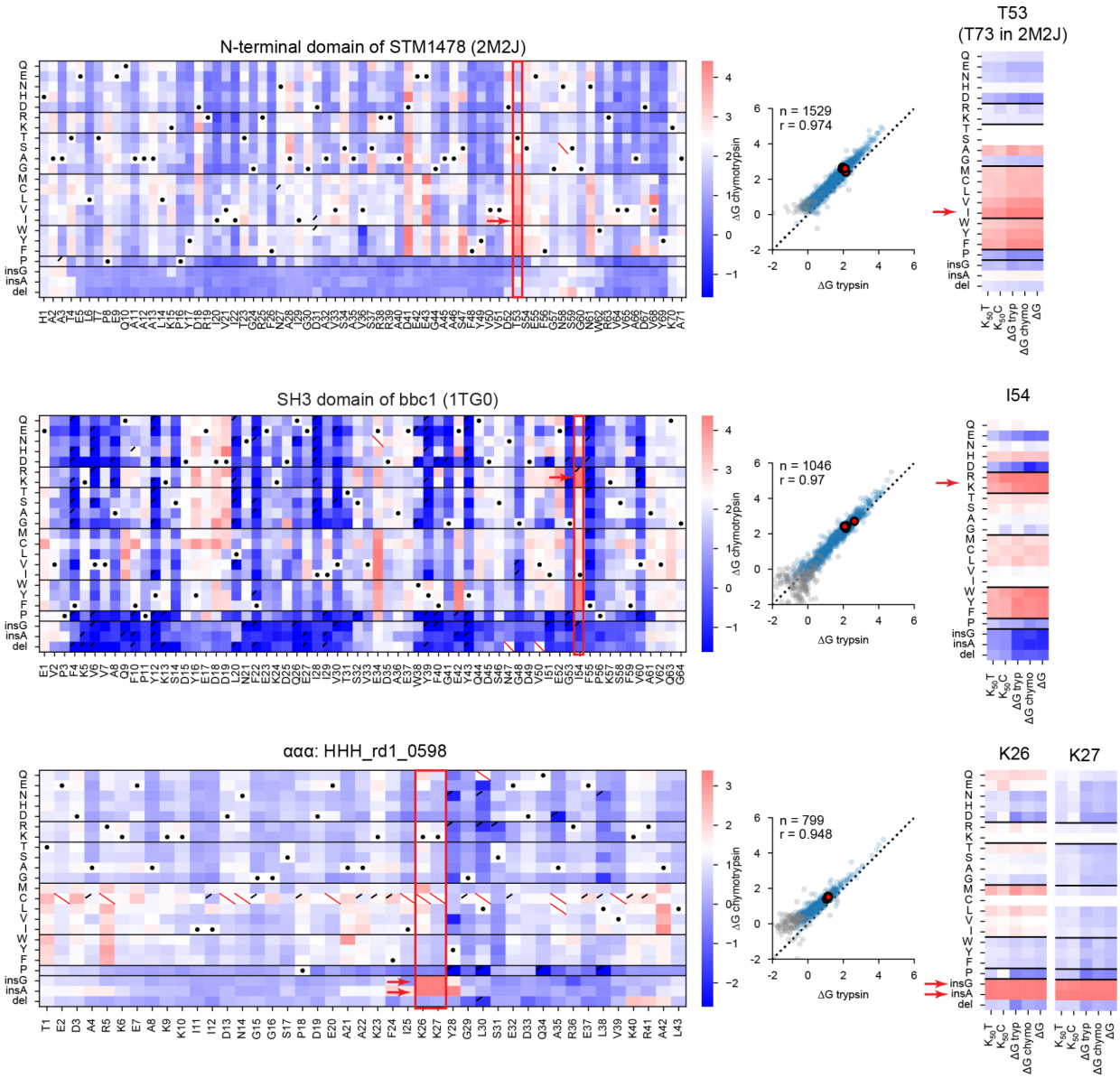


Fig. S19. Heat maps for three domains with notable stabilizing mutations

Left: Heat maps show ΔG for substitutions, deletions, and Gly and Ala insertions at each residue. White indicates the wild-type stability and red/blue indicates stabilizing/destabilizing. Black dots indicate the wild-type amino acid, red slashes indicate missing data, and black corner slashes indicate lower confidence ΔG estimates, (95% confidence interval > 0.5 kcal/mol), including ΔG estimates near the edges of the dynamic range. The red boxes and arrows highlight sites with notable stabilizing mutations. Middle: Agreement between variant ΔG values independently determined using assays with trypsin (x-axis) and chymotrypsin (y-axis). Multiple codon variants

of the wild-type sequence are shown in red, reliable ΔG values in blue, and less reliable ΔG estimates (same as above) in gray. The black dashed line represents Y=X. Each plot shows the number of reliable points and the Pearson r-value for the blue (reliable) points. Right: For four positions with stabilizing mutations, heatmaps show five experimental metrics: the trypsin (T) and chymotrypsin (C) K_{50} values, the ΔG values inferred from trypsin and chymotrypsin experiments, and the overall ΔG inferred from both trypsin and chymotrypsin experiments together.

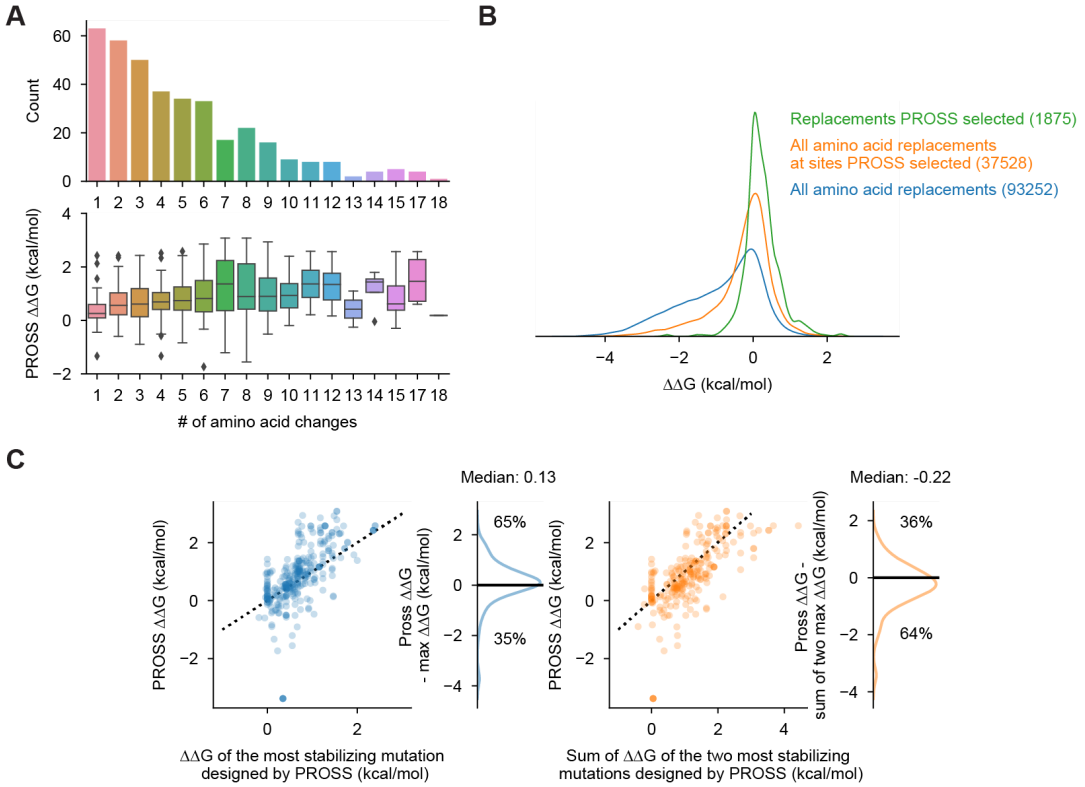


Fig. S20. Global analysis of PROSS designs

(A) All 727 PROSS designs grouped according to the number of amino acid substitutions in each design. Top: the number of designs with each different number of substitutions. Bottom: the distribution of design results for each group. ΔΔG indicates the stability of the PROSS design (ΔG) minus the stability of the original wild-type sequence; positive ΔΔG indicates the design stabilized the domain.

(B) ΔΔG distributions for all amino acid substitutions in wild-type domains used as input to PROSS (blue), all amino acid substitutions at sites modified in PROSS designs (orange), and all PROSS-designed substitutions (green). All ΔΔG measurements are in the original wild-type background; positive ΔΔG indicates stabilizing substitutions.

(C) Relationship between ΔΔG of PROSS designs and ΔΔG of the most stabilizing mutations designed by PROSS. At left, we compare PROSS designs to the single most stabilizing mutation (in the original wild-type background) out of all the substitutions in the PROSS design. At right, we compare PROSS designs to the sum of the two most stabilizing mutations (each measured individually in the original wild-type background without considering thermodynamic coupling). The density plots show the distribution of PROSS designs that were better (positive) or worse (negative) than the single best mutation (left) or sum of the two best mutations (right). Two-thirds of designs are better than the best single designed mutation, although the difference is small. Likewise, two-thirds of designs are worse than the additive effect of the two best designed mutations (assuming no thermodynamic coupling).

Table S1. List of amino acid sequence number for each group

Dataset name	# of total sequences	Sequence group	# of sequence groups	# of sequences
Dataset #1	586,938	Single amino acid replacement, deletion, and insertion	396 wild-types	434,556
		Double amino acid replacement	458 pairs	152,382
Dataset #2	851,552	Single amino acid replacement, deletion, and insertion	560 wild-types	629,287
		Double amino acid replacement	595 pairs	222,265
Dataset #3	1,844,548	Single amino acid replacement, deletion, and insertion	983 wild-types	1,046,752
		Double/Triple amino acid replacement	725 pairs (including 36 triples)	416,274
		Scrambles for unfolded model	-	68,427
		Rocklin 2017 rd1-3	-	36,707
		Others	-	276,388

Table S2. Conditions for measuring folding stability in the previous papers.

Protein	PDB ID	Pubmed ID	Reference	Offset in Fig.1G (kcal/mol)	Temperature (°C)	Buffer	pH
Protein GB1	1PGA	31371509	(Nisthal et al., 2019)	-0.8	25	NaPi	6.5
NTL9	2HBB	28494951	(Sato et al., 2017)	1.9	25	20 mM NAOAc, 100 mM NaCl	5.5
SAP domain	2WQG	26073259	(Dodson and Arbely, 2015)	1.2	10	50 mM MES pH 6.0, 500 mM NaCl	6
hPin1 WW domain	2M8I	19565466	(Jäger et al., 2009)	-1.9	40	10 mM NaPi	7
hYap65 WW domain	1K9Q	11420447	(Jiang et al., 2001)	-2.3	60	20 mM KPi, 100 mM NaCl	7
hYap65 WW domain	1K9Q	23035249	(Araya et al., 2012)	-2.3	50	20 mM NaPi	7
Villin HP35	1VII	23798426/19354264	(Xiao et al., 2009, 2013)	0.1	25	10 mM NaOAc, 150 mM NaCl	5
BBL	2WAV	19445954	(Neuweiler et al., 2009)	0.3	10	50 mM KPi 200 mM KCl	7
FF domain	1UZC	15935381	(Jemth et al., 2005)	0.8	10	50 mM NAOAc, 100 mM NaCl	5.7
ADA2h activation domain	1O6X	9799641	(Villegas et al., 1998)	-0.2	25	50 mM NaPi	7
hFyn SH3 domain	1SHF	9819209/12079394/16142914	(Maxwell and Davidson, 1998; Northey et al., 2002; de los Rios et al., 2005)	-0.1	25	10 mM Tris, 0.2 mM EDTA and 250 mM KCl	8

Table S3. Description of metrics in Fig. S8

Metrics	Description
dg_corr	Correlation between ΔG from trypsin challenge and chymotrypsin challenge
frac_pos_with_hydrophobic_stabilzing_muts	Fraction of stabilizing mutations which replace hydrophobic amino acids
frac_stabilzing_mut	Fraction of stabilizing mutations
lowconf_frac_lowss	Fraction of mutations with low reliable data (95CI > 0.5 kcal/mol)
NA_frac	Fraction of data without sufficient NGS reads
raw_corr	Correlation between trypsin K50 and chymotrypsin K50
slope	Slope of linear fitting line between ΔG from trypsin challenge and chymotrypsin challenge
width_KC	Standard deviation of ΔG from chymotrypsin challenge
width_KT	Standard deviation of ΔG from trypsin challenge
wt_d_from_line	Distance between WT data points and linear fitting line from mutational data
wt_dg	ΔG of WT
wt_dg_std_max	Standard deviation of ΔG of WT
y_intercept	Y-intercept of linear fitting line between ΔG from trypsin challenge and chymotrypsin challenge

Table S4. Description of columns in K50_dG_Dataset1_Dataset2

Column name	Description
name	sequence name
dna_seq	DNA sequence
log10_K50_t	Median of posteriors of K_{50} trypsin in log10 scale (μM)
log10_K50_t_95CI_high	Top 2.5%ile of posterior of K_{50} trypsin in log10 scale (μM)
log10_K50_t_95CI_low	Top 97.5%ile of posterior of K_{50} trypsin in log10 scale (μM)
log10_K50_t_95CI	log10_K50_t_95CI_high - log10_K50_t_95CI_low
fitting_error_t	Absolute error between the observed counts and the expected counts for a given sequence (based on all model parameters related to trypsin data), averaged over 24 conditions and normalized by the observed counts in the no-protease samples for that sequence
log10_K50unfolded_t	K_{50} unfolded trypsin in log10 scale (μM)
deltaG_t	ΔG calculated from log10_K50_t (kcal/mol)
deltaG_t_95CI_high	ΔG calculated from log10_K50_t_95CI_high (kcal/mol)
deltaG_t_95CI_low	ΔG calculated from log10_K50_t_95CI_low (kcal/mol)
deltaG_t_95CI	deltaG_t_95CI_high - deltaG_t_95CI_low (kcal/mol)
log10_K50_c	Median of posteriors of K_{50} chymotrypsin in log10 scale (μM)
log10_K50_c_95CI_high	Top 2.5%ile of posterior of K_{50} chymotrypsin in log10 scale (μM)
log10_K50_c_95CI_low	Top 97.5%ile of posterior of K_{50} chymotrypsin in log10 scale (μM)
log10_K50_c_95CI	log10_K50_c_95CI_high - log10_K50_c_95CI_low
fitting_error_c	Absolute error between the observed counts and the expected counts for a given sequence (based on all model parameters related to chymotrypsin data), averaged over 24 conditions and normalized by the observed counts in the no-protease samples for that sequence
log10_K50unfolded_c	K_{50} unfolded chymotrypsin in log10 scale (μM)
deltaG_c	ΔG calculated from log10_K50_c (kcal/mol)

deltaG_c_95CI_high	ΔG calculated from log10_K50_c_95CI_high (kcal/mol)
deltaG_c_95CI_low	ΔG calculated from log10_K50_c_95CI_low (kcal/mol)
deltaG_c_95CI	deltaG_c_95CI_high - deltaG_c_95CI_low (kcal/mol)
deltaG	Median of posterior of ΔG from trypsin+chymotrypsin data (kcal/mol)
deltaG_95CI_high	Top 2.5%ile posterior of ΔG from trypsin+chymotrypsin data (kcal/mol)
deltaG_95CI_low	Top 97.5%ile posterior of ΔG from trypsin+chymotrypsin data (kcal/mol)
deltaG_95CI	deltaG_95CI_high - deltaG_95CI_low
aa_seq_full	Amino acid sequence including padding linker sequence
aa_seq	Amino acid sequence without linker sequence
mut_type	Mutation type (like WT, substitution, insertion, deletion, or double mutants)
WT_name	Name of wild-type domain
WT_cluster	Cluster number of wild-type domain
log10_K50_trypsin_ML	K_{50} trypsin in log10 scale for machine learning (μM) ('-' means lacking or unreliable data)
log10_K50_chymotrypsin_ML	K_{50} chymotrypsin in log10 scale for machine learning (μM) ('-' means lacking or unreliable data)
dG_ML	ΔG for machine learning (kcal/mol) ('-' means unreliable data)
ddG_ML	ΔΔG for machine learning (kcal/mol) ('-' means unreliable data)
Stabilizing_mut	True if ΔΔG>1 (kcal/mol) and ΔΔG is reliable

Files for Supplementary Materials

Raw_NGS_count_tables.zip
- NGS_count_lib1.csv
- NGS_count_lib2.csv
- NGS_count_lib3.csv
- NGS_count_lib4.csv

K50_dG_tables.zip
- K50_dG_lib1.csv
- K50_dG_lib2.csv
- K50_dG_lib3.csv
- K50_dG_lib4.csv

Processed_K50_dG_datasets.zip
- K50_dG_Dataset1_Dataset2.csv
- K50_Dataset3.csv
- Single_DMS_list.csv
- Double_DMS_list.csv
- Triple_DMS_list.csv
- Heat_maps_single_DMS.pdf
- Heat_maps_double_DMS.pdf

Data_tables_for_figs.zip
- dG_extdG_data_Fig1.csv
- dG_site_feature_Fig3.csv
- dG_for_double_mutants_Fig4.csv
- dG_non_redundant_natural_Fig5.csv
- dG_GEMME_non_redundant_natural_Fig6.csv

Pipeline_qPCR_data.zip
- Raw_qPCR_data_FigS1.csv
- Process_qPCR_data.ipynb

Pipeline_K50_dG.zip
- STEP1_module.ipynb
- STEP1_run.ipynb
- STEP2_run.ipynb
- STEP3_run.ipynb
- STEP4_module.ipynb
- STEP4_run.ipynb
- STEP5_module.ipynb
- STEP5_run.ipynb
- Raw_NGS_counts_overlapped_seqs_STEP1_lib1_lib2.csv
- Raw_NGS_counts_overlapped_seqs_STEP1_lib2_lib3.csv
- Raw_NGS_counts_overlapped_seqs_STEP1_lib1_lib4.csv
- Raw_NGS_counts_overlapped_seqs_STEP1_lib2_lib4.csv

Raw_NGS_counts_overlapped_seqs_STEP1_lib3_lib4.csv
K50_scrambles_for_STEP3.csv
STEP1_out_protease_concentration_trypsin
STEP1_out_protease_concentration_chymotrypsin
STEP3_unfolded_model_params

Pipeline_figure_model.zip
Burial_side_chain_contact_Fig3_Fig6.ipynb
Additive_model_Fig4.ipynb
Classification_model_Fig5.ipynb

AlphaFold_model_PDBs.zip

Blueprints_for_EEHH.zip
eehh_EA_GBB_AGBB.bp
eehh_GG_GBB_AGBB.bp
eehh_XX_XXX_XXXX.bp