

Supplementary Figures

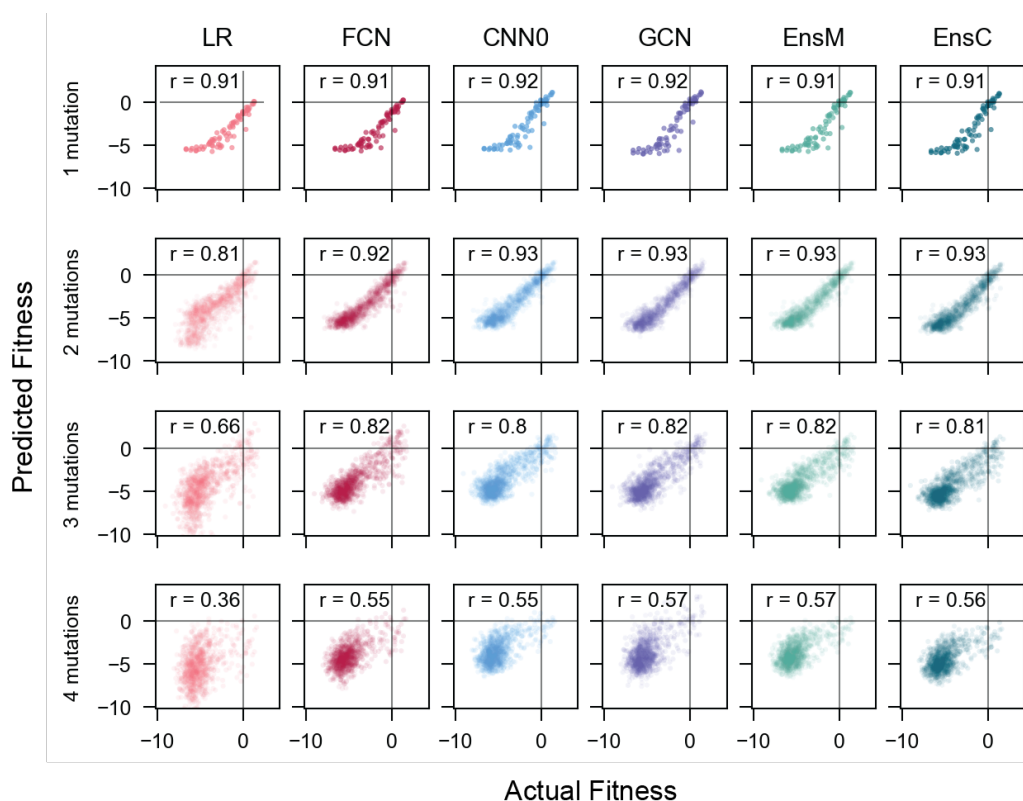


Figure S1. Correlation between actual and predicted fitness of 1-4 mutant GB1s. We used models trained by Gelman et al.¹ on single and double mutant GB1 fitness data from Olson et al.² and used each model to predict the fitness of all single, double, triple, and quadruple mutants from a 4-site GB1 combinatorial library from Wu et al.³ Correlation between experimental and predicted fitness decreases as distance from the training dataset increases. For each mutation distance, we calculated Pearson's r to assess correlation between the experimental and predicted fitness. If number of sequences was $> 1,000$, we selected a subset of 1,000 sequences. Sample sizes for mutation distances: 1, $N=76$; 2, $N=1,000$; 3, $N=1,000$; 4, $N=1,000$. Source data are provided as a Source Data file.

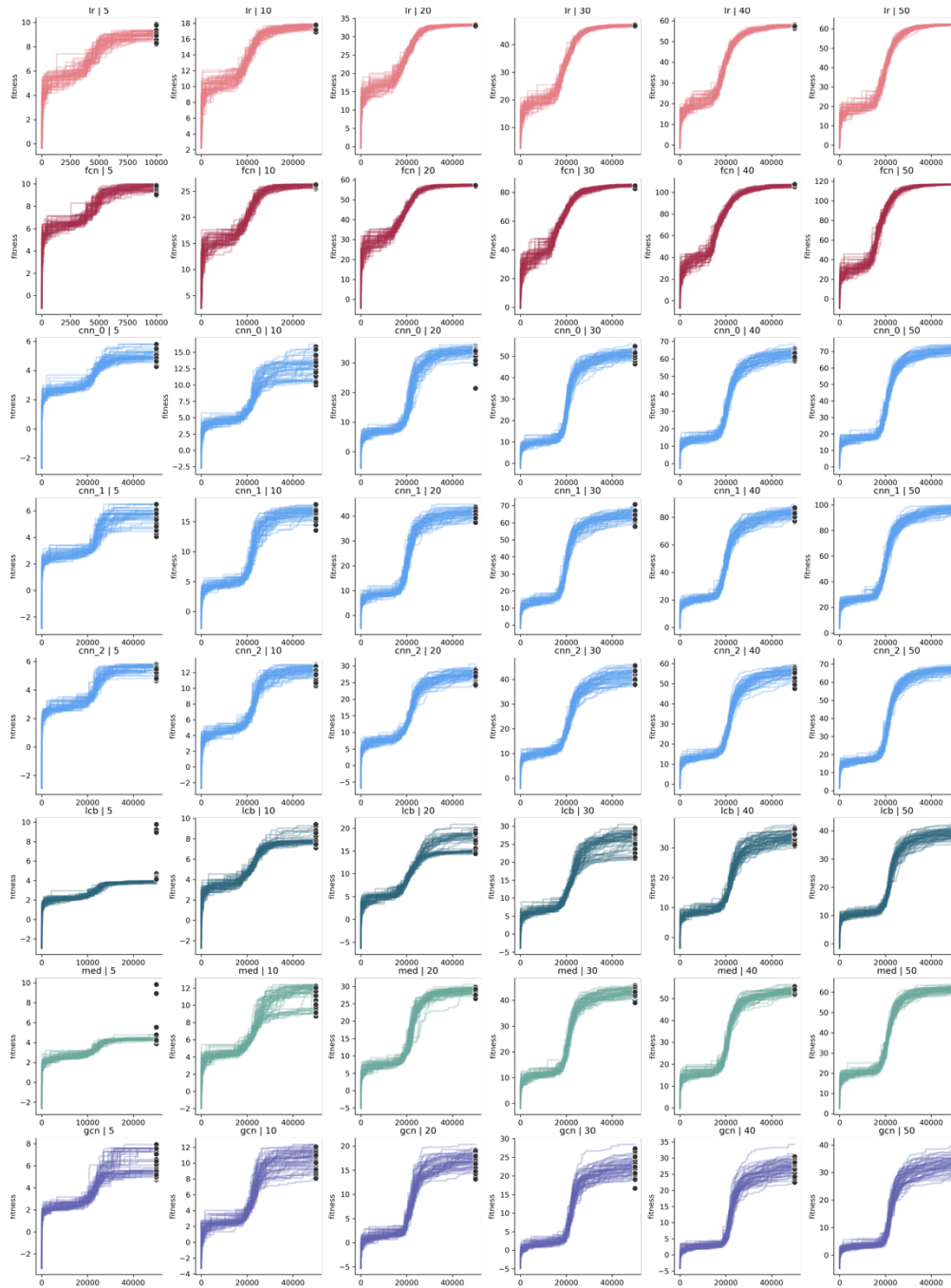


Figure S2. Simulated annealing optimization trajectories for each design condition. Each line represents a single optimization trajectory. Black dots represent the score of the 41 designs chosen for the experiments. Plots show 50 trajectories representative of the range of optimized fitnesses achieved after 500 simulations. Source data are provided as a Source Data file.

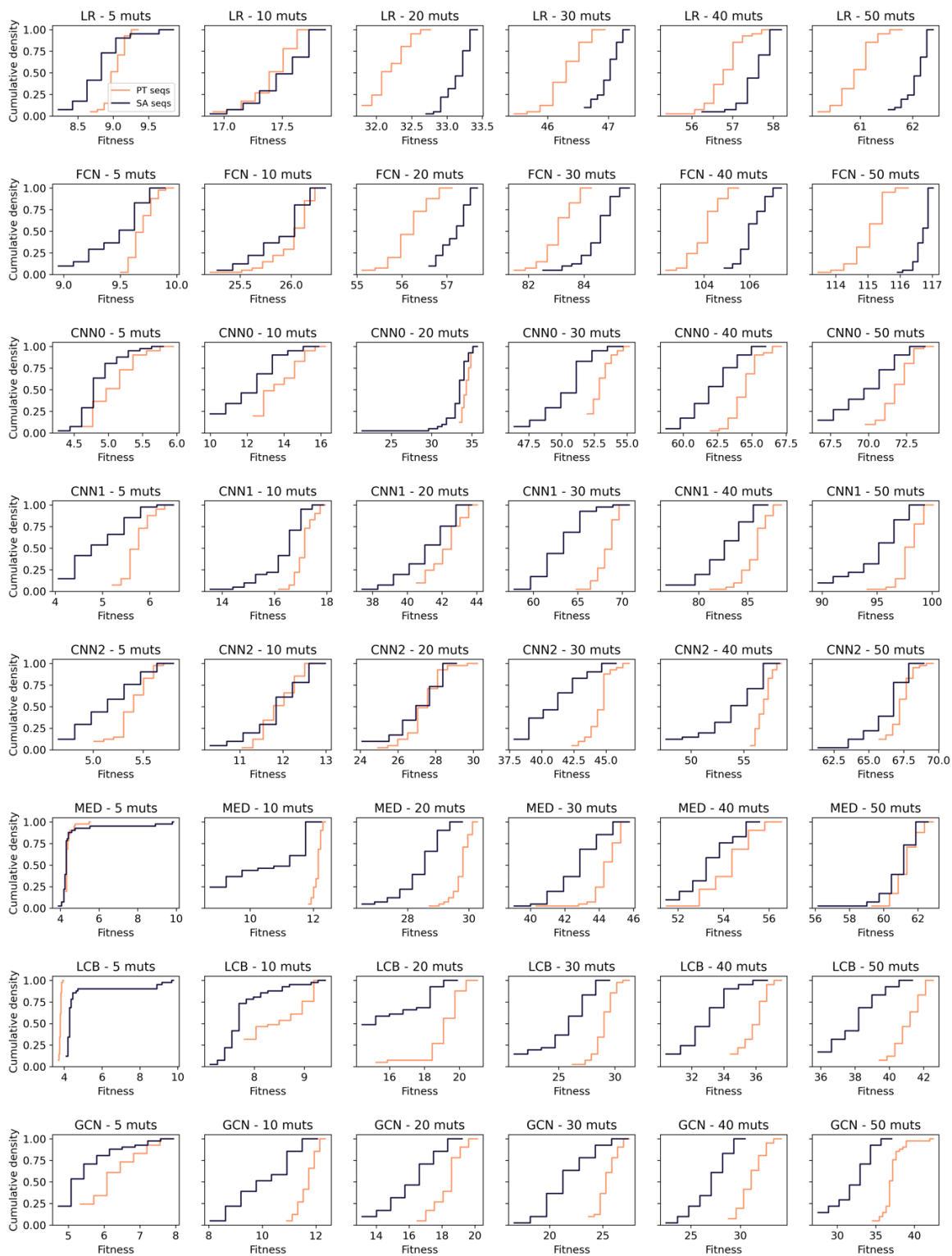


Figure S3. Parallel tempering achieves comparable fitness values across design categories. We used parallel tempering as an alternative method for sequence design. Number of parallel tempering runs and temperatures were optimized for each model and mutation distance such that sequence fitness was roughly near those designed by simulated annealing. Both methods achieve roughly the same range of fitness values for each design categories. Simulated annealing fitness values are shown in navy and parallel tempering fitness values are shown in orange. N=41 designs for each fitness distribution. Source data are provided as a Source Data file.

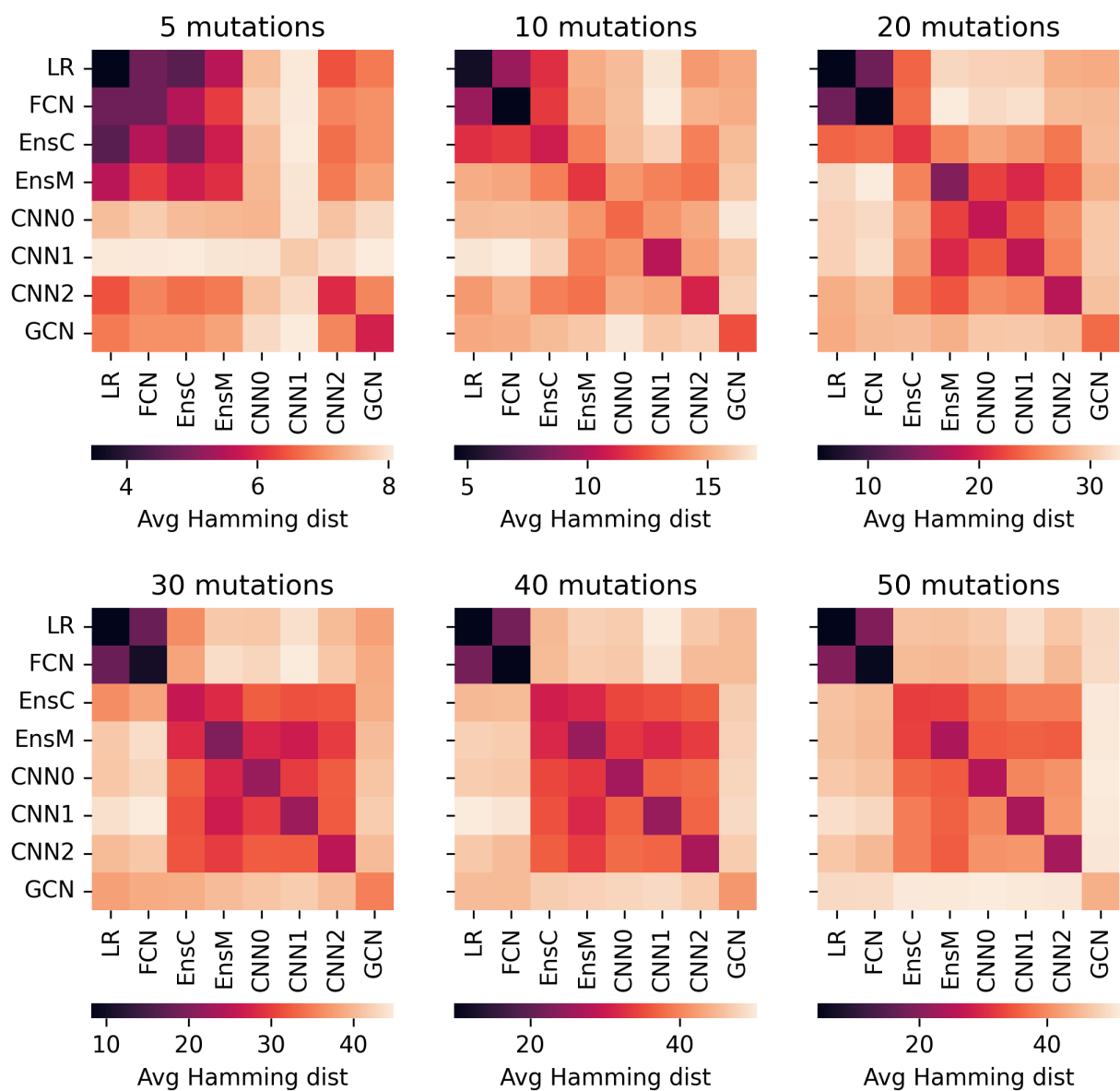


Figure S4. Designs produced by each model show distinct similarity patterns across mutational distances. We calculated the average pairwise Hamming distance between all designs from each pair of models (N=41 sequences for each design category) and performed hierarchical clustering to group design strategies based on their sequence similarity. Source data are provided as a Source Data file.

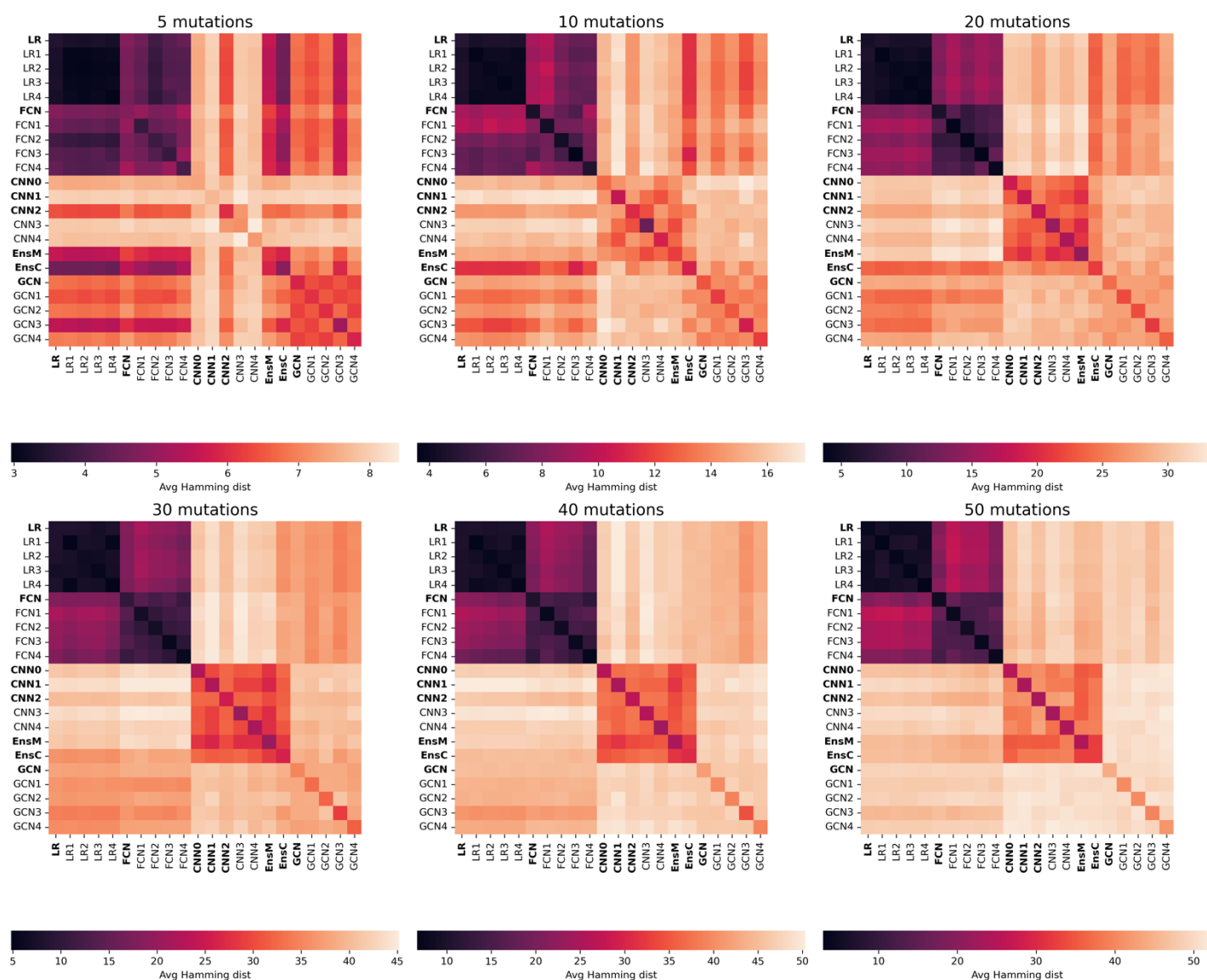


Figure S5. Effect of random parameter initialization on designs. We retrained multiple LR, FCN, and GCN models with different initializations. We designed sequences using additional LR(1-4), FCN(1-4), CNN(3 and 4), and GCN(1-4) models and compared the similarity of the designs within and between model architectures. For each model, we used the same design settings and methods that were used to design the original library sequences. We report the average pairwise hamming distance between all designs for each pair of models ($N=41$ sequences for each design category). Models used to design variants in GB1 library are bolded. We find variation within each model, but the sequences designed by a given architecture are generally more similar to other initializations of that architecture than other architectures. In other words, within architecture similarity is greater than between architecture similarity. The GCN model was an exception, where sequence variability was high between model initializations. Sequence dissimilarity between architectures increases with mutation distance from WT. Source data are provided as a Source Data file.

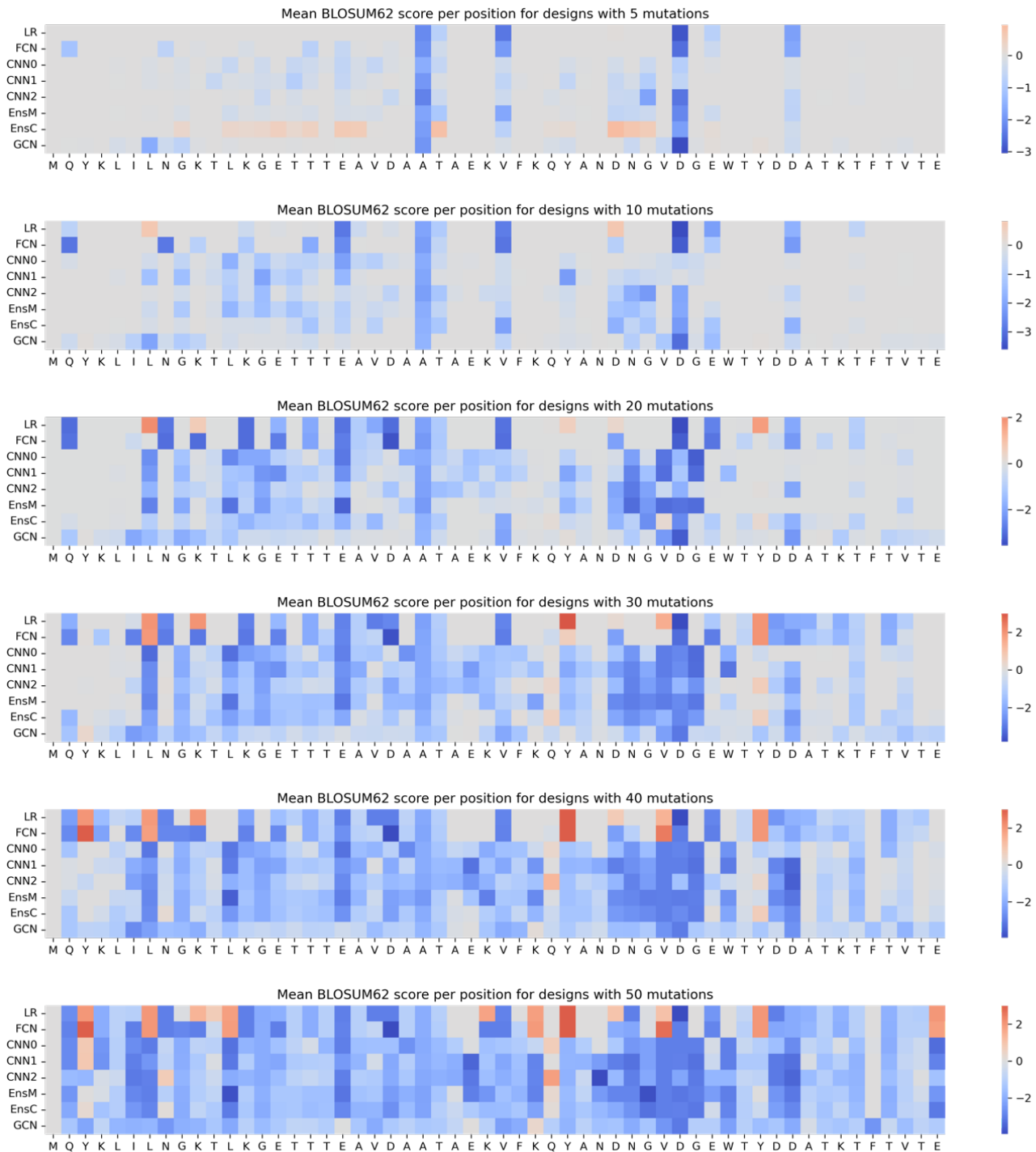


Figure S6. BLOSUM62 analysis of sequence conservation in GB1 library designs. We calculated the mean BLOSUM62 score for mutations at each position in the GB1 sequence by model and mutation regime (N=41 for each design category). Only mutations were included in this calculation, unmutated residues were excluded. Most mutations tend to be non-conservative on average at low mutational distances. There is one exception where the EnsC at 5 mutations generally makes conservative mutations to GB1. At higher mutational distances, select positions favor conservative mutations. This trend is more pronounced for the LR and FCN. Red indicates a conservative mean BLOSUM62 score while blue indicates a non-conservative BLOSUM62 score. Source data are provided as a Source Data file.

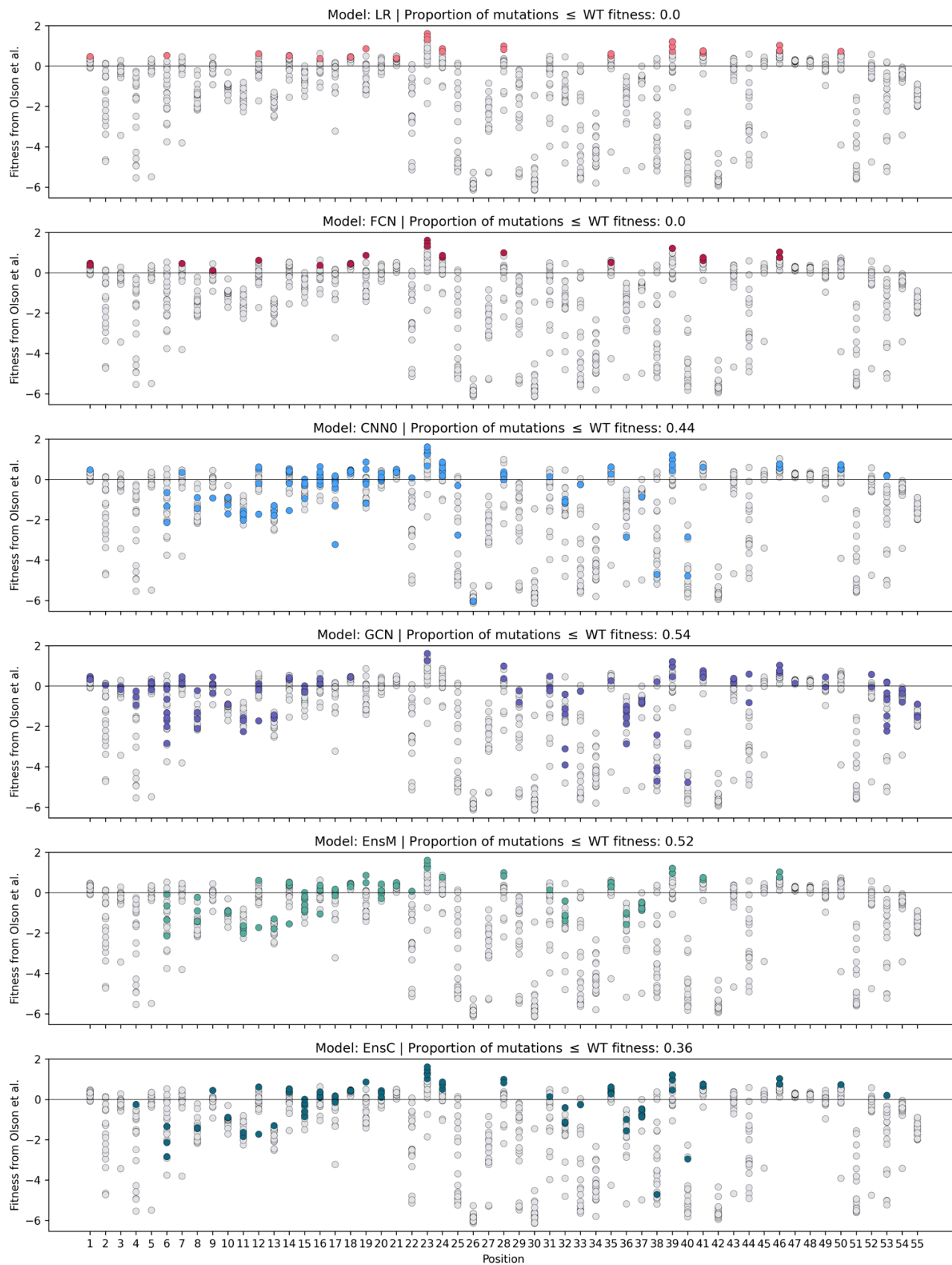


Figure S7. Experimental values for all single mutations in designs with 10 mutations. We show the original enrichment scores from Olson et al.² for all single mutants in grey. For each panel, we color the mutations found in each model's 10-mutant designs (N=41 sequences), highlighting the differences in fitness of single mutations chosen by each model. Source data are provided as a Source Data file.

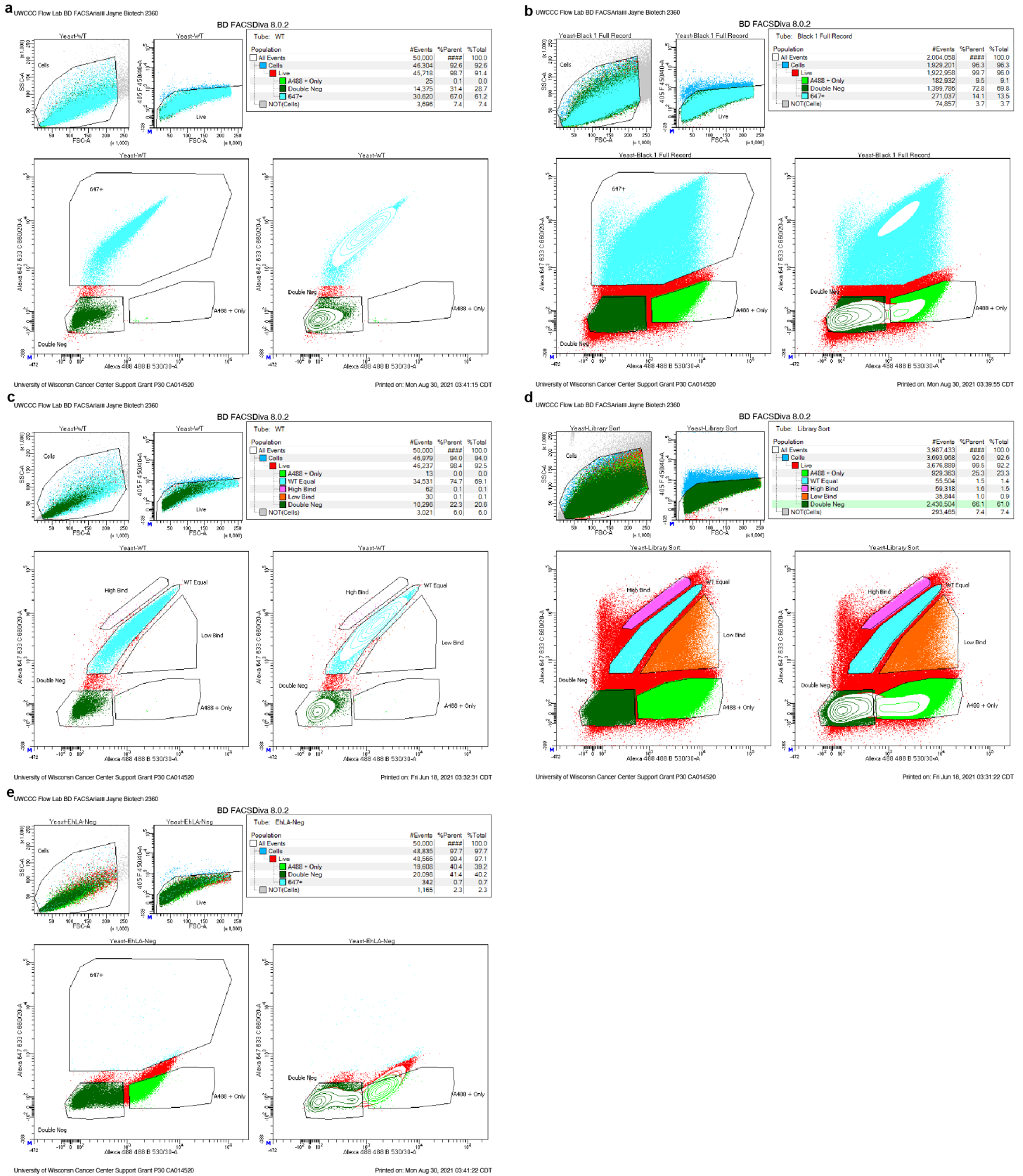


Figure S8. FACS sort records for binary (qualitative) and binned (quantitative) GB1 binding assays. (a) clonal wildtype GB1 with binary sort gates. N=50k. (b) Full sort record for binary FACS experiment. N~2M. (c) clonal wildtype GB1 with binned sort gates. N=50k. (d) Full sort record for binned FACS experiment. N~4M. (e) FACS sort with EHLA-negative yeast with binary sort gates. N=50k.

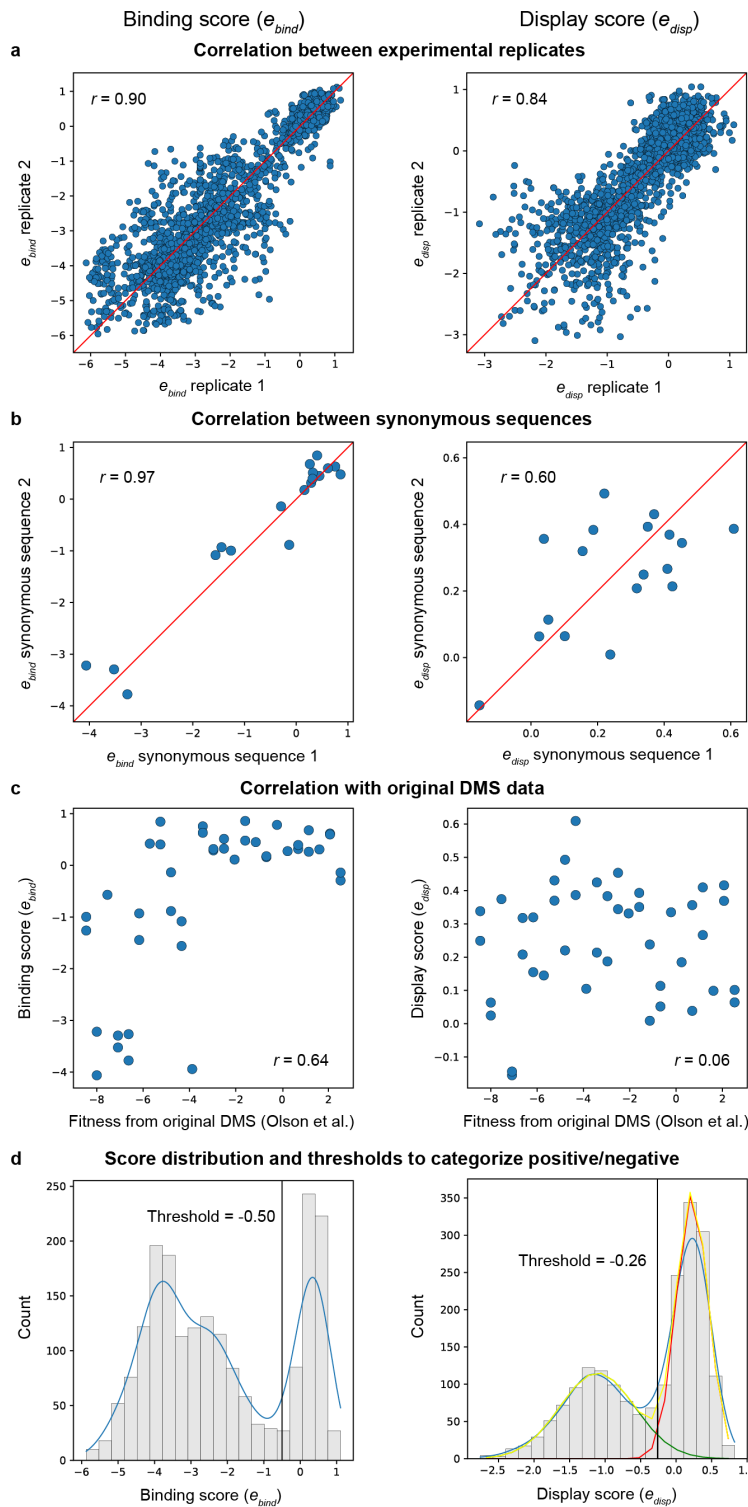


Figure S9. Reproducibility of binding and display enrichment scores. (a) The e_{bind} and e_{disp} scores show strong correlation between two independent experimental replicates. All subsequent analysis in the paper was performed on the average e_{bind} and e_{disp} between these two replicates (N=1,956 sequences). (b) We included synonymous sequence pairs as to ensure reproducible fitness measurements independent of nucleotide sequence. The e_{bind} and e_{disp} scores for these synonymous sequence pairs (N=18) show strong correlation indicating the assay is reliably measuring the protein's fitness. (c) In our experiment we also included 25 calibration sequences (including 18 synonymous sequence duplicates) from the original deep mutational scanning (DMS) dataset to use as a reference. Our e_{bind} score shows a moderate correlation with the DMS fitness, while e_{disp} shows no correlation (N=43 sequences). (d) We manually categorized sequences as binding/not binding based on the distribution of e_{bind} scores and manually setting a threshold to separate the two modes. We manually categorized sequences as displaying/not displaying based on the distribution of e_{disp} scores, fitting a bi-Gaussian distribution, and identifying where the two Gaussians' densities cross. The green Gaussian fits non-display designs, the red Gaussian fits display designs, and the yellow Gaussian is a combined bi-modal distribution of the green and red Gaussians. N=1,950 sequences. Source data are provided as a Source Data file.

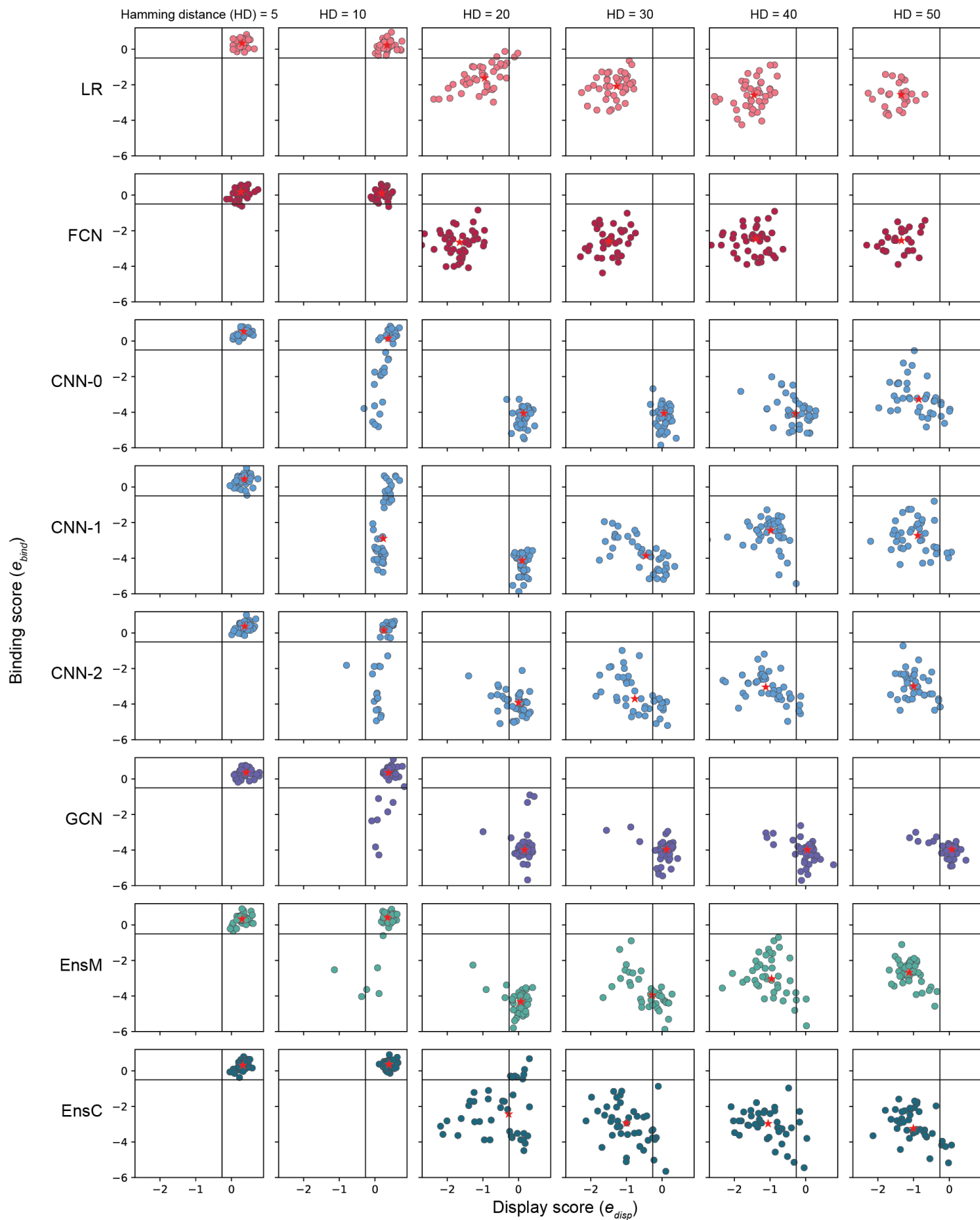


Figure S10. Distribution of binding and display scores of designs for each model at each mutational distance. Each design is shown as a colored point, plotted by its e_{bind} and e_{disp} scores. The median e_{bind} and e_{disp} score for each model at each mutational distance is shown as a red star. Black lines indicate the threshold to determine design binding and display activity. Source data are provided as a Source Data file.

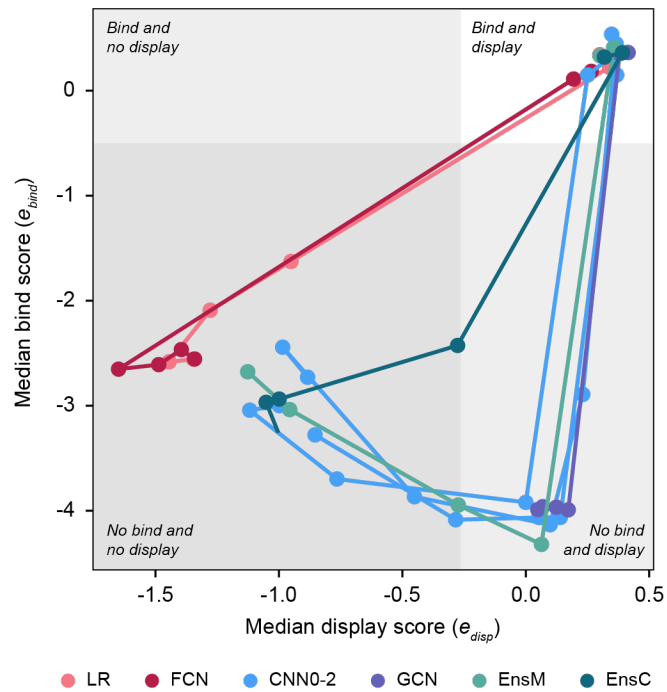


Figure S11. Median e_{bind} and e_{disp} score trajectory for increasing mutational distance. We calculated the median e_{bind} and e_{disp} scores for each model-mutational distance combination and plotted the trajectory with increasing mutational distance for each model. We overlay the e_{bind} and e_{disp} thresholds for bind/no bind and display/no display to separate the space into quadrants. Source data are provided as a Source Data file.

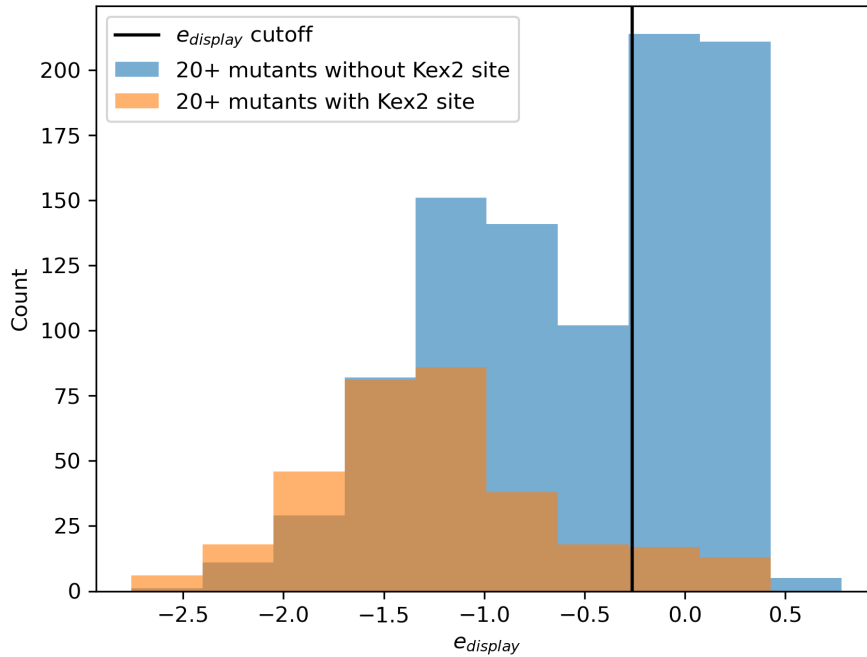


Figure S12. 20+ mutants with Kex2 sites display less than 20+ mutants without Kex2 sites. We examine e_{disp} scores as they relate to presence or absence of Kex2 sites for 20+ mutants since designs with many mutations can be prone to unfolding. Kex2 can cleave unfolded proteins in the yeast endoplasmic reticulum if they have a Kex2 cleavage site. We classify sequences as having a Kex2 site if they contain the consensus sequences KR and RR⁴. Variants without Kex2 sites may still display or not display depending on other factors. Source data are provided as a Source Data file.

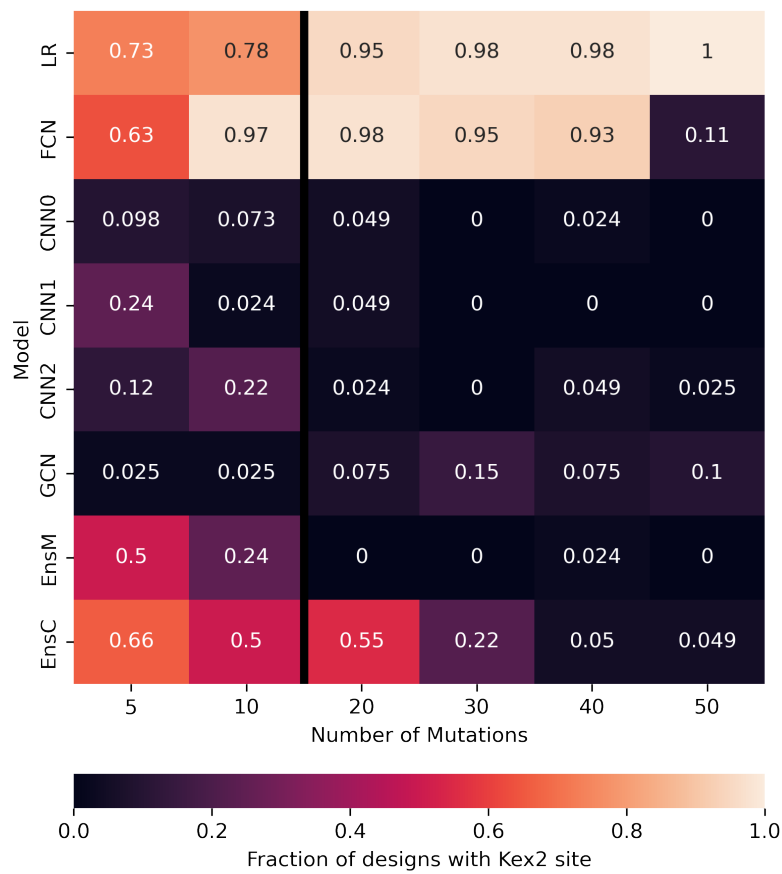


Figure S13. LR and FCN produce higher proportions of designs with Kex2 sites. We broadly determine possible Kex2 sites with the consensus sequences KR and RR and find the proportion of designs from each category with these consensus sequences. The black line separates the designs that are more likely to fold/not fold, based on the number of mutations, since folding hinders Kex2 cleavage even if a Kex2 cleavage site is present. Source data are provided as a Source Data file.

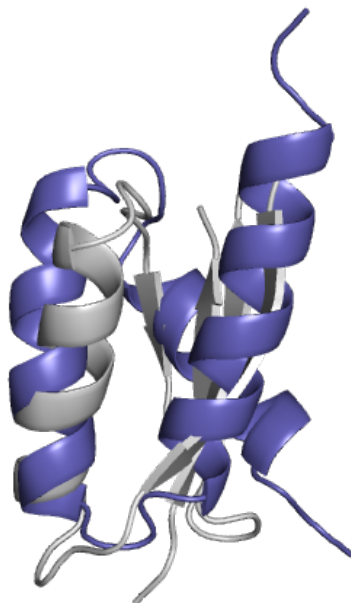


Figure S14. GCN-40 is predicted to fold into a new topology distinct from GB1. GCN-40, shown in purple, displays 2.3-fold higher than wildtype GB1, shown in grey, and is predicted by AlphaFold to fold into a triple helix. Source data are provided as a Source Data file.

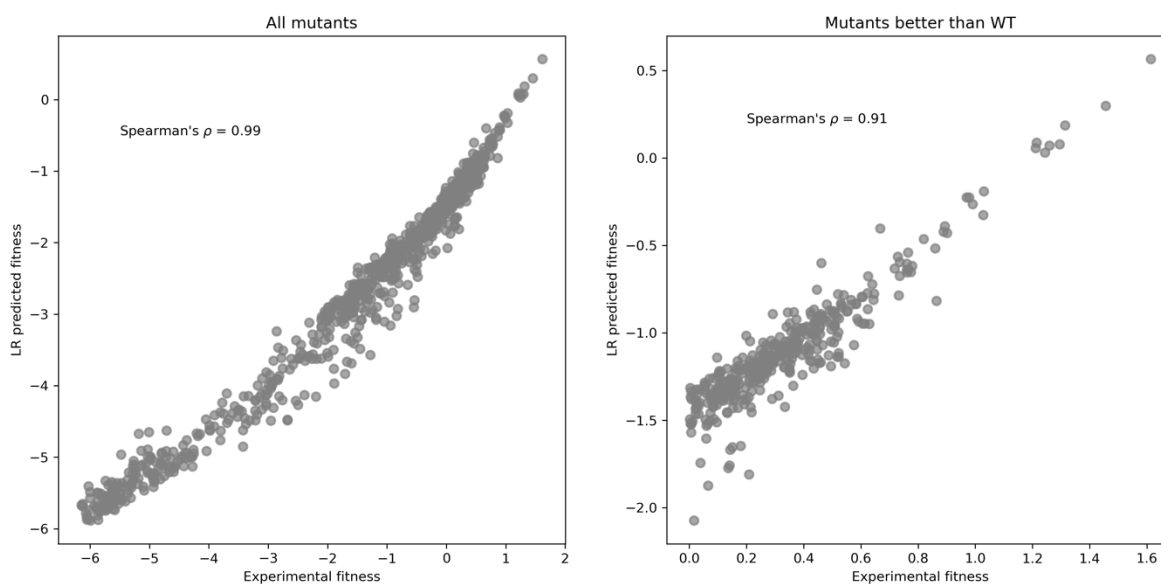


Figure S15. Comparison between the Addition model and linear regression. The Addition model simply considers the mutant's observed fitness difference from wildtype, while LR estimates an amino acid's contribution, marginalized over all observed sequence contexts. N=1,045 single mutants, N=367 single mutants better than WT GB1. Source data are provided as a Source Data file.

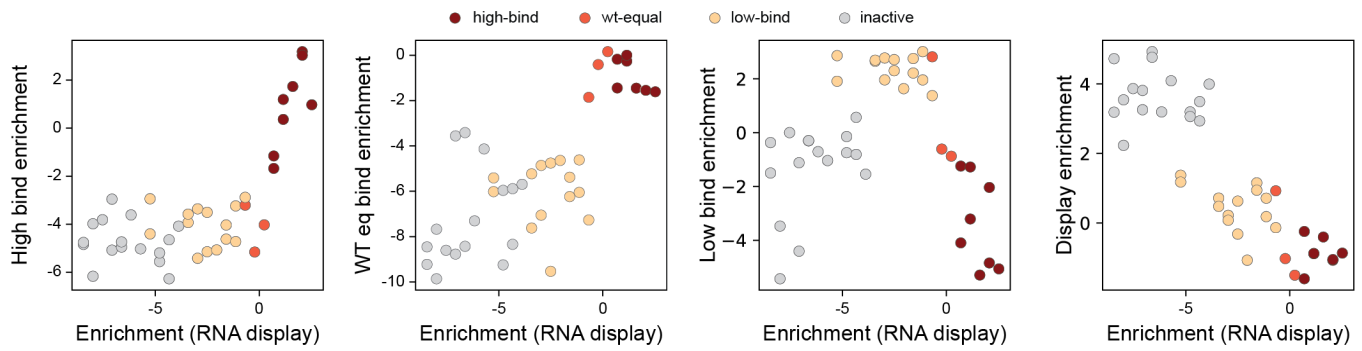


Figure S16. Categorical binding characterization correlation with binding enrichment from RNA display. To confirm that designs are placed in the correct category, we examine the enrichments of our 25 calibration sequences for each sorted population and for Olson et al.'s RNA display binding assay. We categorize each design as high-bind, wt-equal, low-bind, or inactive if the design has high enrichment in one of the populations, beyond our manually set threshold. N=43 sequences (including 18 synonymous sequences). Source data are provided as a Source Data file.

Table S1. Selected designs characterized with low-throughput yeast display binding and display assay. We selected twenty designs to validate high throughput measurements with additional binding assays. Designs 1-10 were selected for high display scores, designs 11-15 were selected for high binding scores in the qualitative experiments (noted in Bin column), and designs 16-20 were selected to test riskier designs with higher mutations that still exhibited some binding activity. Binding activity was assessed in both the qualitative (Bin column) and quantitative (Bind column) experiments.

Design	seq_id	Model	num_mut	ebind	edisp	Bin	Bind	Display
1	seq_1353	GCN	40	-3.7279055	0.155085636	No function	FALSE	TRUE
2	seq_1354	GCN	40	-4.5201031	0.415189322	No function	FALSE	TRUE
3	seq_1369	GCN	40	-4.320192	0.238841472	No function	FALSE	TRUE
4	seq_1370	GCN	40	-3.7390646	0.153355152	No function	FALSE	TRUE
5	seq_1392	GCN	40	-4.2974153	0.315622319	No function	FALSE	TRUE
6	seq_1524	CNN0	40	-5.1671984	0.116679435	No function	FALSE	TRUE
7	seq_1545	CNN0	40	-4.1887537	0.260701614	No function	FALSE	TRUE
8	seq_1689	GCN	50	-3.718207	0.17009947	No function	FALSE	TRUE
9	seq_1708	GCN	50	-4.1993973	0.213689676	No function	FALSE	TRUE
10	seq_1720	GCN	50	-3.5153361	0.193907131	No function	FALSE	TRUE
11	seq_40	FCN	5	-0.1622821	0.447811911	High bind	TRUE	TRUE
12	seq_184	EnsC	5	0.22656655	0.188751177	High bind	TRUE	TRUE
13	seq_195	EnsC	5	0.47332886	0.206546502	High bind	TRUE	TRUE
14	seq_308	CNN2	5	0.47960438	0.313436964	High bind	TRUE	TRUE
15	seq_559	CNN0	10	0.61749758	0.399715506	High bind	TRUE	TRUE
16	seq_769	LR	20	-0.2281441	-0.019846918	No function	TRUE	TRUE
17	seq_827	EnsC	20	-0.4550737	0.118095903	No function	TRUE	TRUE
18	seq_833	EnsC	20	0.68505174	0.304710547	WT Eq bind	TRUE	TRUE
19	seq_851	EnsC	20	-0.3566557	-0.167700417	No function	TRUE	TRUE
20	seq_853	EnsC	20	-0.2035726	0.149719922	WT Eq bind	TRUE	TRUE

Supplementary References

1. Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A. & Gitter, A. Neural networks to learn protein sequence-function relationships from deep mutational scanning data. *Proc Natl Acad Sci U S A* **118**, (2021).
2. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology* **24**, 2643–2651 (2014).
3. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, 1–21 (2016).
4. Li, Q. *et al.* Profiling Protease Specificity: Combining Yeast ER Sequestration Screening (YESS) with Next Generation Sequencing. *ACS Chemical Biology* **12**, 510–518 (2017).