Designed active-site library reveals thousands of functional GFP variants

Jonathan Yaacov Weinstein[1], Carlos Martí-Gómez[2], Rosalie Lipsh-Sokolik[1], Shlomo Yakir Hoch[1], Demian Liebermann[3], Reinat Nevo[1], Haim Weissman[4], Ekaterina Petrovich-Kopitman[5], David Margulies[6], Dmitry Ivankov[7], David McCandlish[2], Sarel Jacob Fleishman[1]

[1]Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel
[2]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, 11724, NY, USA
[3]Department of Chemical and Biological Physics, Weizmann Institute of Science, Rehovot 7610001, Israel
[4]Department of Molecular Chemistry and Materials Science, Weizmann Institute of Science, Rehovot, 7610001 Israel
[5]Life science Core facilities, Weizmann Institute of Science, Rehovot 7610001, Israel
[6]Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot 7610001, Israel
[7]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia

Abstract

Mutations in a protein active site can lead to dramatic and useful changes in protein activity. The active site, however, is extremely sensitive to mutations due to a high density of molecular interactions, drastically reducing the likelihood of obtaining functional multipoint mutants. We introduce an atomistic and machine-learning-based approach, called htFuncLib, to design a sequence space in which mutations form low-energy combinations that mitigate the risk of incompatible interactions. We applied htFuncLib to the GFP chromophore-binding pocket, and, using fluorescence readout, recovered >16,000 unique designs encoding as many as eight active-site mutations. Many designs exhibit substantial and useful diversity in functional thermostability (up to 96 °C), fluorescence lifetime, and quantum yield. By eliminating incompatible active-site mutations, htFuncLib generates a large diversity of functional sequences. We envision that htFuncLib will be useful for one-shot optimization of activity in enzymes, binders, and other proteins.

Introduction

Protein active sites comprise molecular-interaction networks that are critical to function. Due to the molecular density of the active site, however, the majority of mutations destabilize the protein[1] or lead to dysfunction[2], and functional multipoint mutants are exceptionally rare[3,4]. Thus, active sites are among the most evolutionarily conserved protein sites[5]. Furthermore, experimental lab-evolution studies that aim to modify protein activity typically discover many more mutations outside the active site than within it[6]; yet, understanding whether and how remote mutations change activity is often elusive[7,8]. Although active-site mutations have the greatest potential to alter function, in practice, sensitivity to mutation has severely limited access to active-site functional variants in natural and lab evolution and in deep mutational scanning[9,10] and computational protein design[11]. Therefore, as a rule, lab-evolution studies comprise multiple cycles of mutagenesis and selection that are customized specifically for each desired functional trait[12–14]. Such iterative processes are time consuming and likely to severely undersample the space of functional sequences.

Furthermore, epistatic interactions between mutations can severely restrict the chances of finding functional multipoint mutants[15]. In epistasis, a mutation may be tolerated only if another position has already been mutated[16–18], drastically reducing the chances for the emergence of beneficial multipoint mutants[15,19]. This dependence also severely limits our ability to predict the functional impact of multipoint mutations even when the effects of single-point mutations are known[20,21], for instance, based on deep mutational scanning[3,4]. Epistasis has critical implications for our understanding of molecular evolution, including the emergence of viral and microbial resistance mutations[22] and the evolution of new enzymatic and binding specificities[23]. It also presents one of the primary obstacles to our ability to design protein activities in basic and applied research[1,24].

Here, we introduce a computational method called high-throughput functional libraries (htFuncLib) to design large libraries of active-site mutants that can be applied, in principle, to any protein. Most current atomistic design methods, including our previously described FuncLib method[24], select designs that optimize desired energy or structure criteria[25,26]. By contrast, htFuncLib searches for a set of active-site point mutations that, when freely combined, yield low-energy multipoint designs. Our approach can be applied to an arbitrarily large set of positions to generate diverse and complex libraries that encode millions of designs. htFuncLib thus accesses sequence spaces that have so far been interrogated through random or semi-random mutagenesis and selection methods. Yet, unlike such methods, htFuncLib generates libraries that are preselected computationally to enrich for stable, folded, and potentially active designs.

Results

Principles for designing combinatorial active-site diversity

We applied htFuncLib to Green Fluorescent Protein (GFP). GFP and other fluorescent proteins have attracted intense interest in evolution studies due to their ubiquitous uses in molecular and cellular biology[27–29] and their straightforward optical readout[30]. GFP fluorescence depends on the chemical environment of the chromophore, including electrostatics and torsional freedom about the bond that links its aromatic rings[31] and is therefore sensitive to mutations in the chromophore-binding pocket. Most previous large-scale screens targeted the entire protein or consecutive segments of it[3,4,30,32]. GFP is a β-barrel, however, and the chromophore is buried within the protein core. Therefore, most mutations targeted solvent-exposed regions that are unlikely to impact spectral properties. Unlike these previous studies, we apply htFuncLib solely to positions that line the chromophore-binding pocket. Because active-site mutations may reduce protein stability, we chose as a starting point a previously designed version of enhanced GFP, PROSS-eGFP, that exhibited elevated resistance to thermal denaturation[33]. In this previous design, active-site positions, except Tyr145Phe and Thr167Ile, were immutable. In applying htFuncLib, we also allowed design in these two positions.

Our working hypothesis is that epistatic interactions most frequently arise from three molecular sources (Supplementary Figure 1): (1) direct molecular interactions between proximal mutated amino acids; (2) indirect interactions between amino acid positions due to backbone conformational changes; and (3) stability-mediated interactions in which destabilizing mutations do not exhibit phenotypic differences when introduced singly but reduce stability or expression levels when combined[1,7].

The htFuncLib approach combines phylogenetic analysis, Rosetta atomistic design calculations[26,34], and a machine-learning analysis to nominate mutually compatible mutations when combined freely with one another (see Methods for details). Using Figure 1 as a visual guide for applying htFuncLib to GFP, we started by manually selecting 27 active-site positions likely to impact functional properties based on previous GFP studies or proximity to the chromophore (Figure 1A). htFuncLib then computed all single-point mutations and selected the ones likely to be tolerated against the background of the original amino acids in all other positions[34]. In this selection step, we retain mutations that are likely to be present in the diversity of sequence homologs and that are moreover predicted not to destabilize the protein native state according to atomistic design calculations[35]. The atomistic calculations contain the chromophore to ensure that the mutations do not abrogate contacts that may be critical to fluorescence. In addition, these calculations apply harmonic coordinate constraints to backbone atoms during whole-structure minimization, thereby penalizing backbone deformations that may lead to indirect epistatic interactions (Supplementary Figure 1B).

After filtering, htFuncLib applies atomistic modeling to evaluate the energy of combinations of tolerated point mutations. Since the space of potential multipoint mutations in a large active site is computationally intractable for enumeration, we focus calculations on combinations of mutations within neighborhoods of proximal positions (Figure 1B & C, Supplementary Tables 1 and 2) which are the most likely to give rise to direct epistatic interactions (Supplementary Figure 1A). In a companion paper, we show how to select combinations of enzyme backbone fragments that form low-energy combinations when freely combined using a new machine-learning-based approach called EpiNNet[36]. Here, we apply EpiNNet to select low-energy combinations of mutations across all spatial neighborhoods within the chromophore-binding pocket. The multipoint mutants within each neighborhood are classified according to their energies into favorable (Rosetta energies lower than PROSS-eGFP) and unfavorable (highest-energy 50%, Figure 1D). We then train the neural network to predict the energy-based classification of favorable and unfavorable designs. Finally, the trained network ranks the single-point mutations according to their likelihood of being found in low-energy multipoint mutants, and the top-ranked mutations are selected for library construction. The resulting library is enriched in mutually compatible mutations, such that both direct and stability-mediated epistasis (Supplementary Figure 1A and C) are addressed. Following design, we clone the library using Golden Gate assembly[37] (Figure 1E) and apply FACS sorting and deep sequencing to identify active designs (Figure 1F).

Multipoint mutants from the EpiNNet-enriched sequence space exhibit, on average, dramatically lower computed energies than those in the original filtered sequence space (Figure 2A), suggesting that EpiNNet increases the fraction of folded and stable designs. Furthermore, in a representative case, following the selection steps, only 14 positions (out of the 27 we selected initially) were selected for design with a sequence space of 10^7, compared to experimentally intractable 10^{35} sequences for the space encompassing every mutation at 27 positions and 10^{19} following the phylogenetic and single-mutation energy filters (Supplementary Tables 3 and 4).

Thus, unlike conventional protein design methods[25,26], htFuncLib does not search for the most optimal mutants according to energy or structural criteria. Instead, the astronomically large space of combinatorial mutations in an active site is reduced to a tractable size through phylogenetic, structural, and energy-based analysis. Then, mutations that may destabilize the protein in combination with others are removed by analyzing the energies of combinatorial mutations. Thus, htFuncLib assumes that active-site stability is a primary constraint for discovering functional multipoint mutants[1,38,39]. Additional functional constraints are encoded by verifying that the mutants form favorable interactions with the chromophore.

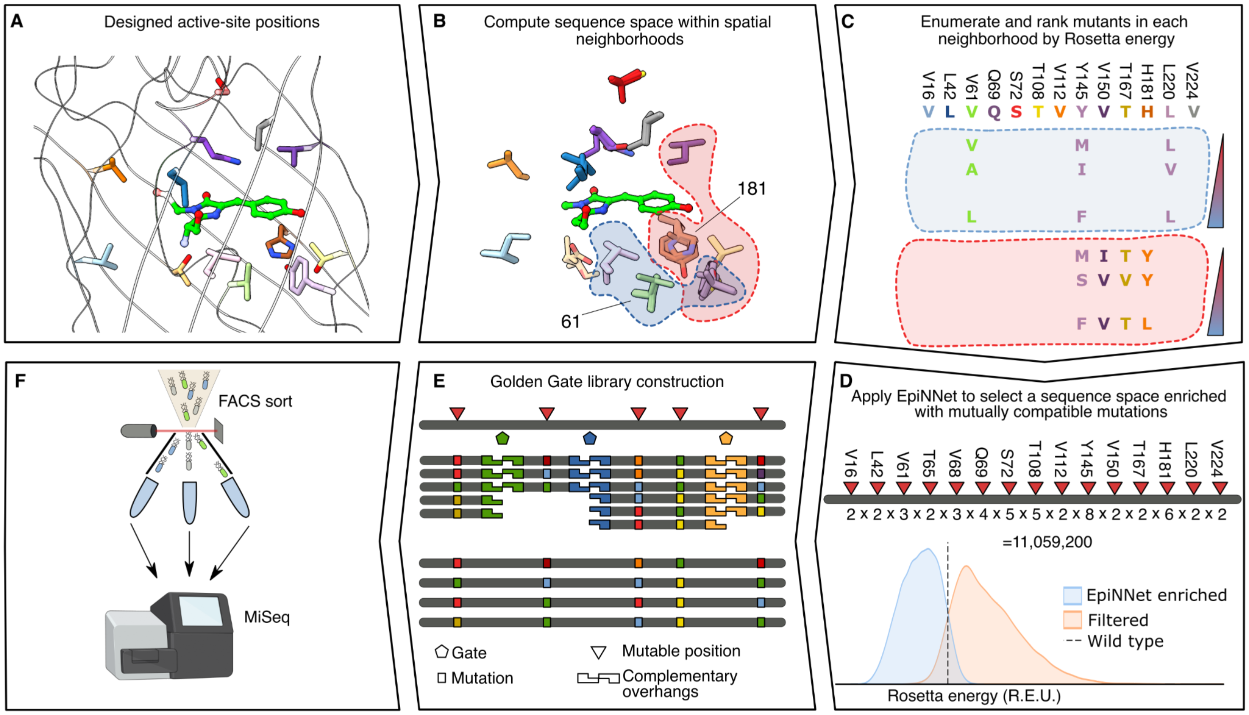

A
Designed active-site positions
B
Compute sequence space within spatial neighborhoods
181
61
C
Enumerate and rank mutants in each neighborhood by Rosetta energy
V16 L42 V61 Q69 S72 T108 V112 Y145 V150 T167 H181 L220 V224
V L V Q S T V Y V T H L V
V M L
A I V
L F L
M I T Y
S V V Y
F V T L
F
FACS sort
MiSeq
E
Golden Gate library construction
Gate
Mutation
Mutable position
Complementary overhangs
D
Apply EpiNNet to select a sequence space enriched with mutually compatible mutations
V16 L42 V61 T65 V68 Q69 S72 T108 V112 Y145 V150 T167 H181 L220 V224
2 x 2 x 3 x 2 x 3 x 4 x 5 x 5 x 2 x 8 x 2 x 2 x 6 x 2 x 2
=11,059,200
EpiNNet enriched
Filtered
-- Wild type
Rosetta energy (R.E.U.)

Figure 1. Steps in applying htFuncLib to GFP. (A) 14 positions designed by htFuncLib are shown (PDB entry: 2WUR). **(B)** Red and blue backgrounds indicate representative neighborhoods centered around GFP amino acid positions 181 and 61, respectively. **(C)** The sequence space of each neighborhood is partially enumerated. Sequence representation of the two neighborhoods shown in (B). Only variable positions are shown for clarity. Color bars represent Rosetta energies. **(D)** EpiNNet top-ranked mutations are selected as the enriched sequence space. An atomistic verification step scores thousands of random combinations from the EpiNNet-enriched and the filtered sequence spaces. Nearly all designs in the EpiNNet sequence space are predicted to be more stable than PROSS-eGFP, compared to almost none in the filtered sequence space. Red triangles mark the mutable positions, and the number of mutations in each position is marked under the bar. **(E)** The designed library is cloned using Golden Gate assembly[37] of oligos that contain the desired mutations, expressed in *E. coli* cells, and **(F)** sorted by FACS.

Design of a multiplexed GFP active-site library

The spectral properties of GFP depend on chromophore packing, electrostatics, and hydrogen-bond networks around the chromophore[27]. Since hydrogen-bond networks are extremely sensitive to structural perturbations, we designed two libraries: nohbonds, which excluded positions that directly hydrogen bond to the chromophore, and hbonds, which included such positions. We manually selected 27 and 24 positions for design in each library, respectively, applied htFuncLib to these positions, and generated 11 million and 930,000 designs for each library, respectively. Both libraries are complex: some positions allow only subtle mutations, and others, including *e.g.,* Gln69 and Tyr145, exhibit high diversity and radical mutations (Supplementary Figures 2 and 3, Supplementary Table 5). According to Rosetta atomistic modeling, both libraries are highly enriched for low-energy mutants compared to the GFP starting point. For instance, nearly 99% and more than 67% of the nohbonds and hbonds designs, respectively, exhibit lower Rosetta energies than the progenitor PROSS-eGFP (Figure 2A). By contrast, the energies of multipoint mutants from the sequence space prior to EpiNNet enrichment are significantly worse than PROSS-eGFP, with >99% and >96% exhibiting higher energies for nohbonds and hbonds, respectively (Figure 2A). The unfavorable energies of combinatorial mutants in the sequence space before EpiNNet selection reflect the high epistasis in the active site. By contrast, the EpiNNet-enriched sequence space significantly improves the fraction of low-energy and, thus, potentially stable and foldable active-site designs. Additionally, combinations of EpiNNet-selected mutations exhibit lower energies than expected from an additive contribution of the constituting point mutations (Figure 2B).

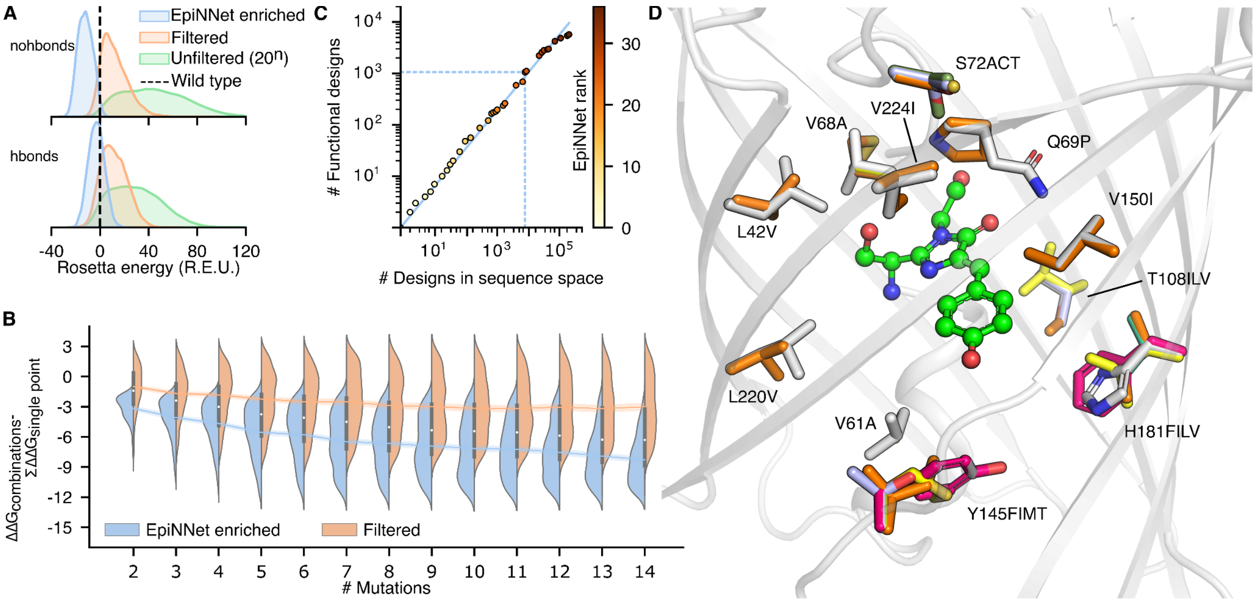


Figure 2. htFuncLib selects mutations that combine to form low-energy designs. (**A**) Energy distributions of the EpiNNet-enriched sequence space, the sequence space filtered by energy and phylogenetic criteria (Filtered), and unfiltered (all 20 amino acids at each position). >95% of mutants in the EpiNNet-enriched combinatorial sequence space exhibit higher stability than PROSS-eGFP, compared to <0.6% for the other spaces. 12,000 randomly selected sequences were modeled to generate each distribution. The dashed line signifies PROSS-eGFP energy. (**B**) Distributions of the energy difference between multipoint mutants and the sum of their constituent point mutations. (**C**) The number of functional designs according to FACS screening of libraries comprising an increasing number of top-ranked EpiNNet mutations is plotted as a function of the total number of designs detected in the deep-sequencing data. Points are color-coded according to the number of mutations that constitute the library. For example, a library of 25 top-ranked EpiNNet mutations that comprise ~10^4 designs would yield approximately 10^3 functional ones (dashed blue lines). The diagonal is the best fit to the data points. (**D**) Overlay of all mutations of the 25 top-ranked EpiNNet mutations from panel C. Despite the relatively small size of this library, it contains radical mutations, including Tyr145Met and Gln69Pro.

The two libraries were cloned using Golden Gate assembly into *E. coli* cells, with transformation efficiency greater than 5 x 10^7. Deep-sequencing analysis of the unsorted libraries shows high uniformity in the distribution of multipoint mutations, verifying that the assembly process exhibits low bias (Supplementary Figure 4). The cells were FACS-sorted using two selection gates: (405 nm excitation, 525 nm emission; referred to as AmCyan$^{405/525}$) and (488 nm, 530 nm; referred to as GFP$^{488/530}$; Supplementary Figure 5). Following selection, plasmids were purified and cloned into fresh cells and resorted using the same gating strategy to reduce sort errors. Following each sort, we collected several individual clones for sequencing and functional measurements, obtaining 62 unique designs, 50 of which were functional. Furthermore, the presorted library and the output from the second sort were subjected to deep sequencing analysis. To determine thresholds for selecting positive hits from the deep sequencing data, we analyzed the enrichment values of the 62 designs we collected during sorts. Relatively loose criteria (enrichment in the selected population relative to the presorted population >1) captured 45 functional designs with only a single false positive (Supplementary Table 6). Applying these thresholds, we identified 14,242 and 1,926 unique designs in the sorted nohbonds and hbonds libraries, respectively (0.13% and 0.21%, respectively; see Supplemental Figure 6 for distribution of read counts in the selected libraries). We also retrospectively evaluated the fraction of functional GFP variants in libraries that were constructed from top-ranked EpiNNet-selected mutations. We found that up to library sizes of 10^4-10^5, approximately 10% of the multipoint mutants were functional, and only above a library size of 10^5 did the fraction of functional variants decay substantially (Figure 2C & D). These results are encouraging as they suggest that focusing htFuncLib on top-ranked mutations may yield highly functional libraries in experimental systems that are not amenable to high-throughput screening.

Combining the positive hits from both libraries yields 16,155 unique, putatively active GFP designs. Remarkably, these include 1,167 designs that exhibit ≥8 mutations relative to GFP (Figure 3A). Strikingly, many of the active designs have radical mutations, including Thr203His (13%), Gln69Met (9%), Ser205Asp (9%), Gln94Leu (8%), and Tyr145Met (8%) (Supplementary Table 7). The large number of functional active-site multipoint mutants is striking compared to previous engineering and design strategies applied to eGFP, which showed a steep decline in active mutants with the number of mutations and no active mutants with ≥5 mutations in the chromophore-binding pocket[4,40] (albeit, these studies did not focus diversity on the active site). The vast majority of the mutations observed in those studies were in the more tolerant solvent-exposed surfaces. By contrast, the current designs are entirely within the chromophore-binding pocket where they are more likely to affect functional properties (Figure 3A). The large number of active high-order multipoint designs in our dataset confirms our working hypothesis that a stable starting scaffold (eGFP-PROSS) and the htFuncLib enrichment of mutually compatible mutations dramatically increase the yield of functional active-site multipoint mutations. Furthermore, htFuncLib generates many more functional multipoint active-site designs relative to random mutagenesis (Figure 3A). Finally, compared to all known descendants of *Aequorea victoria* GFP (avGFP) in the fluorescent protein database (FPBase)[35] and variants characterized in focused and high-throughput studies, we find that htFuncLib explored different regions of the sequence space (Figure 3B).

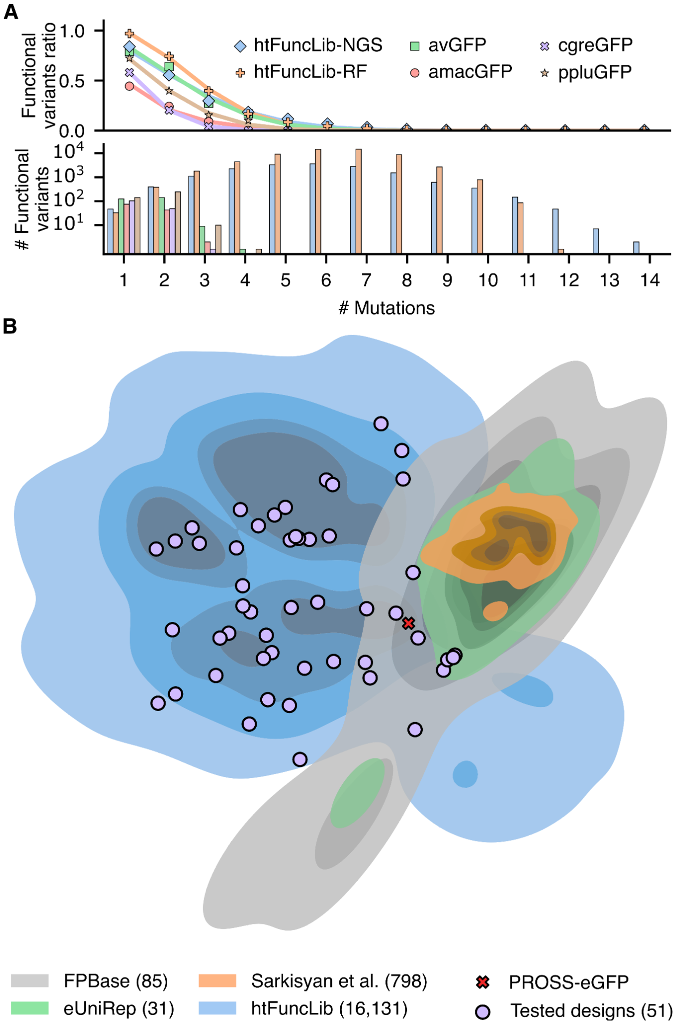


Figure 3. htFuncLib exposes a large space of functional multipoint active-site GFP variants. Deep sequencing of htFuncLib libraries sorted by fluorescence revealed over 16,000 potentially active designs. (**A**) Frequency and number of functional variants with a given number of mutations (top and bottom, respectively). htFuncLib-NGS - all sequences obtained from deep sequencing of the sorted designs; htFuncLib-RF - the entire sequence space labeled by the random forest. The avGFP dataset was derived from Sarkisyan *et al.*[4]. The amacGFP, cgreGFP, and ppluGFP datasets were derived from Somermeyer *et al.*[3]. Lines represent fits to the data (points) according to *Eq*. 2 (see Methods and Supplementary Table 8). Data excluded sequences with mutations outside of the chromophore pocket. (**B**) Distance-preserving dimensionality reduction analysis shows the relationships between GFP variants in FPBase[35], Sarkisyan *et al.*[4], eUniRep[40], and htFuncLib. The plot approximates the number of mutations between any pair of mutants[40,41]. PROSS-eGFP (and eGFP, which are nearly identical in the designed positions, Supplementary Table 12) are marked by a cross for reference. Individually characterized htFuncLib designs are marked by purple circles. The number of sequences represented for each category is marked in parentheses. Variants with mutations outside the chromophore pocket were included, but these mutations were ignored when calculating distances.

Random forest modeling of GFP genotype-phenotype map

To gain insight into what determines the functional outcome of multipoint mutants in the htFuncLib designs, we trained a random forest model using the functional annotations derived from the deep sequencing data for the nohobnds library. We chose this type of analysis because it is easily interpretable, less prone to overfitting than other approaches, and well suited for mixed categorical and numerical data. As features for training, we used the mutation identities, geometric and physicochemical properties, and conservation scores. The best-performing model exhibits 84% accuracy in predicting functional versus non-functional designs in a balanced test (Supplementary Figure 7, Supplementary Table 9). The most important single feature for predicting functionality is the mean conservation score, calculated as the sum of differences in the conservation scores between PROSS-eGFP and mutated identities (ΔPSSM, Supplementary Figure 8). In fact, this single parameter exhibits an area under the ROC curve of 87%, compared to 93% for the random forest. This result provides a compelling verification for the approach of combining sequence conservation with atomistic protein design which underlies htFuncLib and other successful protein design methods developed in recent years[26].

To further understand the qualitative features of the sequence-function relationship learned by the random forest, we used a technique for visualizing complex fitness landscapes[42]. In this technique, the distance between sequences reflects the time it would take for a population to evolve from one sequence to another under selection to maintain a fluorescent phenotype as predicted by the random forest model (see Methods). We found that the main structure of the landscape could be represented by a two-dimensional visualization, where each axis captures a different qualitative feature of the GFP genotype-phenotype map (Figure 4A). The first axis (diffusion axis 1) mainly distinguishes functional from non-functional sequences (79% of sequences with diffusion axis 1 values greater than 1 were functional, while only 0.01% were functional if they had diffusion axis 1 values less than -1), capturing the fact that the functional sequences are highly connected with each other and localized in sequence space rather than consisting of isolated fitness peaks separated by valleys. The contiguity between functional sequences suggests that the htFuncLib selection of mutations that increase stability may generate a highly evolvable library in which active variants are connected via mutational trajectories that maintain function[38,39]. Additionally, the second axis (diffusion axis 2) then largely separates functional AmCyan$^{405/525}$ sequences from functional GFP$^{488/530}$ sequences.

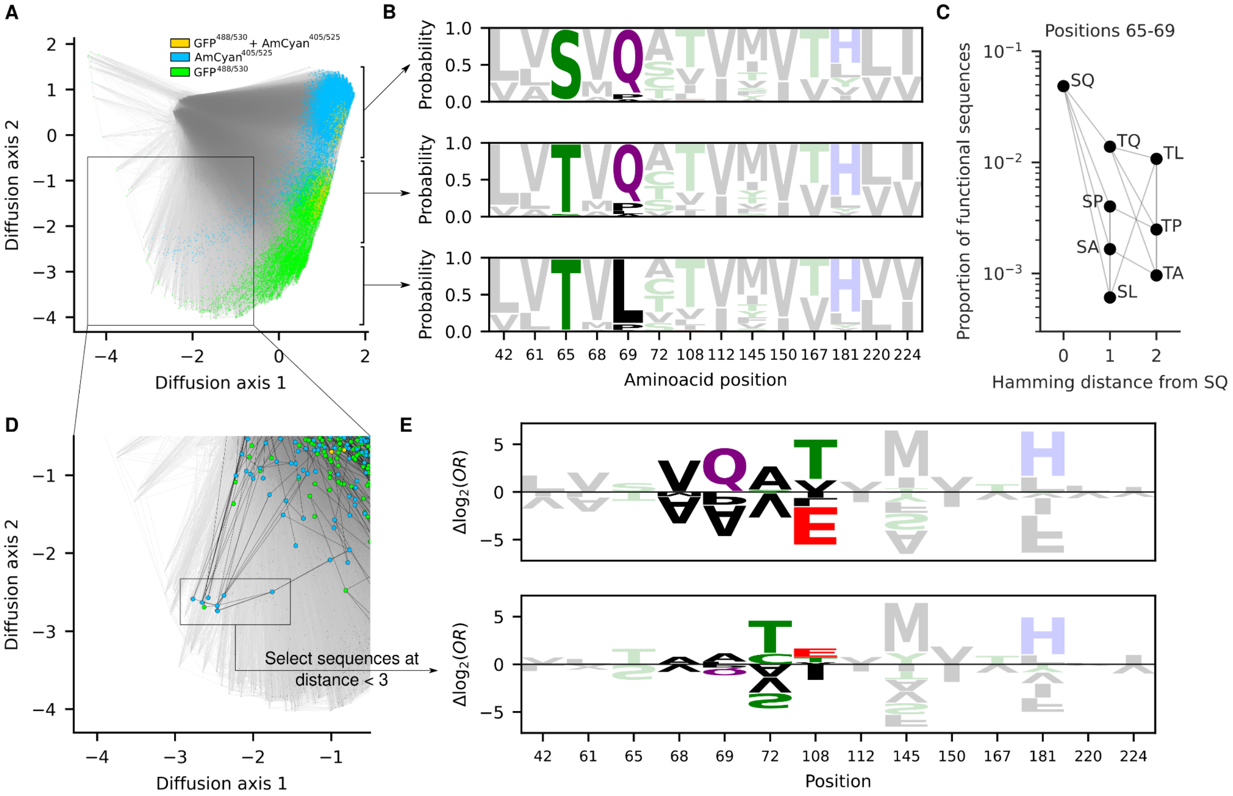


Figure 4. Global analysis of the GFP genotype-phenotype map shows high mutational contiguity among functional sequences. (A) Low-dimensional visualization of the sequence-function relationship predicted by the random forest model (see Methods). Functional sequences are highlighted in different colors according to whether they are predicted to fluoresce in the GFP$^{488/530}$ channel (green), AmCyan$^{405/525}$ channel (blue), or both (gold). Lines join genotypes that are separated by a single amino acid substitution. **(B)** Site-frequency logos of functional sequences based on position along diffusion axis 2 (the three logos correspond to diffusion axis 2 coordinates greater than -0.5, between -0.5 and -2.25, and less than -2.25). **(C)** The proportion of functional sequences changes depending on the amino acids at positions 65 and 69. Gray lines indicate single amino acid substitutions. **(D)** Close-up of the region containing a cluster of observed sequences with unusual sequence properties. Highlighted dots indicate sequences that were directly characterized as functional in the high-throughput experiments, and black lines indicate single amino acid substitutions between these experimentally characterized sequences (see Supplementary Figure 9 for a visualization of all sequences enriched in the high-throughput experiment). **(E)** Sequence logo representing the coefficients of the logistic regression models trained on random forest predictions to identify changes in allelic preferences when using all sequences for training (top) or only sequences within two mutations of the genotypes highlighted in panel D (bottom). Coefficients are expressed as additive allelic contributions (*i.e.*, $\Delta\log_2$ odds ratios) that have been mean-centered by site.

Figure 4B provides more detail on the interpretation of diffusion axis 2 by showing site frequency logos for three different regions of the fitness landscape. These frequency logos indicate that the main set of functional sequences is largely separated into three groups: one group with Thr65Ser and Gln69 consisting of AmCyan$^{405/525}$ designs; one group with Thr65 and Gln69 consisting of designs that fluoresce a mixture of

AmCyan$^{405/525}$, GFP$^{488/530}$, or both; and one group with Thr65 and Gln69Leu that consists of GFP$^{488/530}$ designs. All three groups are strongly supported by many different sequences directly assessed in the sorting experiment (Supplementary Figure 9). Strikingly, Thr65Ser and Gln69Leu are highly incompatible: sequences that contain both these mutations have a much lower chance of being functional (Figure 4C). As a consequence, evolutionary trajectories from Ser65/Gln69 to Thr65/Leu69 that maintain functionality would tend to pass through a Thr65/Gln69 intermediate.

In addition to these three main groups, which comprise the large majority of the functional designs, the random forest analysis predicts two long parallel tails of functional sequences spreading along diffusion axis 1 and sweeping up along diffusion axis 2. The AmCyan$^{405/525}$ tail is well-supported by the experimental data and is not an artifactual prediction of our model, as we observe a cluster of highly mutationally connected designs that were also among the most strongly enriched in AmCyan$^{405/525}$ sorted cells (Figure 4D, Supplementary Figure 9, Supplementary Table 10). Moreover, in this cluster, all sequences contain an unusual and rarely functional pairing of alleles Thr65/Gln69Ala (Figure 4C), and all except one contain Thr108Glu, which is also unusual among other functional sequences (Figure 4B). To investigate what distinguishes these designs from the other fluorescent proteins in the library, we fit an additive logistic regression model to the random forest output using only sequences up to two mutations away from the cluster highlighted in Figure 4D (see Methods). We then compared the estimated mutational effects on the probability of activity to those obtained by fitting the same logistic regression model to the full genotypic space (Figure 4E). Although there are some commonalities in the inferred mutational effects (*e.g.,* Tyr145Met, which is the strongest single-site predictor of functionality based on the random forest, greatly increases the probability that a sequence is fluorescent under both logistic regression models), positions 68, 69, 72 and 108 show marked differences in amino acid preferences. For example, Thr72Ala increases the odds ratio for functionality by approximately fourfold in the general model but reduces the odds ratio by 13-fold in this alternative context. These results suggest that variants within this cluster also differ in their functional constraints as compared to the majority of fluorescent designs, although more detailed experiments would be required to validate this qualitatively different solution to GFP fluorescence.

Designs exhibit large and useful functional diversity

The above results, based on flow cytometry, identify designs that maintain fluorescence, but they do not provide information on other changes in functional properties, including finer-scale changes in excitation and emission spectra. To examine these aspects of functional diversity, we expressed, purified, and characterized a total of 88 unique designs, exhibiting at least two mutations from PROSS-eGFP and typically at least two mutations from one another, and three controls (eGFP, PROSS-eGFP, and superfolder GFP (sfGFP); Supplementary Tables 11 and 12). Twenty-four designs are cluster representatives of the hits observed in the deep-sequencing data, 17 of which (71%) were active. We also selected three designs with mutations rarely found in the sorted populations, Glu222Gln/Leu and Leu44Met, one of which was active. As an especially stringent test, we selected six designs with the maximal number of mutations (12-14), but none of these was functional. Furthermore, we selected 19 designs that were predicted to be functional by the random forest analysis but were not observed among the positive hits in the deep sequencing analysis. Surprisingly, 15 (79%) were active, confirming that a random forest analysis based on deep sequencing data of htFuncLib designs can be used to recover false negatives — active designs that were missed by the experimental workflow. Additionally, we isolated designs from FACS sorts that were gated for higher brightness or spectral shifts (Supplementary Table 11) by applying sorting gates that combine two channels (Supplementary Figure 10). We also verified that 19 designs could be transferred to the superfolder GFP (sfGFP) background[43] to demonstrate that the designs are compatible with a different chassis (Supplementary Table 12).

Although we did not explicitly guide the design process to improve any functional property (except native-state stability (Figures 1D and 2A)), we hypothesized that the large diversity in active-site sequences would lead to observable functional differences. We first analyzed GFP functional thermostability or the temperature at which its fluorescence deteriorates to 50% of the maximal value, a critical property for high-temperature or long-term experiments and “real-world” applications[44,45]. Functional thermostability is remarkably variable among the designs, 46-96 ℃, compared to 84 ℃ for eGFP (Figure 5A and Supplementary Figure 11). We noticed that the PROSS-eGFP parental design is less stable than eGFP when functional thermostability is measured (Figure 5A) rather than thermal denaturation as in the PROSS-eGFP design study[33]. Apparently, the PROSS-eGFP design is more resistant to heat denaturation, but its fluorescence is more sensitive to heat than eGFP. Quantum yield, which measures the efficiency of emitting light absorbed by the chromophore, was also extremely variable, 0.16-0.82, compared to 0.55 for eGFP (Figure 5A and Supplementary Figure 12). Surprisingly, across the all designs we tested, functional thermostability and quantum yield were correlated (Pearson’s r=0.53, Supplementary Figure 13). This correlation probably stems from the fact that both chromophore brightness and resistance to unfolding increase with core packing density[31,46]. To our knowledge, this is the first observation of such a correlation, demonstrating how a large set of active-site variants can yield insights even in a well-studied protein. Moreover, the designs we sorted specifically for spectral shifts indeed displayed significant shifts in excitation spectra (Figure 5B, Supplementary Tables 11 and 12, and Supplementary Figure 14).

We examined the design models for a molecular explanation of the large observed differences in stability and quantum yield. For instance, Tyr145Phe, seen previously to enhance stability and quantum yield[47], appeared in all five high stability/brightness designs but only in one of the bottom 26 designs. Similarly, Thr203His, likely to stabilize the chromophore through π-π stacking interactions[48], is seen in all top designs and none of the bottom ones. Ser205Thr is in three of the top-five designs and none of the bottom. By contrast to the two mutations above, Ser205Thr is enriched in designs with high thermostability and quantum yield though we are unaware of previous studies that pointed to its significance.

We also observed large variability in photostability, which is the resistance of the chromophore to bleaching by bright light. Bleaching is often a limitation in long-term live-imaging studies[49], whereas it is an advantage in assays such as fluorescence recovery after photobleaching (FRAP), in which fast fluorescence decay enhances signal[50]. We isolated two designs that exhibited higher photostability than GFP (photostable.1 & photostable.2, with seven mutations each from PROSS-eGFP) and many significantly less photostable designs (Figure 5A and Supplementary Figure 15). At the extremes, design fast.4 (6 mutations) photobleaches tenfold faster than GFP, while the design photostable.1 requires 122% of that time. Finally, we also noted large differences in fluorescence lifetime (Supplementary Figure 16) and pH sensitivity (Supplementary Figure 17). Furthermore, several mutations enriched in designs with altered pK_a are either adjacent to His amino acids or introduce a novel His (Thr203His). Interestingly, seven designs exhibit different pH sensitivity profiles when excited at either 405 or 488nm (Supplementary Figure 17).

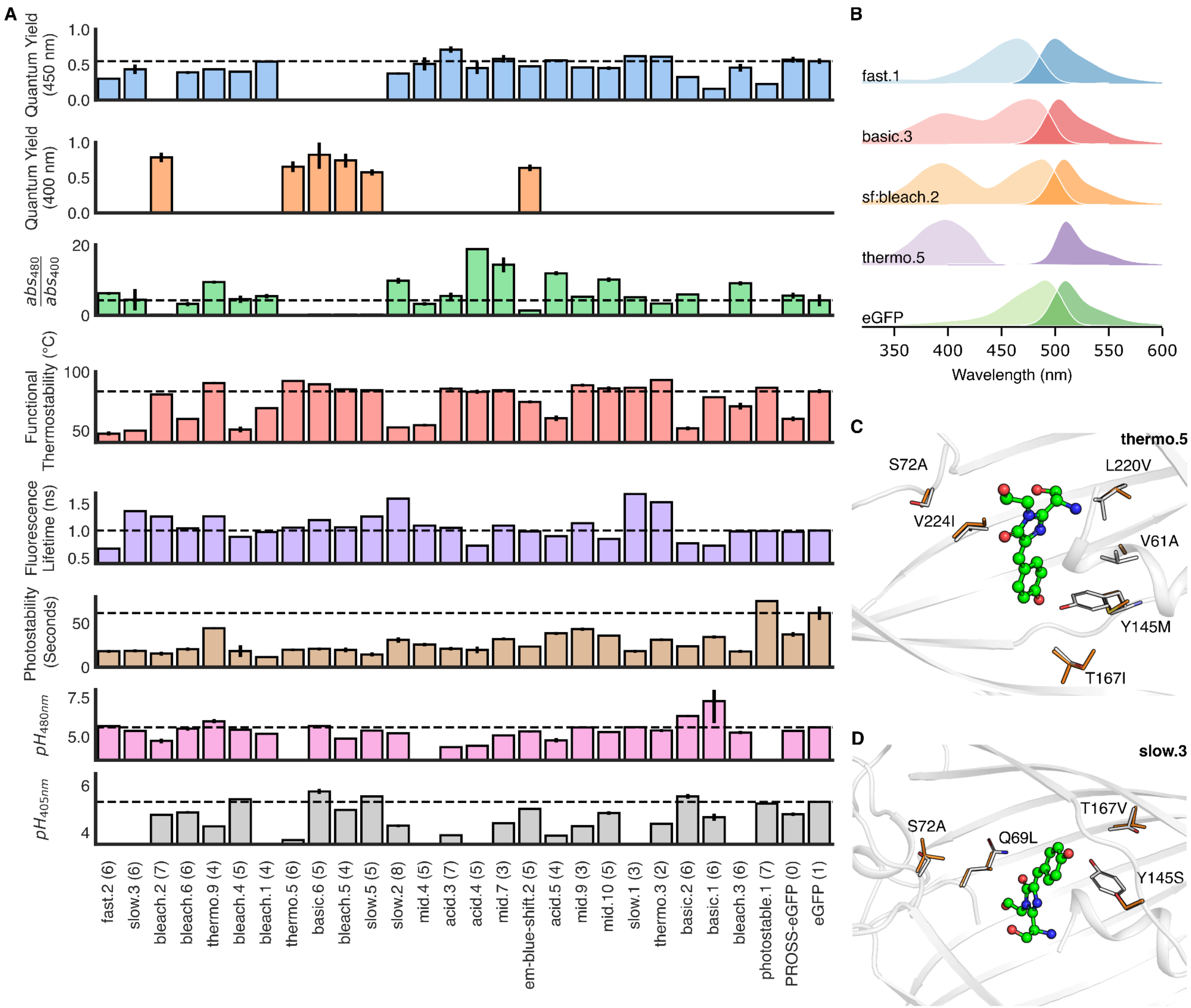


Figure 5. Functional diversity among htFuncLib designs. (**A**) A subset of tested designs clustered by sequence similarity. The dashed line marks eGFP, error bars mark standard deviations. (**B**) Selected excitation and emission spectra in light and dark hues, respectively. The excitation spectra of several designs are considerably different from eGFP (Supplementary Figure 14). (**C & D**) Structural view of thermo.5 and slow.3. Each design exhibits six mutations from PROSS-eGFP). PROSS-eGFP and designs colored gray and orange, respectively.

Discussion

Epistasis is a significant constraint on the emergence of new activities in proteins and other biomolecules[15]. Until now, experimental methods to address epistasis have relied on iterative cycles of diversification and selection, but such processes do not efficiently cover the space of functional variants. Computational methods have used evolutionary couplings among pairs of positions[51], but such analyses require deep and diverse sequence alignments, which are not generally available. Other approaches have used machine-learning models trained on high-quality and large-scale mutational data to recommend mutations[52–54]. By contrast, htFuncLib only requires a molecular structure (or model) and a limited sequence alignment of homologs. Its success in generating an order of magnitude more functional active-site mutants than were previously known for GFP verifies our underlying assumption that energetically incompatible mutations are a significant source of epistasis. Furthermore, because the designs are diverse and only target the chromophore-binding pocket, they exhibit potentially useful functional diversity in each of the properties we assayed.

Our implementation of htFuncLib did not target a specific functional outcome, except for protein stability. This implementation is especially suitable if multiple variants for different and potentially incompatible goals are desired. For example, FRAP experiments require fluorescent proteins that bleach quickly, whereas long-term imaging experiments require slow bleaching, and we recovered designs that exhibited both properties from a single library. If a specific functional goal is desired and the molecular underpinnings of that goal are known, they can be imposed during the design process to focus diversity such that it exhibits these essential molecular features. The high stability and brightness of the eGFP starting point are likely to be key to obtaining so many functional variants[3,55]. Further research is needed to determine whether the combination of PROSS stability design[56] and htFuncLib can access such large spaces of functional variants in less robust starting points.

In a companion paper, we demonstrate that the EpiNNet strategy is general and can be extended to design large and highly functional enzyme libraries comprising substantial backbone conformational diversity, including insertions and deletions[36]. We envision that htFuncLib will provide a platform for designing high-yield multipoint mutation libraries in a range of applications, including optimizing binding affinity and enzyme catalytic rate and selectivity.

Acknowledgments

We thank Moshe Goldsmith, Robert Campbell, and Yi Shen for helpful advice and Olga Khersonsky and Shiran Barber-Zucker for critical reading. Research in the Fleishman lab was supported by the Volkswagen Foundation (94747), the Israel Science Foundation (1844), the European Research Council through a Consolidator Award (815379), the Dr. Barry Sherman Institute for Medicinal Chemistry, and a donation in memory of Sam Switzer. Research in the McCandlish lab was supported by NIH grant R35GM133613, an Alfred P. Sloan Research Fellowship, and additional funding from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

Methods

Library design

Phylogenetic analysis

A phylogenetic analysis was conducted as previously[24,56,57] using all sequences in the lineage of avGFP according to FPBase[35]. A total of 153 seqeunces were retrieved from FPBase, all synthetic variants of avGFP. Briefly, sequences were clustered by cd-hit[58] and aligned using MUSCLE[59]. The resulting multiple sequence alignment (MSA) was segmented by secondary structure elements. A position-specific scoring matrix (PSSM)[60] was then derived from the MSA segments and concatenated to create a PSSM for the whole sequence. The PSSM is used to filter mutations absent in the PSSM at each position and to bias the Rosetta energy function towards mutations favored by the PSSM (high PSSM score).

Refinement and mutational scan

PROSS-eGFP was modeled based on a high-resolution X-ray structure of eGFP (PDB code: 2WUR). The eGFP-PROSS model was subsequently refined in Rosetta as described before[56]. Chromophore pocket positions were then manually selected, 14, 16, 18, 42, 44, 46, 61, 64, 66, 68, 69, 72, 108, 110, 112, 119, 123, 145, 150, 163, 165, 167, 181, 185, 201, 220, 224 and 42, 44, 61, 62, 69, 92, 94, 96, 112, 121, 145, 148, 150, 163, 165, 167, 181, 183, 185, 203, 205, 220, 222, 224 for the hbonds and nohbonds libraries, respectively. All positions are within 8 Å from the chromophore, and their side chains are buried within the GFP β-barrel. All mutations with PSSM scores > -2 were then scanned *in silico*, as previously described[24,56]. Briefly, each mutation is modeled, refined, and scored separately on the PROSS-eGFP background. This step calculates the $\Delta\Delta G$ between the mutant and eGFP-PROSS.

Spatial partitioning and sequence space selection

We split the chromophore pocket into spatial neighborhoods, with each selected position as a center of a distinct neighborhood. In order to capture direct epistatic interactions, each neighborhood is extended to all positions that interact directly with the neighborhood's center. Here, direct interaction is defined as having at least two heavy atoms within 6 Å of the neighborhood's central residue. Neighborhoods were manually examined, and positions that did not interact directly with the neighborhood's center were removed. By selecting neighborhoods this way, we ensure overlap between proximal neighborhoods. These overlaps ensure that no position-position interactions are missed.

Partial modeling and scoring

The number of designs encoded in each neighborhood is calculated for each $\Delta\Delta G_{mut-wt}$ threshold. The energy threshold is selected to limit the number of unique variants to under 1 million. In this particular case, the $\Delta\Delta G$ thresholds were set to +5.5 and +6.0 Rosetta energy units (R.e.u.) for nohbonds and hbonds, respectively. Neighborhoods with a sequence space smaller than 10,000 designs were fully modeled. For larger neighborhoods, only 10% of the sequence space was modeled. RosettaScripts[61] and command-line arguments for modeling calculations are in the Supplementary Information.

Data aggregation and EpiNNet training

We train an EpiNNet neural net model to predict which designs are more stable than PROSS-eGFP. Specifically, designs that score better than the wild-type are labeled as success (1), and the worse 50% are labeled as failed (0). Intermediate designs are considered undetermined and discarded from subsequent analysis. The resulting data are split into a training (80%) and a test (20%) set. We then train a multi-layer perceptron classifier with a single hidden layer the size of the number of selected positions. The classifier is trained on a one-hot encoded representation of the sequence data to classify whether a sequence is more or less stable than PROSS-eGFP. The classifier is trained up to 2,000 iterations. Next, we rank single-point mutations according to the trained model: each single-point mutation in the tolerated sequence space is fed into EpiNNet separately, and its score is recorded. The mutations are then ranked from top to bottom according to their scores. Mutations are selected for the library by iteratively adding the top-ranked mutations until the resulting sequence space reaches the experimental limit of several million sequences.

In silico testing of the enriched versus the original sequence spaces

To ensure the resulting sequence space is enriched for low-energy sequences, 10,000 random sequences from both the original and enriched sequence spaces were modeled and scored (using the same protocol as in the modeling step). The resulting score distributions were compared (Figure 1D and Figure 2A).

Random forest

To augment the sequence data for machine learning prediction, we added several features based solely on the sequence and not requiring atomistic calculations. These include the amino acid identity at each variable position, the total number of mutations compared to PROSS-eGFP, the number of mutations at each spatial neighborhood, and the number of mutations in specific areas. In addition, for every variable position, the difference in the surface accessible solvent area (SASA), PSSM score, and amino acid category were also assigned (comparing the mutated amino acid and the PROSS-eGFP identity). The mean and max values of each of these parameters were added as well. Non-informative features and features with low importance in initial random forest training were removed. A prediction pipeline with two consecutive elements was trained. The first predictor classifies sequences as either functional or non-functional. The subsequent predictor classifies all functional sequences as either GFP, AmCyan, GFP/AmCyan, or non-functional. Both models are gradient-boosting random forests from the LightGBM library[62].

Visualization methods

Visualization method as previously described[42]. Briefly, we construct a model of molecular evolution where a population evolves via single amino acid substitutions, and the rate at which each possible substitution becomes fixed in the population reflects its selective advantage or disadvantage relative to the currently fixed sequence. More specifically, in our model, the rate of evolution from sequence *i* to any mutationally adjacent sequence *j* is given by

$$Q_{ij} = \frac{S_{ij}}{1 - e^{S_{ij}}}$$

where S_{ij} is the scaled selection coefficient (population size times the selection coefficient of *j* relative to *i*), time is measured relative to the amino acid mutation rate (each possible amino acid mutation occurs at rate 1), and the total leaving rate from each sequence *i* is given by $Q_{ii} =- \sum_{j \neq i} Q_{ij}$. In the current context, sequences are either predicted to be fluorescent or not, and so we set $S_{ij} = c$ if *j* is fluorescent and *i* is not, $S_{ij} =- c$ if *i* is fluorescent and *j* is not, and otherwise $S_{ij} = 0$ so that $Q_{ij} = 1$, corresponding to neutral evolution. For this analysis, we choose c so that in the long-term, a population spends 60% of its time at functional sequences, representing a roughly 60-fold enrichment of functional sequences due to natural selection.

Given the rate matrix Q for our evolutionary model, we then construct the visualization by using the subdominant right eigenvectors associated with the smallest magnitude non-zero eigenvalues of this rate matrix as coordinates. This produces a visualization that reflects the long-term barriers to diffusion in sequence space, and, in particular, clusters of sequences in the visualization correspond to sets of initial states from which the evolutionary model approaches its stationary distribution in the same manner, and multi-peaked fitness landscapes appear as broadly separated clusters with one peak in each cluster. Moreover, by scaling the axes appropriately, as is done here, these axes can be given units of sqrt(time), and it can be shown that the resulting distances reflect evolutionary times under this model. In particular, using these coordinates, the squared Euclidean distance between arbitrary sequences *i* and *j* optimally approximates (in a specific sense) the sum of the expected time to evolve from *i* to *j* and the expected time to evolve from *j* to *i*. See ref. [42] for details.

Logistic regression and sequence logos

Calculations and plots were performed using gpmap-tools python library (https://gpmap-tools.readthedocs.io/en/latest/). Sequence logos were plotted using logomaker[63], and L2-penalized logistic regression models were fit using scikit-learn [64]. Specifically, the global model using all sequences was fit using non-penalized regression, while the model in the neighborhood of the alternative functional sequences highlighted in Figure 4D was fit using L2-penalization under one-hot encoding, using 10-fold cross-validation to optimize the hyperparameter controlling the strength of the regularization. The regularization constant was chosen to be C=0.5 as the strongest regularization before a drop in the cross-validated AUROC.

Experimental procedures

Library cloning

Each designed library was cloned separately using a Golden Gate assembly (manuscript in preparation). A computational algorithm optimizes a set of Golden-Gate gates to minimize the total cost of ordered oligos required to cover all mutations in the library without introducing unwanted mutations. This results in several variable and constant segments, with and without mutations. Constant segments were PCR amplified with primers adding *BsaI* recognition sites. These and all other DNA fragments were purified using (HiYield Gel/PCR DNA Fragments Extraction Kit, Real Genomics). Variable segments were ordered as degenerate oligos (IDT). The single-stranded oligos were double-stranded by a short PCR with a single primer and purified. The concentration of each segment was measured using NanoDrop One (Thermo Scientific). A Golden-Gate assembly was conducted using the manufacturer's specifications. Briefly, all segments are added at an equal amount, without the vector, and assembled using T4-ligase and BsaI-HF-v2 using cycles of 16℃ and 37℃ (New-England Biolabs). The resulting assembly is PCR amplified to add the final gates and assembled into a pBAD vector with appropriate gates.

FACS sorting

E. coli BL21 (DE3) (E. cloni EXPRESS BL21 (DE3), Lucigen) cells were transformed with the pBAD plasmids containing the libraries and grown overnight. Transformation efficiency was estimated by plating serial dilutions of the transformed bacteria, ensuring that, for each library, the number of transformed cells was at least tenfold higher than the designed library size. 1 µl from each transformation was plated in dilution to estimate transformation efficiency. Cells were diluted 1:200 in 2YT media, grown to 0.6-8OD, induced using 0.2% arabinose, and shaken at 20℃ overnight. Induced cultures were transferred to 4℃ for another night to allow maturation. Cells were centrifuged at 3000RPM for 10 minutes, decanted, and resuspended with cold PBS twice. The cells were then sorted using a FACS AriaIII (BD Biosciences) with a 70 µm diameter nozzle and a cell flow rate of 10,000-20,000 events per second. A preliminary sorting gate was done on forward scatter (FSC) Vs. side scatter (SSC) parameters to select single bacteria cells alongside the AlexaFluor488 (excitation at 488 nm, emission detection at 530±15 nm) and AmCyan (excitation at 405 nm, emission detection at 525±25 nm) channels. Sorted cells were collected in SOC media (2.5 mM KCl, 20 mM glucose, 0.5% yeast extract, 2% tryptone, 10 mM MgCl2, 10 mM MgSO4, and 10 mM NaCl), grown overnight at 37℃ and transferred to 2YT supplemented with ampicillin. Plasmids were extracted by mini-prep (Qiagen).

Plasmids from sorted populations were extracted by min-prep, transformed and sorted again (using the smae procedure) to reduce false-positives.

Deep sequencing

Plasmids from presorted and sorted populations were PCR amplified using primers to generate 590bp amplicons, containing all variable positions excluding position 16 (forward primer: GGGCGATGCCACCTACGGCAAG and Reverse primer: GAGTGATCCCGGCGGCCTC). Amplicon libraries were prepared at the Weizmann Institute’s Israel National Center for Personalized Medicine. Libraries were prepared from 20 ng of DNA, as previously described[65]. Libraries were quantified by Qubit (Thermo fisher scientific) and TapeStation (Agilent). Sequencing was done on a Miseq instrument (Illumina) using a V3 600 cycles kit, using paired-end sequencing. Sequences were analyzed using the LAST software package and python[66,67]. Fastq sequences were aligned to all designed oligos using the LAST align function. Sequences were consequently filtered for low LAST scores, and assigned to the best aligned oligo. Pair-end reads were identified using MiSeq UMIs (unique molecular identifiers). Enrichment values were calculated as the ratio between read frequencies in the sorted and appropriate unsorted samples. The presorted libraries are too large to be completely covered by the deep sequencing analysis. We, therefore, did not expect to detect all combinations, specifically in the nohbonds library. However, given that the

transformation efficiency was greater than 5x10^7, and > 10^8 cells were sorted by FACS, it is likely that the majority of the functional designs were recovered. We thus considered all sequences found solely in the sorted samples to be enriched.
The sorted library will be deposited in AddGene upon publication.

Cloning of single designs

Genes encoding for selected designs were ordered from Twist Bioscience and codon-optimized for *E. coli.* Genes were inserted in the pET28 vector using *BsaI* restriction sites previously cloned using QuickChange. All plasmids were sequence verified. Designs selected directly from FACS sorting were transferred from the pBAD vector into pET28 by PCR amplifying the insert with primers and adding *BsaI* recognition sites. Amplicons were purified and inserted into a pET28 vector with *BsaI* insertion sites using Golden Gate assembly. All plasmids were individually verified using Sanger sequencing.

Protein expression & purification

pET28 plasmids containing the relevant insert were transformed into BL21 (DE3) cells and grown overnight. Overnight cultures were diluted 1:100 in 10ml conical tubes containing 2YT and 50ug/ml kanamycin, grown to OD=0.6-8, induced using 1mM IPTG, and shaken overnight at 20℃. After expression, samples were shaken at 4℃ to maximize chromophore maturation. Samples were centrifuged at 4000RPM for 20 minutes at 4℃ and resuspended in 1ml lysis buffer containing PBS, 0.01% Triton x100, 0.02% Benzonase, 0.1% protease inhibitor cocktail, and 0.1mg/ml lysozyme. Samples were then sonicated and centrifuged at 14,000RPM at 4°C for 45 minutes. 500µl Ni-NTA beads per sample were resuspended in PBS and allocated into an appropriate number of 1.7ml tubes. The supernatant of each sample was transferred to a tube containing 500µl Ni-NTA beads and 10mM imidazole. Samples were shaken at room temperature for two hours for binding, centrifuged at 3,000 RPM for 3 minutes, and the supernatant was removed. Beads were resuspended in PBS with 20mM imidazole and shaken for 30 minutes at room temperature. Samples were centrifuged again at 3,000RPM for 3 minutes, and the supernatant was removed. Samples were eluted using PBS with 500mM imidazole, shaken for 1 hour at room temperature, and centrifuged at 3,000 RPM for 5 minutes. The supernatant was recovered and kept at 4℃. Protein purity was estimated by SDS-PAGE gel electrophoresis, and protein concentration was determined using NanoDrop One (Thermo Scientific).

Functional thermostability

Functional thermostability was measured similarly to SYPRO orange measurements[68]. 10µM of each design were diluted in PBS in triplicates and placed in a 96-well plate (20

MicroAmp Fast Optical 96W Reaction Plate, Thermo Fisher, and MicroAmp Optical Adhesive Film). A ViiA7 real-time PCR instrument (Applied Biosystems) was used to measure fluorescence during heating from 25-99.9℃ at 0.05℃/second. Raw data were analyzed using Python to find the temperature at which fluorescence was 50% of the max for each well.

Fluorescence lifetime

Fluorescence lifetime measurements were performed using a MicroTime200 optical setup. GFP samples were placed as drops on top of 175 µm glass slides (Precision Cover Glass No:1.5H, Marienfeld), mounted on an inverted microscope (IX83 inverted, Olympus) with a 60X water immersion objective (UPlanSApo, Superapochromat, Olympus. A 485 nm pulsed-interleaved excitation laser (LDH-D-C-485, PicoQuant) with a repetition rate of 20MHz (50 ns) was directed via a dichroic mirror (ZT473/594rpc, Chroma) and focused ~10 µm into the sample. The fluorescence emission signal passed through a 50 µm pinhole and an emission filter (HC520/35, Semrock). Photons were focused into a single-photon avalanche diode (SPCM-AQRH-14-TR, Excelitas) coupled to a counting module (PicoHarp 400, PicoQuant), and time-correlated single-photon counting (TCSPC) histograms were generated. Each sample was measured for 1-5 min with laser intensities between 2-20 µW, adjusted using OD filters to reach a photon count rate of ~20 kHz. The profile for the instrument response function (IRF) was obtained by measuring scattered light from a mirror. The fluorescence decay curves were fitted with a bi-exponential fluorescence decay model by iterative IRF-reconvolution to extract the characteristic lifetimes and weights of the GFP designs.

Photobleaching

Photobleaching was measured similarly as previously described[69]. A final concentration of 1µM of each variant was embedded in polyacrylamide gels (168µl 30% acrylamide/bis-acrylamide, 25µl PBS, 0.5µl TEMED, and 3µl 10% APS and 57µl fluorescent protein in PBS) inside 8-well microscope slides (ibidi No. 80826). Slides were mounted to Eclipse TI-E Nikon inverted microscope (Nikon Instruments Inc., Melville, NY) with Plan Apo DIC 60X/1.4 NA objective and equipped with a cooled electron-multiplying charge-coupled device camera (IXON ULTRA 888; Andor). The measurement consisted of six repetitions of exposure to the strongest available LED light at either 405 or 488nm for 15 minutes while capturing an image every five seconds. Images were analyzed using ImageJ to recover the mean intensity from each frame. A bi-exponential function was fitted to normalized brightness as a function of exposure time. The weighted average of the exponential coefficients was calculated. Outliers were removed, and at least three measurements were used to calculate means and standard deviations.

Fluorescence spectra and quantum yield

Proteins were diluted in PBS to OD 0.05 at either 450 or 400nm in disposable fluorescence cuvettes (ordered from Alex Red No CUV010015) in triplicates. OD was measured on a Cary 60 UV-Vis spectrophotometer (Agilent Technologies) from 300 to 650nm. Both emission and excitation spectra were measured with the same samples on a fluorescence spectrophotometer (Varian Cary Eclipse). Quantum yield was calculated using the relative method described in the literature[69,70]. Briefly, the ratio between absorbance at the excitation wavelength and the integral of emission spectra are measured for each sample and a standard with known quantum yield. Fluorescein and 1-aminoanthracene were used as standards for measurements at excitation wavelengths 450 and 400nm, respectively.

pH sensitivity

Buffers at pH ranging from 3.0 to 10 were prepared as previously reported[69]. 100µl of pH buffer were placed in black flat-bottom 96-well plates (Greiner Bio-One, No 655090), and 2µg of fluorescent proteins were added. Samples were incubated at room temperature for one hour, and fluorescence at both 405 and 488mn and emission at 520nm was measured for all wells (Infinite M Plex, Tecan).

Multipoint mutants from other GFP datasets

To compare the GFP variants considered here with that studied earlier, we extracted from previous works (refs. 17,18) sequences and fluorescences of variants having mutations in the chromophore pocket positions only (corresponding to 2WUR GFP positions 14, 16, 18, 42, 44, 46, 61, 62, 63, 64, 66, 68, 69, 72, 92, 94, 96, 108, 110, 112, 119, 121, 123, 145, 148, 150, 163, 165, 167, 181, 183, 185, 201, 203, 205, 220, 222, 224). The four GFP variants are *Aequorea victoria* GFP (avGFP[4]), *Aequorea macrodactyla* GFP (amacGFP[3]), *Clytia gregaria* (cgreGFP[3]), and *Pontellina plumata* GFP (ppluGFP[3]). Our reference GFP sequence was aligned to the alignment of the avGFP, amacGFP, cgreGFP, and ppluGFP.

Fluorescence Vs. the number of mutations in the active site

The fluorescence F of different GFP variants was fit to the exponential-decline and negative-epistasis function[39]:

$$F = exp\left(-\alpha n - \beta n^{2}\right), \quad (1)$$

To get a clearer interpretation of the coefficients α and β, we have rewritten the Eq. (1) as:

$$F = exp\left(- An - B\frac{n(n-1)}{2}\right), \tag{2}$$

Where $A = \alpha - \beta$ is responsible for robustness, and $B = 2\beta$ is responsible for the epistasis. We required $A \geq 0$ and $B \geq 0$. The analysis is provided as a Jupyter notebook 'GFP_threshold_epistasis.ipynb'.

Data availability

All sorted and presorted libraries, together with designs of special interest, were deposited to AddGene (deposit number 81660). Deep-sequencing data is available on figshare at 10.6084/m9.figshare.21922365.

Code availability

Jupyter notebooks for the evolutionary analysis can be found at https://bitbucket.org/cmartiga/gfp_core/src/master. Jupyter notebooks, including data, for other analyses and the htFuncLib algorithm, are available at https://github.com/Fleishman-Lab/htFuncLib.

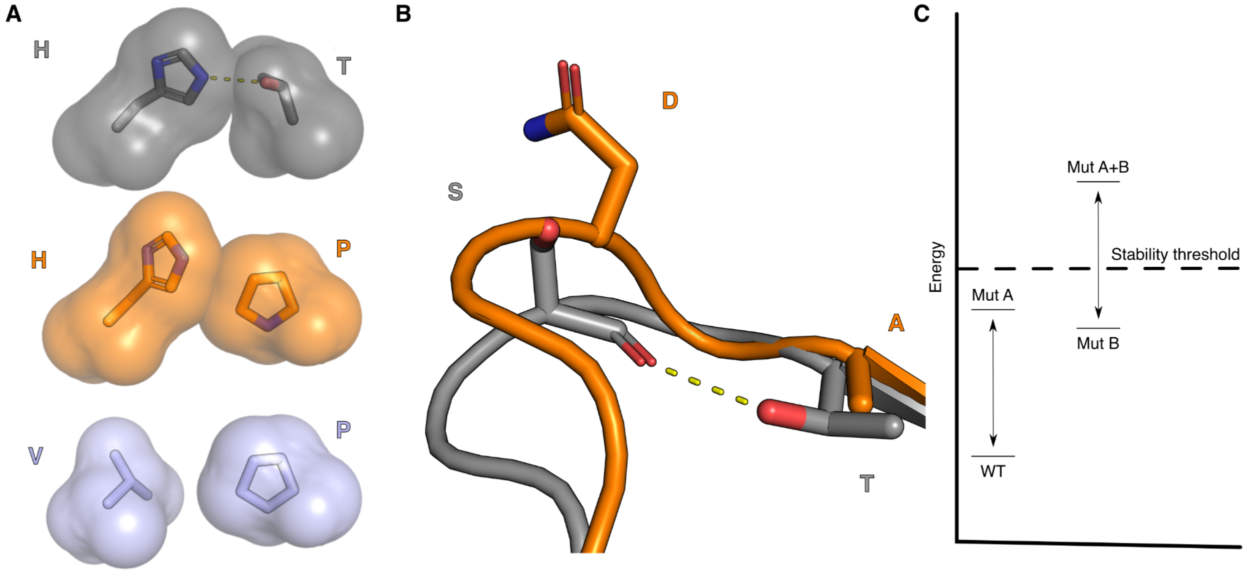
Supplementary Information

RosettaScripts script and flags for modeling combinations of mutations. Every combination of mutations was modelled using a command based on:

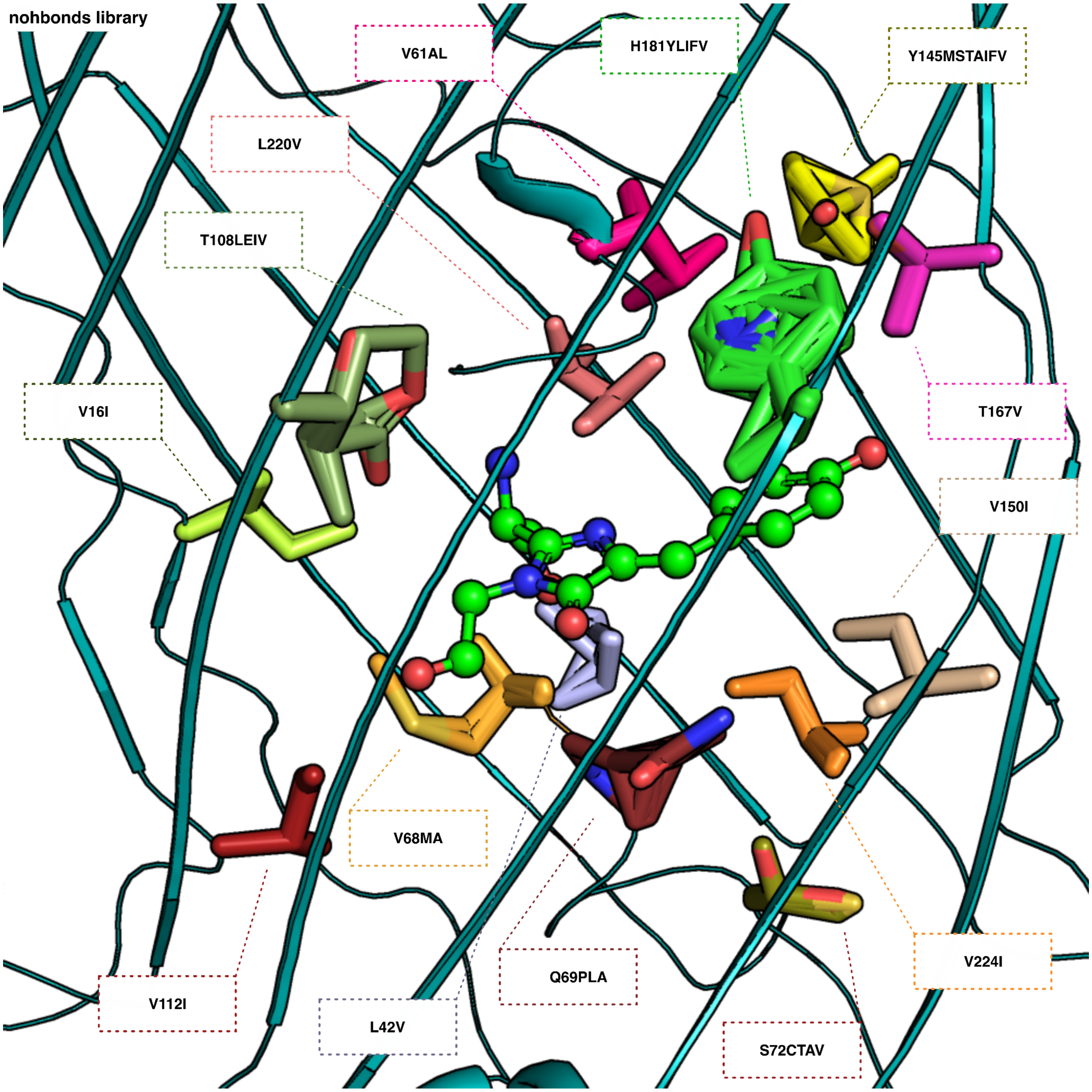
```
rosetta_scripts_executable -database path_to_database -pdb_gz -overwrite -use_input_sc
-extrachi_cutoff 5 -ignore_unrecognized_res
-chemical:exclude_patches LowerDNA UpperDNA Cterm_amidation SpecialRotamer VirtualBB ShoveBB
VirtualDNAPhosphate VirtualNTerm CTermConnect sc_orbitals pro_hydroxylated_case1
pro_hydroxylated_case2 ser_phosphorylated thr_phosphorylated tyr_phosphorylated tyr_sulfated
lys_dimethylated lys_monomethylated lys_trimethylated lys_acetylated glu_carboxylated
cys_acetylated tyr_diiodinated N_acetylated C_methylamidated MethylatedProteinCterm -linmem_ig
10 -ignore_zero_occupancy false -no_nstruct_label true -in:file:native refined_pdb
-extra_res_fa LG.params -nstruct 30 -out:prefix NAME_
-s refined_pdb -use_occurrence_data -parser:protocol mutate_all_poss.xml -mute all
-parser:script_vars res_to_fix=94A,96A,121A,148A,203A,205A,222A,1X cst_full_path=ref_coord.cst
ignore_pose_profile_length_mismatch=1 min_aa_probability=-2 keep_n=1
all_ress=14A,16A,18A,42A,44A,46A,61A,64A,68A,69A,72A,108A,110A,112A,119A,123A,145A,150A,163A,1
65A,167A,181A,185A,201A,220A,224A,42A,44A,61A,62A,69A,92A,94A,96A,112A,121A,145A,148A,150A,163
A,165A,167A,181A,183A,185A,203A,205A,220A,222A,224A -parser:script_vars target1=POS
new_res1=3_LETTER_AA
```

Where every mutation is listed as a separate target# and new_res#, the LG.params files is a parameters file a small ligand, in GFP's case, it is the chromophore. The script is:

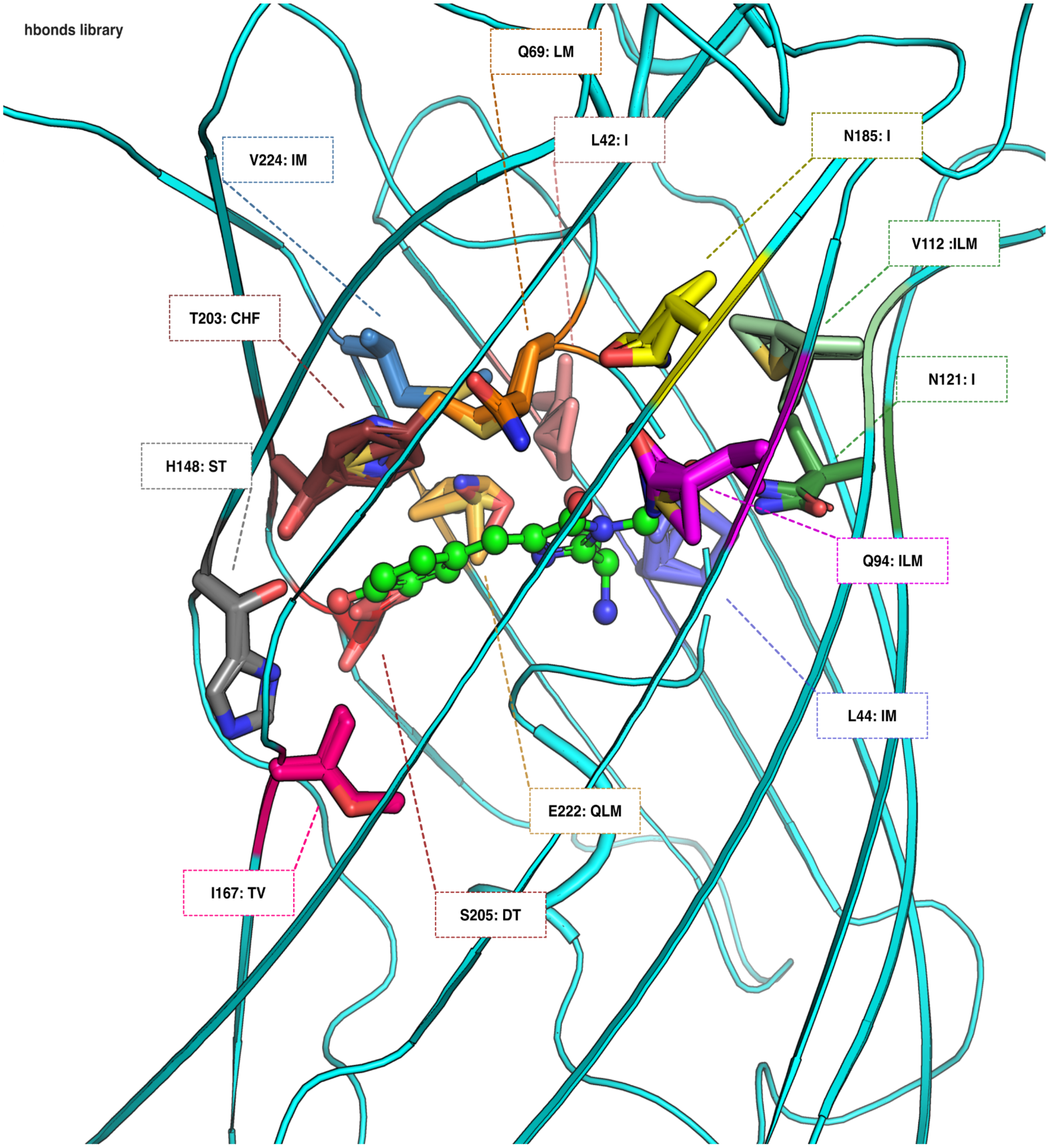
```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="scorefxn_full" weights="ref2015">
            <Reweight scoretype="coordinate_constraint" weight="0.1"/>
        </ScoreFunction>
        <ScoreFunction name="soft_rep_full" weights="soft_rep">
            <Reweight scoretype="coordinate_constraint" weight="0.1"/>
        </ScoreFunction>
    </SCOREFXNS>
    <RESIDUE_SELECTORS>
            <Index name="ress_fix" resnums="%%res_to_fix%%"/>
    </RESIDUE_SELECTORS>
    <TASKOPERATIONS>
        <RestrictToRepacking name="rtr"/>
        <OperateOnResidueSubset name="fix_not_neighbor">
            <Not>
            <Neighborhood distance="8">
                    <Index resnums="%%all_ress%%"/>
            </Neighborhood>
            </Not>
            <PreventRepackingRLT/>
        </OperateOnResidueSubset>
        <InitializeFromCommandline name="init"/>
        <IncludeCurrent name="include_curr"/>
        <OperateOnResidueSubset name="fix_res" selector="ress_fix">
                <PreventRepackingRLT/>
        </OperateOnResidueSubset>
        <OperateOnResidueSubset name="not_to_cst_sc">
            <Not selector="ress_fix"/>
            <PreventRepackingRLT/>
        </OperateOnResidueSubset>
    </TASKOPERATIONS>
    <FILTERS>
        <DesignableResidues name="designable" task_operations="fix_not_neighbor"
designable="0" packable="1"/>
    </FILTERS>
    <MOVERS>
          <MutateResidue name="mutres0" new_res="%%new_res0%%" target="%%target0%%"
preserve_atom_coords="1"/>
        <ConstraintSetMover name="add_CA_cst" cst_file="%%cst_full_path%%"/>
        <AtomCoordinateCstMover name="fix_res_sc_cst" coord_dev="0.5" bounded="false"
sidechain="true" task_operations="not_to_cst_sc"/>
        <PackRotamersMover name="prm"
task_operations="init,include_curr,rtr,fix_not_neighbor,fix_res" scorefxn="scorefxn_full"/>
        <RotamerTrialsMinMover name="rtmin"
task_operations="init,include_curr,rtr,fix_not_neighbor,fix_res" scorefxn="scorefxn_full"/>
        <MinMover name="min" bb="1" chi="1" jump="1" scorefxn="scorefxn_full"/>
        <PackRotamersMover name="soft_repack" scorefxn="soft_rep_full"
task_operations="init,include_curr,rtr,fix_not_neighbor,fix_res"/>
    </MOVERS>
    <PROTOCOLS>
        <Add mover="add_CA_cst"/>
        <Add mover="fix_res_sc_cst"/>
        <Add mover="mutres0"/>
        <Add mover="soft_repack"/>
        <Add mover="min"/>
        <Add mover="prm"/>
        <Add mover="min"/>
        <Add filter="designable"/>
    </PROTOCOLS>
    <OUTPUT scorefxn="scorefxn_full"/>
</ROSETTASCRIPTS>
```



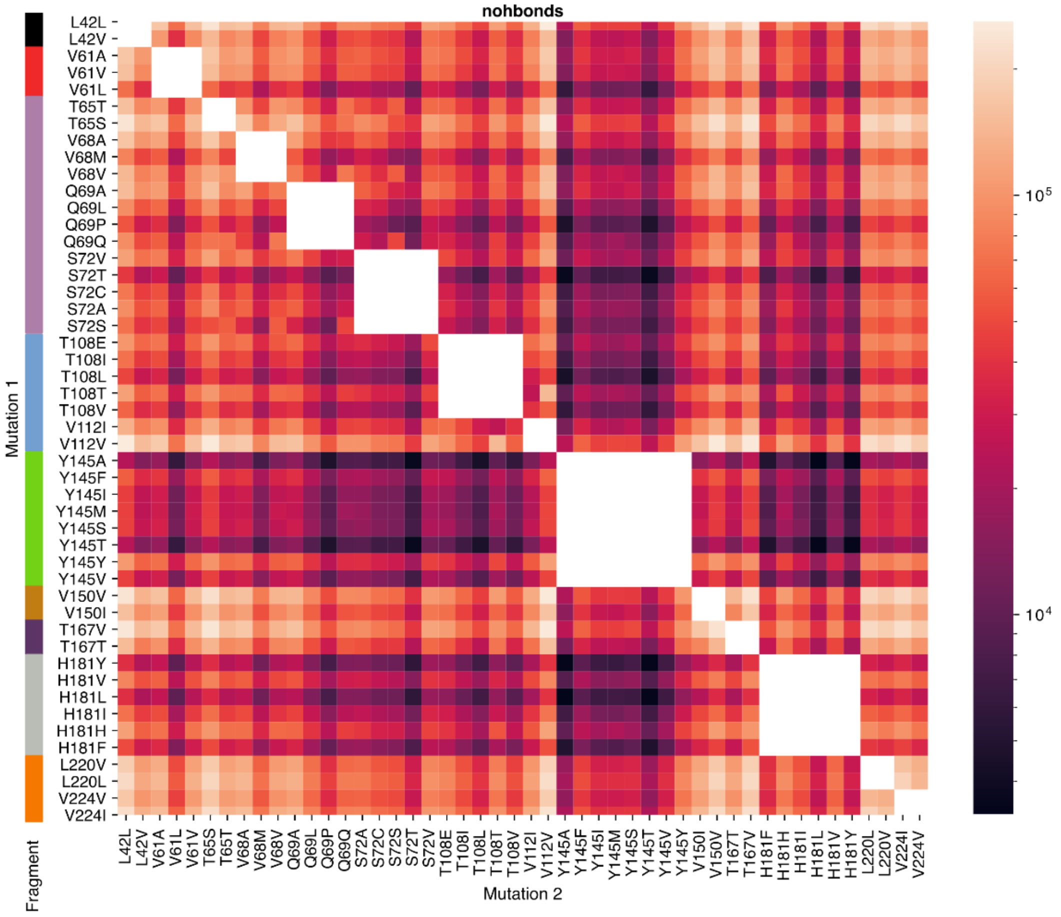
Supplementary Figure 1. Examples of direct and indirect epistasis. (**A**) In direct epistasis, interacting amino acids form favorable contacts (e.g., a hydrogen bond between the Thr and His residues). The double mutant Val/Pro pair is also favorable, but the point mutant Thr→Pro (middle) sterically overlaps with the His. (**B**) In indirect epistasis, a mutation (Thr→Ala) eliminates a hydrogen bond to the backbone (dashed line), leading to a conformational change across several non-interacting amino acids. This example is taken from a comparison of the structures of human and computationally designed variant acetylcholinesterase. Protein Data Bank entries 4EY4 and 5HQ3 are in gray and orange, respectively[56]. Showing positions 110 and 112. (**C**) Schematic explanation of stability-mediated interactions, the wild type (WT), mutant A, and mutant B are stable (below the stability threshold). The energy of the double mutant is a linear sum of the two energies of the two mutants, but it is not expressible as its stability has crossed the stability threshold (marked by a dashed line)[39].

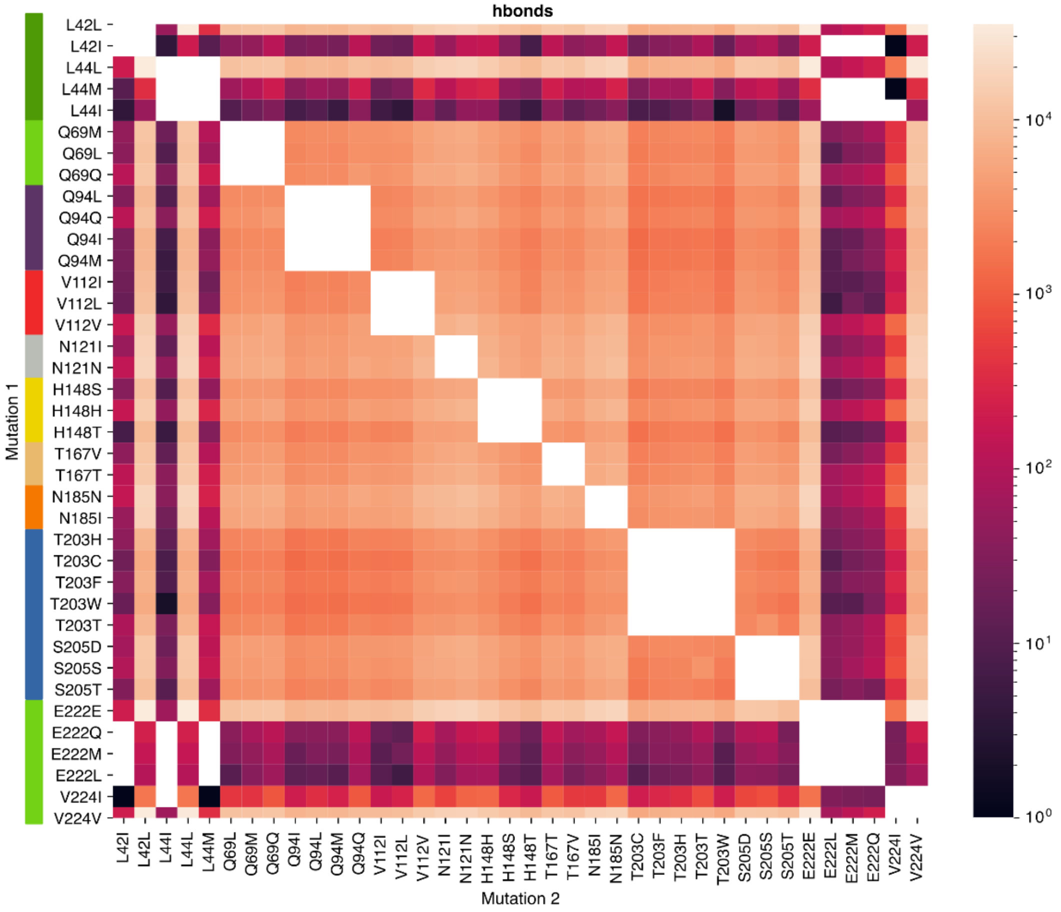


Supplementary Figure 2. Structural overview of the nohbonds library. All mutations are overlayed in stick representation, colored by position. The total library size is 11,059,200 designs.

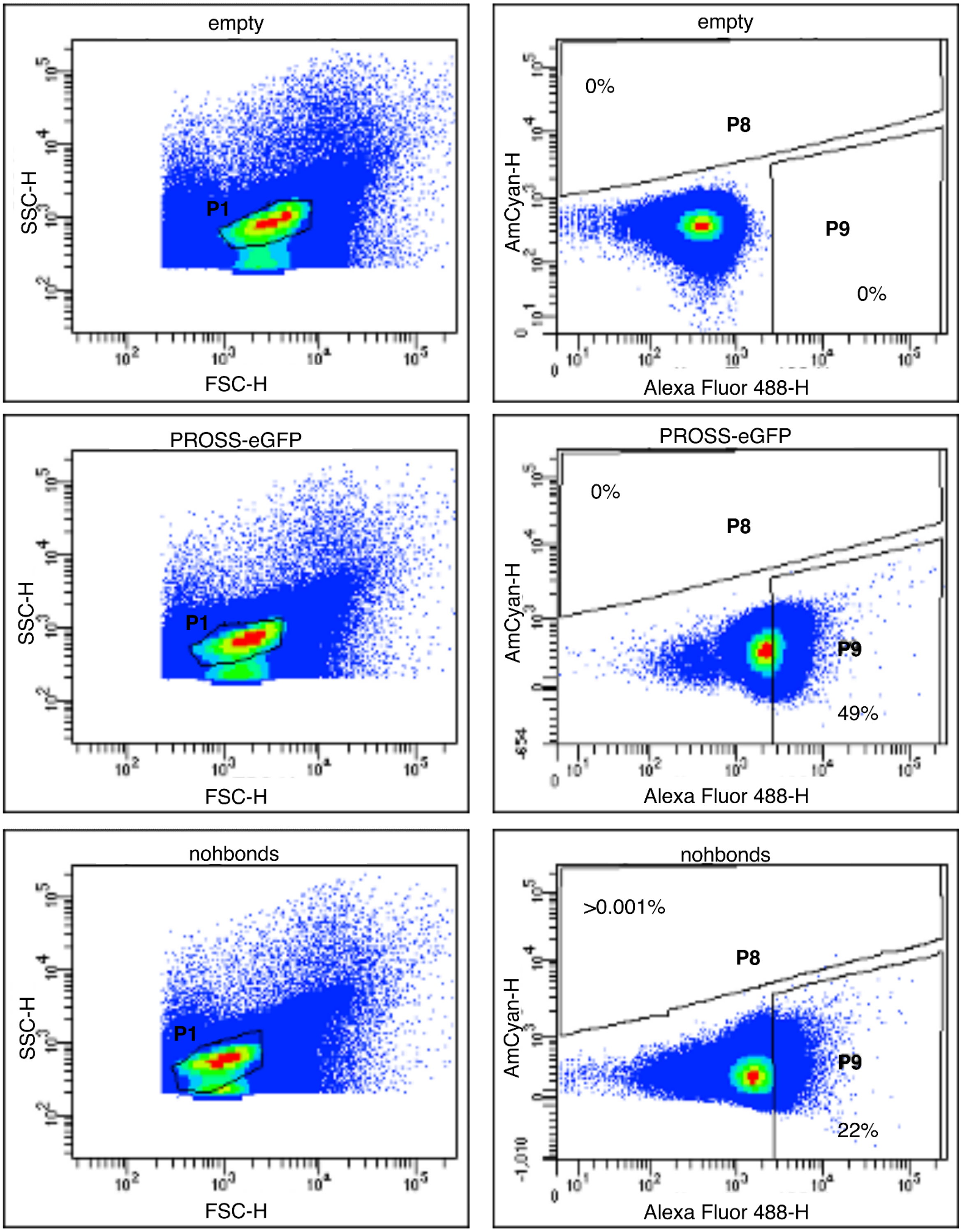


Supplementary Figure 3. Structural overview of the hbonds library. All mutations are overlayed in stick representation, colored by position. The total library size is 933,120 designs.

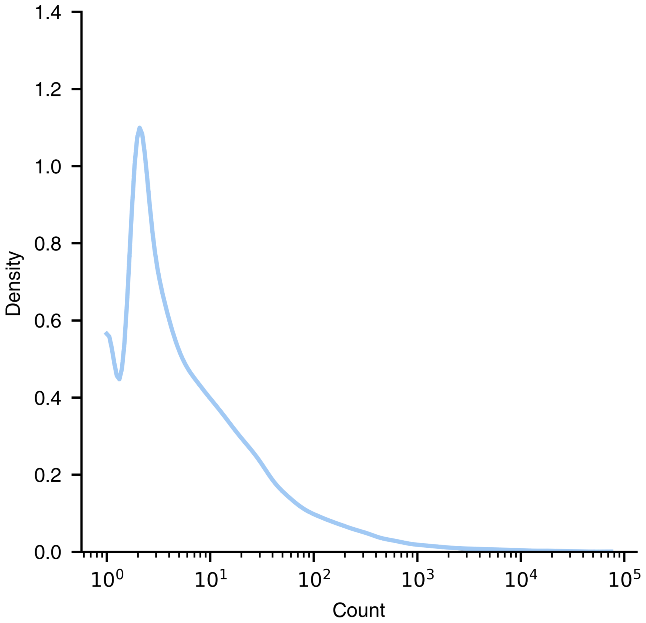

nohbonds
Mutation 1
Fragment
L42L
L42V
V61A
V61V
V61L
T65T
T65S
V68A
V68M
V68V
Q69A
Q69L
Q69P
Q69Q
S72V
S72T
S72C
S72A
S72S
T108E
T108I
T108L
T108T
T108V
V112I
V112V
Y145A
Y145F
Y145I
Y145M
Y145S
Y145T
Y145Y
Y145V
V150V
V150I
T167V
T167T
H181Y
H181V
H181L
H181I
H181H
H181F
L220V
L220L
V224V
V224I
L42L
L42V
V61A
V61L
V61V
T65S
T65T
V68A
V68M
V68V
Q69A
Q69L
Q69P
Q69Q
S72A
S72C
S72S
S72T
S72V
T108E
T108I
T108L
T108T
T108V
V112I
V112V
Y145A
Y145F
Y145I
Y145M
Y145S
Y145T
Y145V
Y145Y
V150I
V150V
T167T
T167V
H181F
H181H
H181I
H181L
H181V
H181Y
L220L
L220V
V224I
V224V
Mutation 2
10^5
10^4


hbonds
Mutation 1
L42L
L42I
L44L
L44M
L44I
Q69M
Q69L
Q69Q
Q94L
Q94Q
Q94I
Q94M
V112I
V112L
V112V
N121I
N121N
H148S
H148H
H148T
T167V
T167T
N185N
N185I
T203H
T203C
T203F
T203W
T203T
S205D
S205S
S205T
E222E
E222Q
E222M
E222L
V224I
V224V
L42I
L42L
L44I
L44L
L44M
Q69L
Q69M
Q69Q
Q94I
Q94L
Q94M
Q94Q
V112I
V112L
V112V
N121I
N121N
H148H
H148S
H148T
T167T
T167V
N185I
N185N
T203C
T203F
T203H
T203T
T203W
S205D
S205S
S205T
E222E
E222L
E222M
E222Q
V224I
V224V
Mutation 2
10^4
10^3
10^2
10^1
10^0

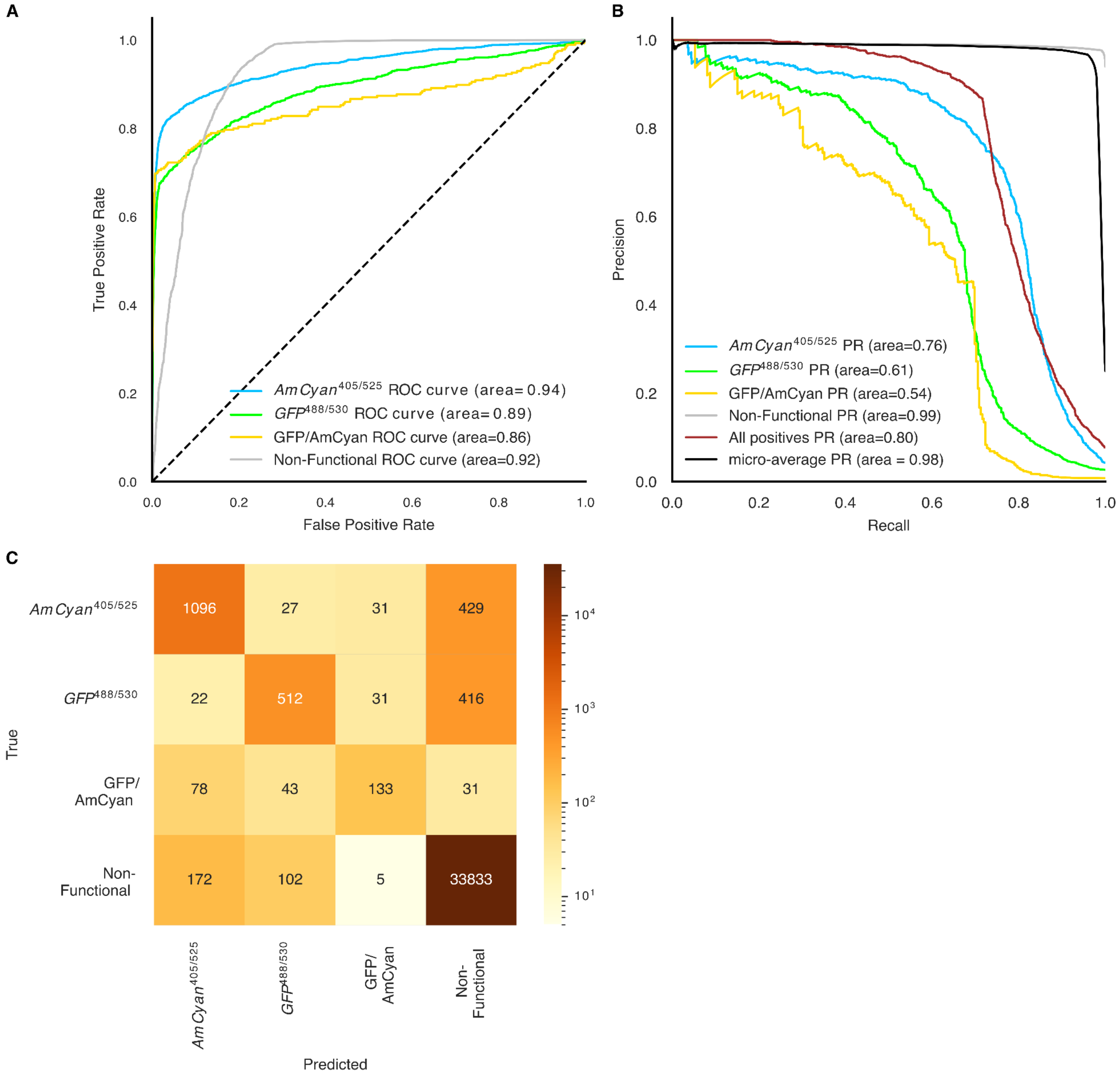
Supplementary figure 4. Golden Gate assembly validation. Both libraries were cloned using Golden-Gate assembly. The number of occurrences of each pair of mutations is shown as a heat map. Sequence positions that were on the same oligonucleotide are marked with a continuous colored bar on the left. There is no obvious linkage between any pair of mutations, which means that mutations are uniformly represented. Additionally, all single and double mutations were present in the nohbonds library. The hbonds library suffered from small diversity at the edges (first and last oligonucleotides), and thus not all pairs of mutations are represented. Mutations in the same position are masked in white as a single sequence cannot have two mutations at the same position.



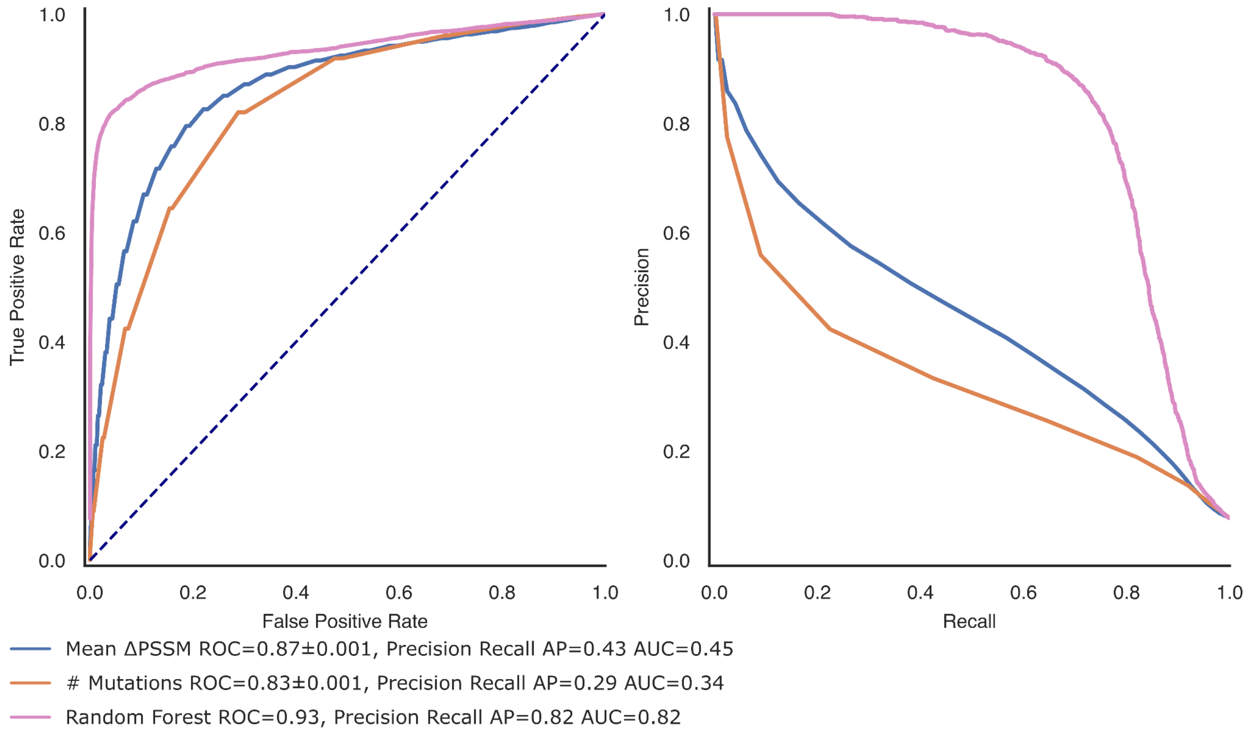
Supplementary Figure 5. Library sorting gates. The hbonds and nohbonds libraries were sorted for excitation at both 405 and 488, with emission at 530 and 525, respectively. Alexa Fluor 488 measures excitation and emission at 488 and 530 nm, respectively. AmCyan measures excitation and emission at 405 and 525 nm, respectively. Each panel shows only 10,000 events.



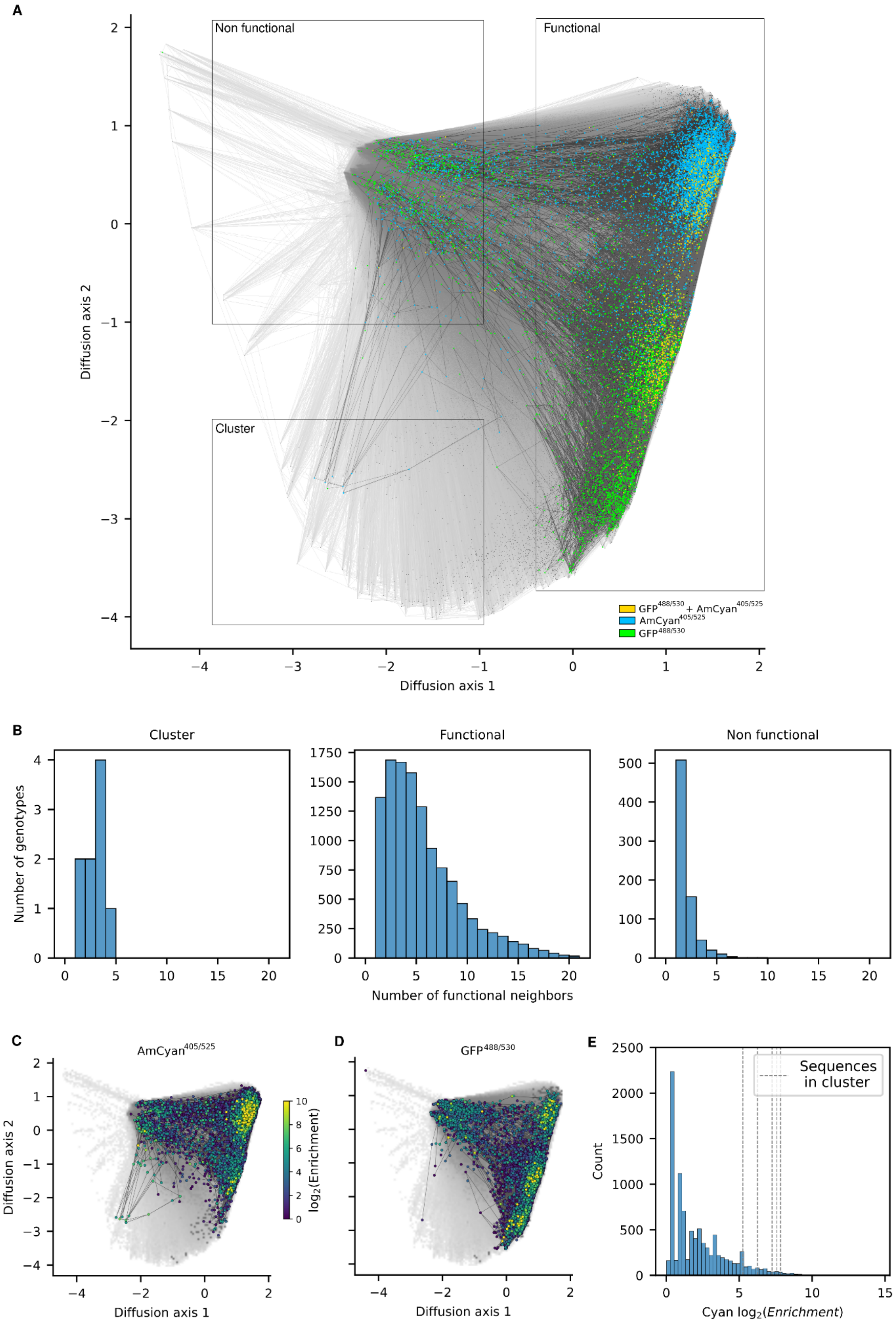
Supplementary Figure 6. deep-sequencing counts across the sorted samples. Number of times each unique sequence in all sorted samples was found in the deep-sequencing data.



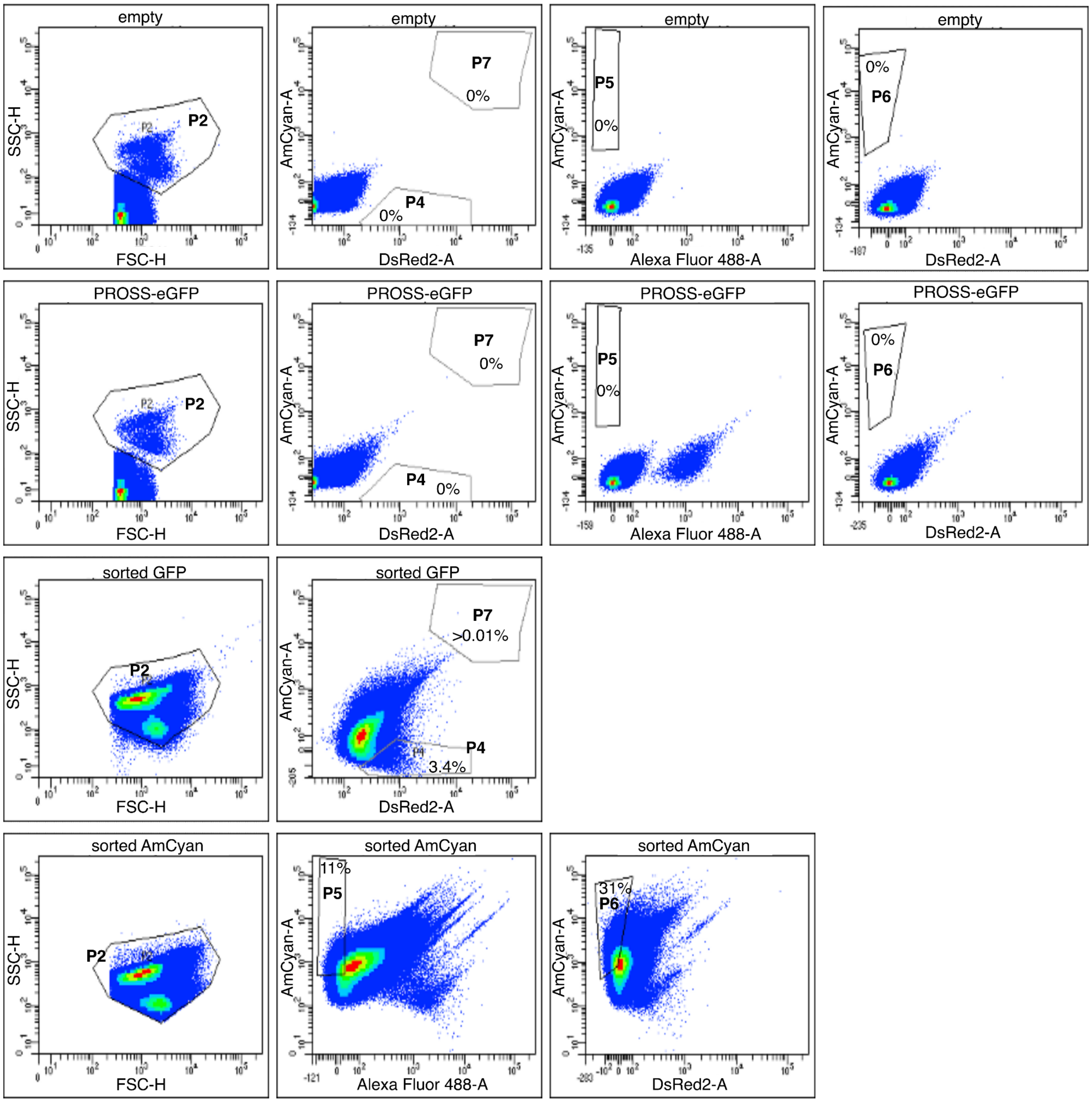
Supplementary Figure 7. Random forest prediction analysis. (A) Receiver Operating Characteristic (ROC) analysis of classification accuracy to all four classes. (B) Precision-Recall analysis for all four classes. "All positives" refers to only the functional classes, and "micro-average" refers to a sliding window that measures the average precision across all classes. The area under the curve (AUC) and average precision (AP) are reported for the ROC and precision-recall analysis, respectively. (C) A confusion matrix of prediction results. All analyses were conducted on an independent test set. The random forest is fairly accurate in determining whether a given sequence is functional, and is somewhat less accurate in assigning a specific functional classification (GFP, AmCyan, or GFP/AmCyan).



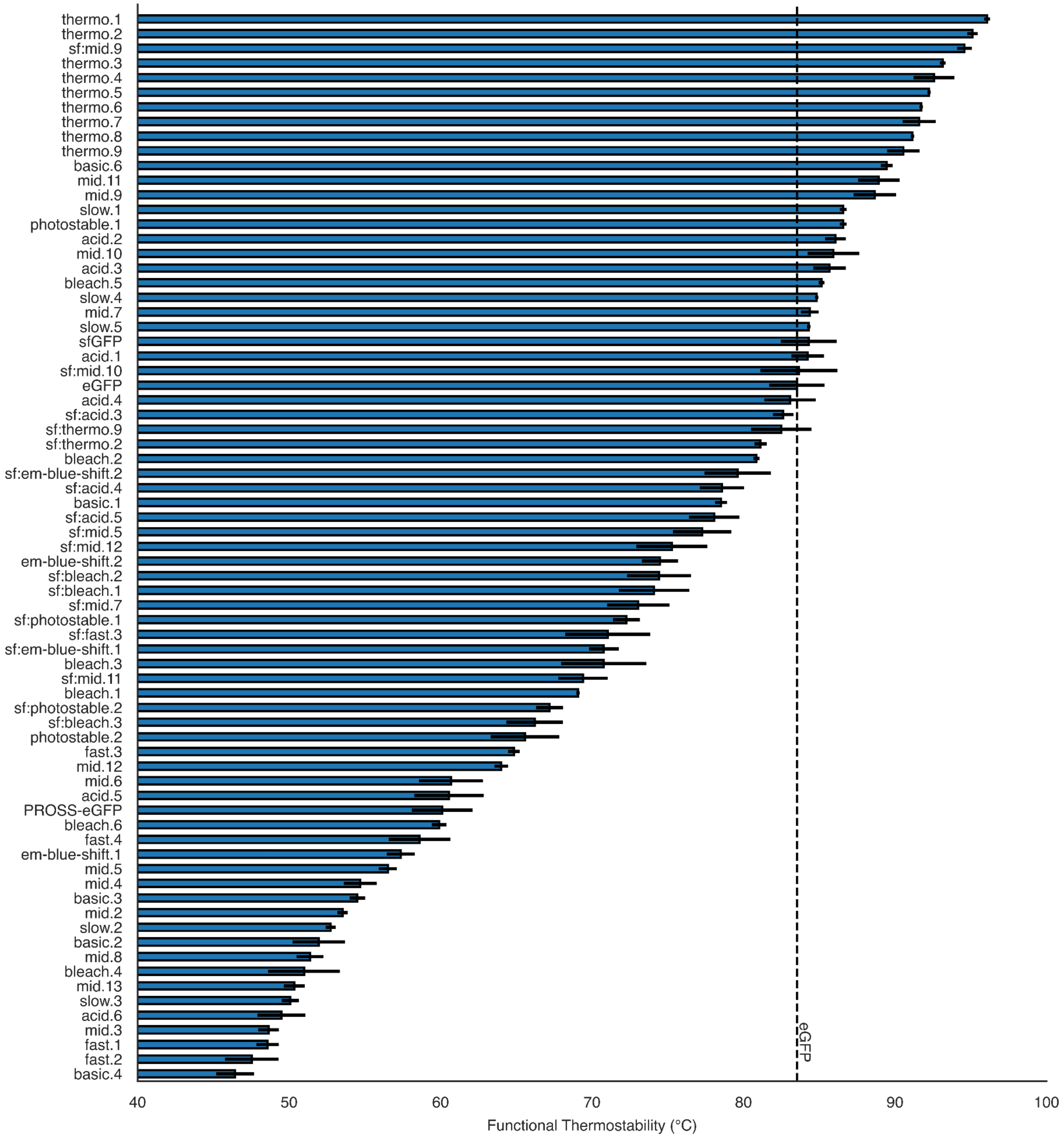
Supplementary Figure 8. Mean ∆PSSM and number of mutations predict design functionality. Prediction accuracy analysis for mean ∆PSSM and number of mutations, receiver operator curve (ROC, left), and Precision-Recall curve (right). The area under the curve (AUC) and average precision (AP) are reported for the ROC and precision-recall analysis, respectively.


A
Non functional
Functional
Cluster
Diffusion axis 2
Diffusion axis 1
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
$GFP^{488/530}$ + $AmCyan^{405/525}$
$AmCyan^{405/525}$
$GFP^{488/530}$
B
Cluster
Functional
Non functional
Number of genotypes
Number of functional neighbors
0
1
2
3
4
250
500
750
1000
1250
1500
1750
100
200
300
400
500
0
5
10
15
20
C
$AmCyan^{405/525}$
D
$GFP^{488/530}$
Diffusion axis 2
Diffusion axis 1
$\log_2$(Enrichment)
0
2
4
6
8
10
E
Sequences in cluster
Count
500
1000
1500
2000
2500
Cyan $\log_2(Enrichment)$
0
5
10
15

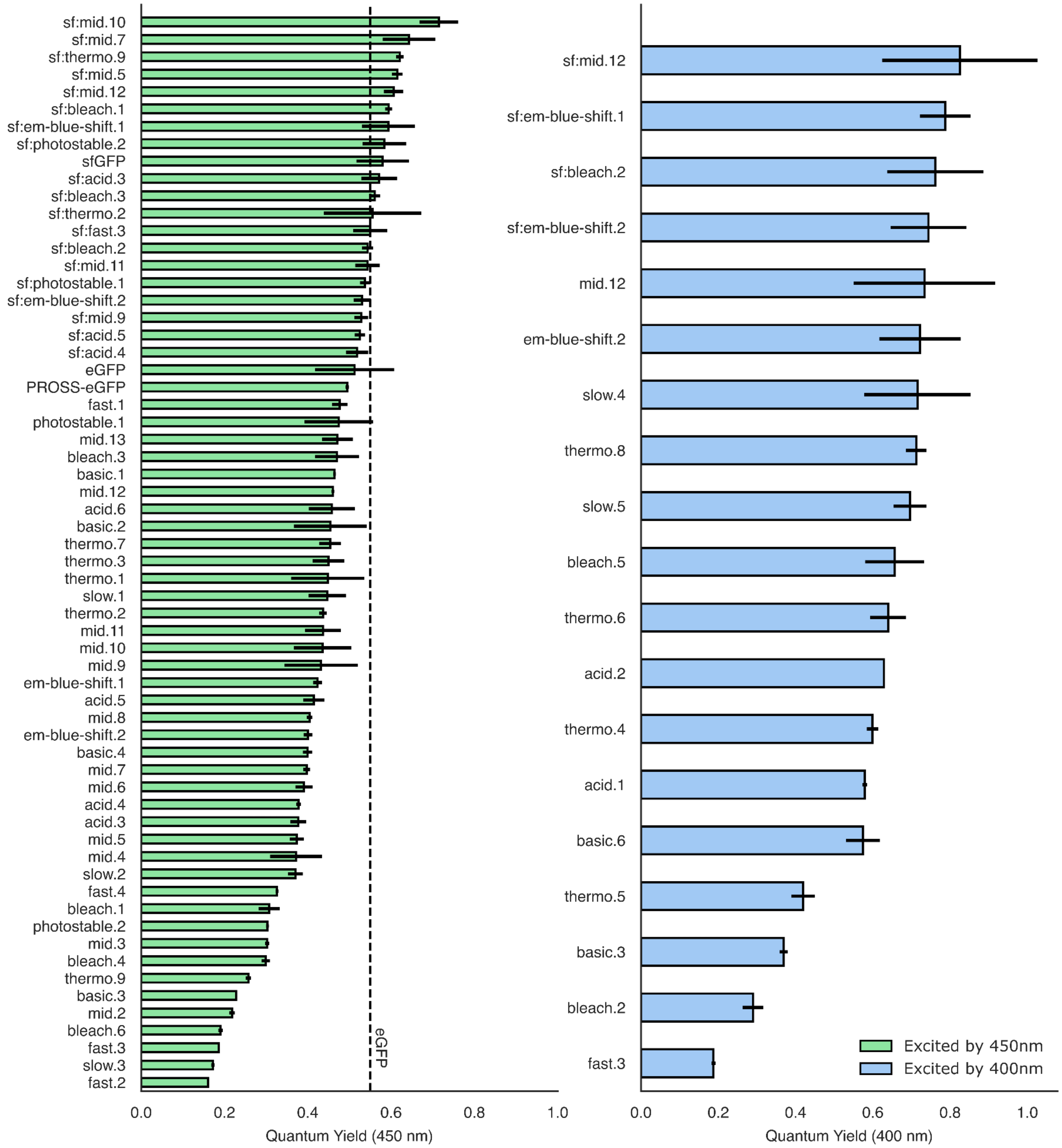
Supplementary Figure 9. Fitness landscape visualizations showing experimentally enriched sequences. (A) Low dimensional visualization of the sequence-function relationship predicted by the random forest model. Sequences are highlighted in different colors according to whether they are detected to be enriched in the GFP (green), AmCyan (blue) or both channels (orange) in the high-throughput data. Dark lines join experimentally enriched genotypes that are separated by single amino acid substitutions. **(B)** Degree distributions for genotypes located at different regions of the visualization as observed directly in the experimental data: in the minor cluster, the main set of functional sequences, and the set of genotypes that the RF predicted to be non-functional. Non-functional genotypes tend to be more poorly connected in the graph of experimentally determined sequences than those in the main set of functional sequences (Mann-Whitney U test, $p\text{-value}<10^{-10}$), further suggesting that, on average, they are false positives correctly smoothed by the RF. The small cluster of functional sequences predicted by the RF shows a higher connectivity than the set of non-functional sequences (Mann-Whitney U test, $p\text{-value} < 0.001$), providing an additional line of evidence for their functionality. **(C, D)** Low dimensional visualization of the sequence-function relationship predicted by the random forest model. Overlaid sequences represent the sequences that were enriched in the experimental data with the color scale from purple to yellow represents is proportional to the $\log_2$(Enrichment) in the sorted vs unsorted fractions for AmCyan (C) and GFP (D). Sequences with higher enrichment values are represented on top and the color scale was truncated at $\log_2(\text{Enrichment})=10$.(E) Histogram of the $\log_2$(Enrichment) in the sorted vs unsorted fractions for AmCyan with vertical lines showing the values for sequences in the Cluster highlighted in (A).



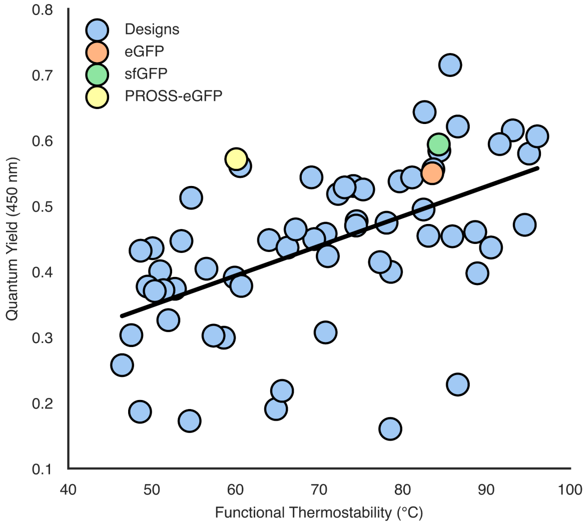
Supplementary Figure 10. Screen for designs with shifted fluorescence spectra. We sorted the GFP and AmCyan pre-sorted libraries for designs that exhibit spectral shifts compared to PROSS-eGFP and eGFP. From top to bottom: empty vector as a negative control, PROSS-eGFP as a positive control, the library sorted for GFP fluorescence and the library sorted for AmCyan fluorescence. Alexa Fluor 488 measures excitation and emission at 488 and 530 nm, respectively. AmCyan measures excitation and emission at 405 and 525 nm, respectively. DsRed2 measures excitation and emission at 561 and 582 nm, respectively.



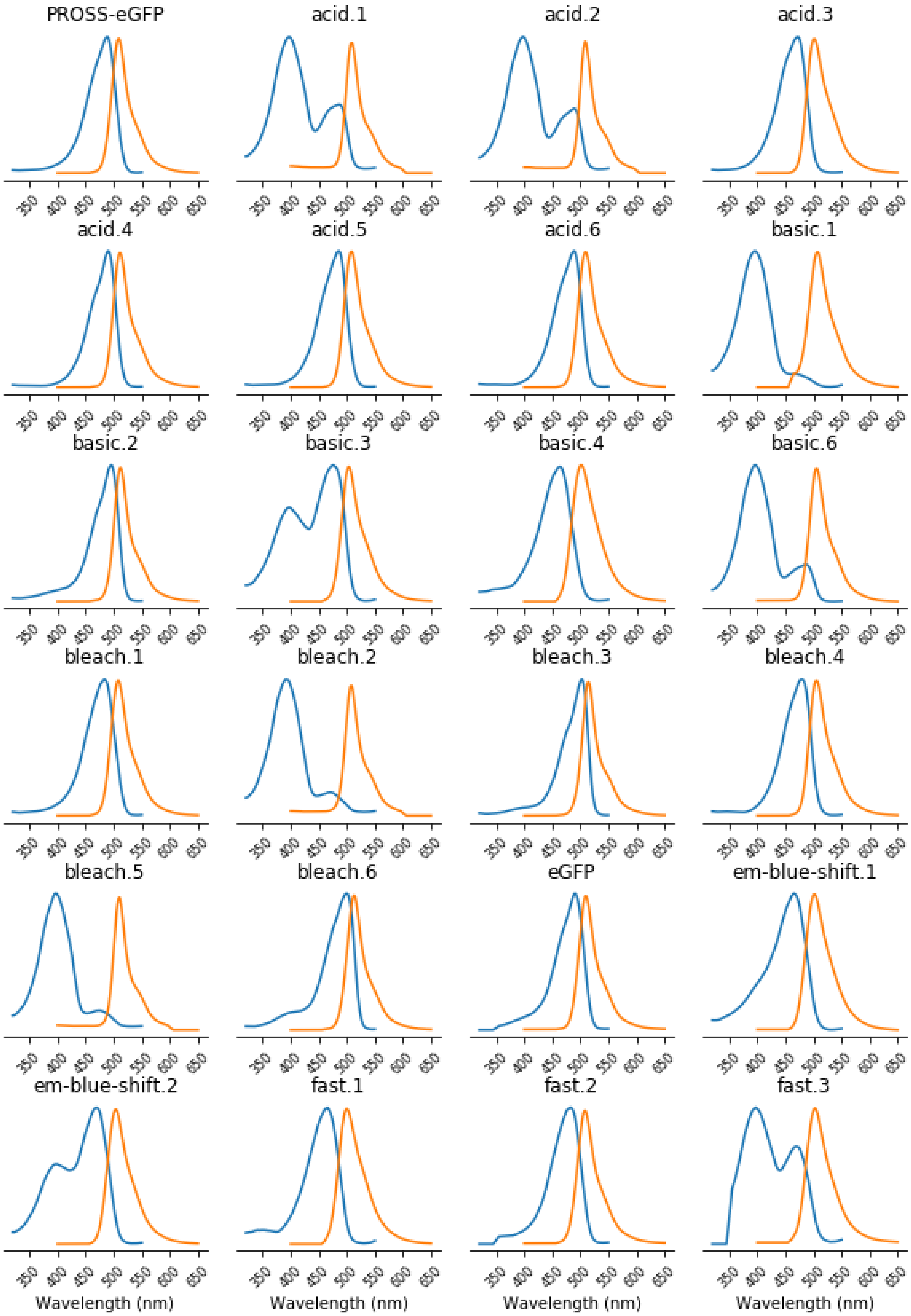
Supplementary Figure 11. Functional thermostability of selected designs. Functional thermostability is the temperature at which fluorescence is at 50% of the max. The dashed line marks the thermostability of eGFP. Error bars mark standard deviations.

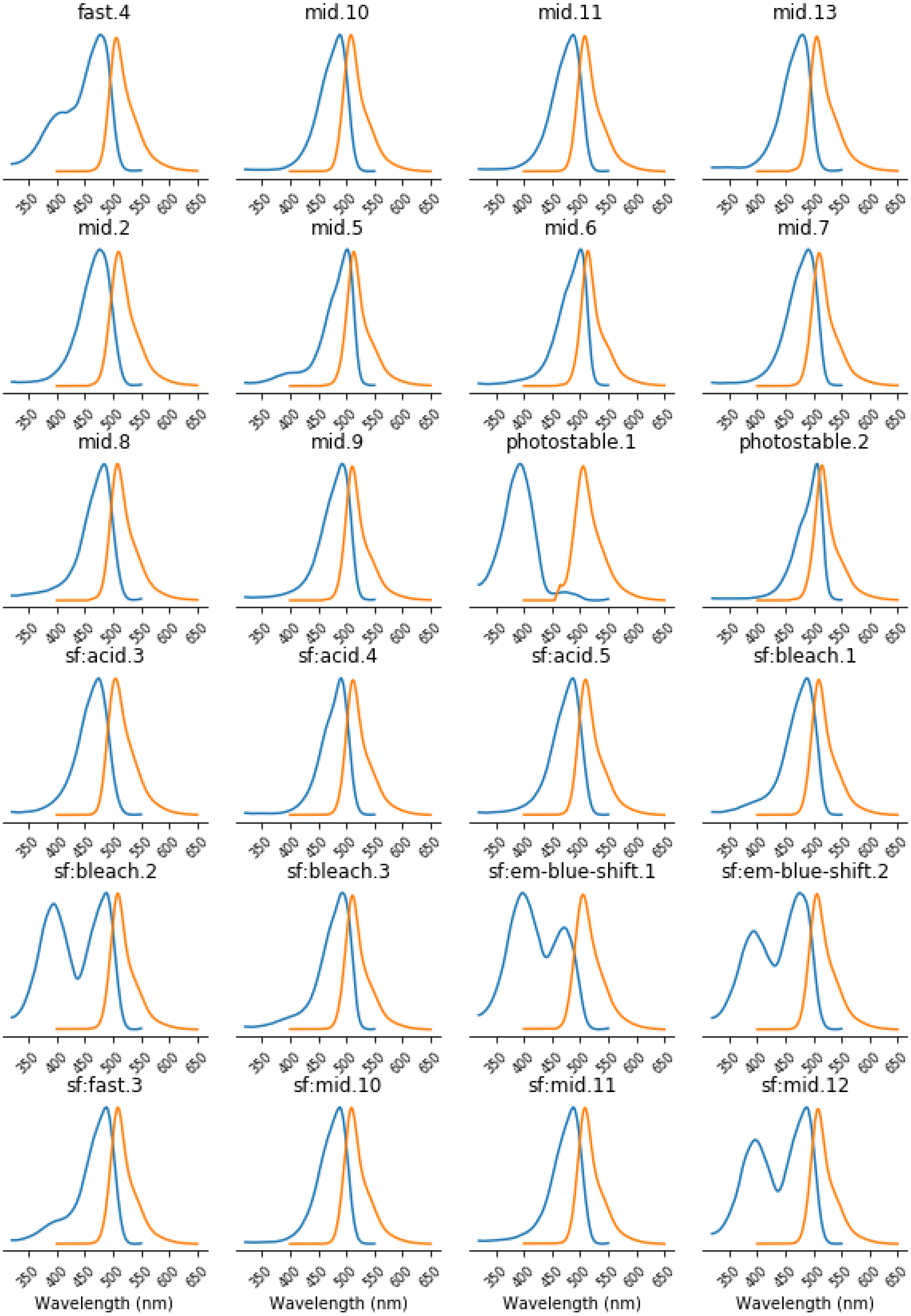


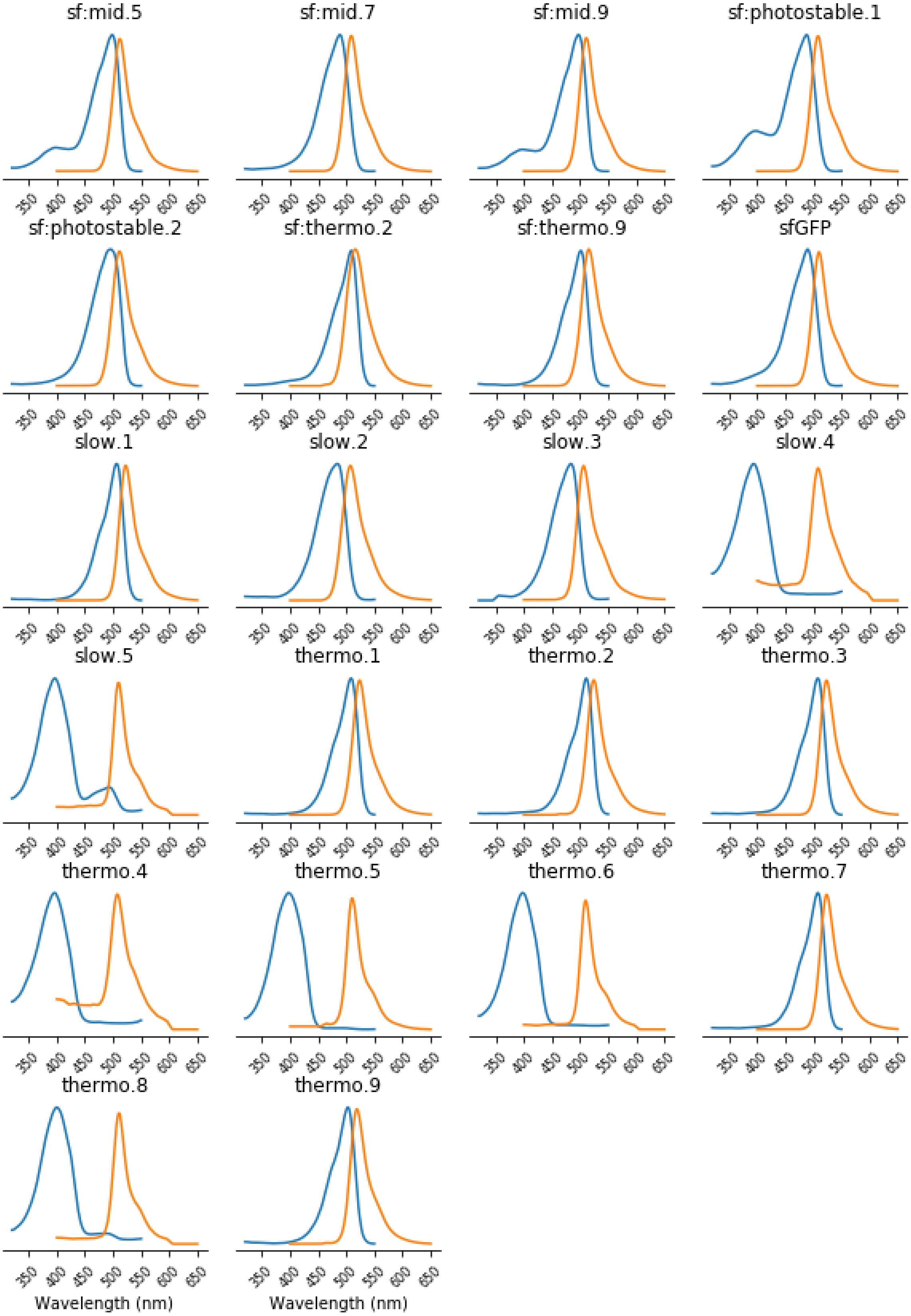
Supplementary Figure 12. Quantum yield measurement for all designs with excitation at either 400 or 450 nm. The dashed line marks the quantum yield of eGFP (which is not excited at 400nm). Error bars mark standard deviations.



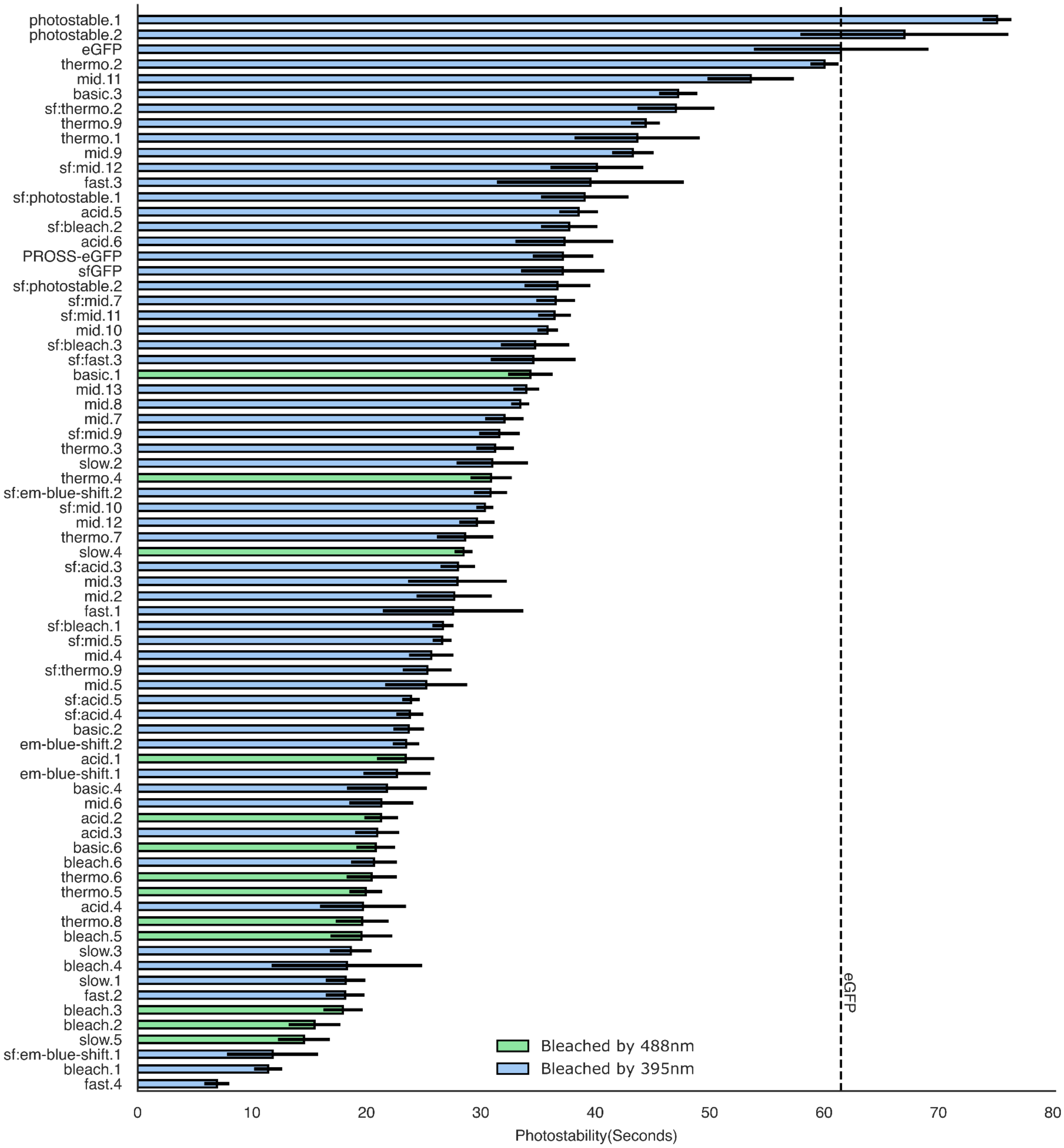
Supplementary Figure 13. Functional thermostability and quantum yield are correlated. Pearson's *r*=0.53 (*p*-value<10^{-5}) between functional thermostability and quantum yield at 450nm excitation.


PROSS-eGFP
acid.1
acid.2
acid.3
acid.4
acid.5
acid.6
basic.1
basic.2
basic.3
basic.4
basic.6
bleach.1
bleach.2
bleach.3
bleach.4
bleach.5
bleach.6
eGFP
em-blue-shift.1
em-blue-shift.2
fast.1
fast.2
fast.3
350
400
450
500
550
600
650
Wavelength (nm)

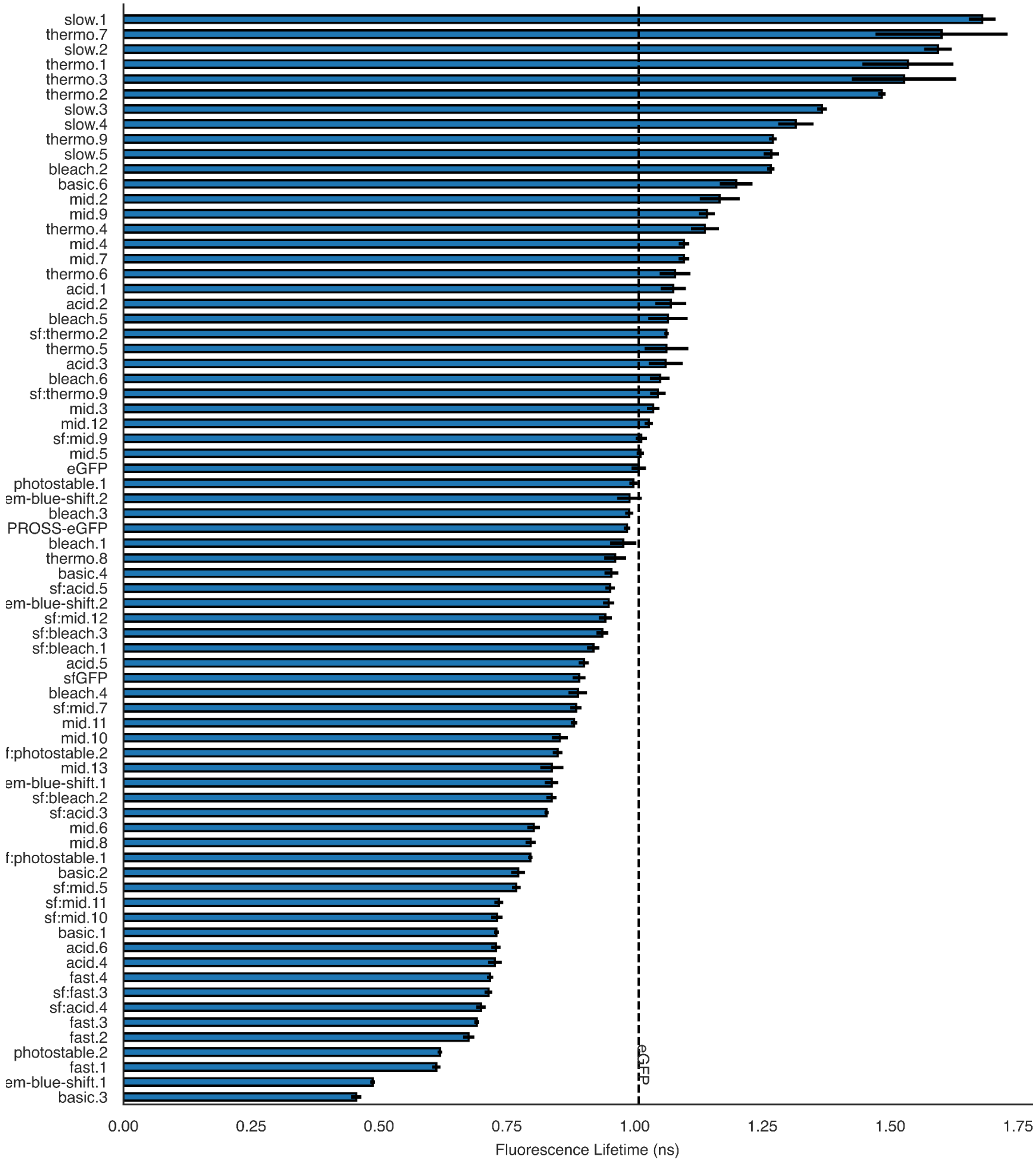

fast.4
mid.10
mid.11
mid.13
mid.2
mid.5
mid.6
mid.7
mid.8
mid.9
photostable.1
photostable.2
sf:acid.3
sf:acid.4
sf:acid.5
sf:bleach.1
sf:bleach.2
sf:bleach.3
sf:em-blue-shift.1
sf:em-blue-shift.2
sf:fast.3
sf:mid.10
sf:mid.11
sf:mid.12
350
400
450
500
550
600
650
Wavelength (nm)
Wavelength (nm)
Wavelength (nm)
Wavelength (nm)


sf:mid.5
sf:mid.7
sf:mid.9
sf:photostable.1
sf:photostable.2
sf:thermo.2
sf:thermo.9
sfGFP
slow.1
slow.2
slow.3
slow.4
slow.5
thermo.1
thermo.2
thermo.3
thermo.4
thermo.5
thermo.6
thermo.7
thermo.8
thermo.9
350
400
450
500
550
600
650
Wavelength (nm)
Wavelength (nm)

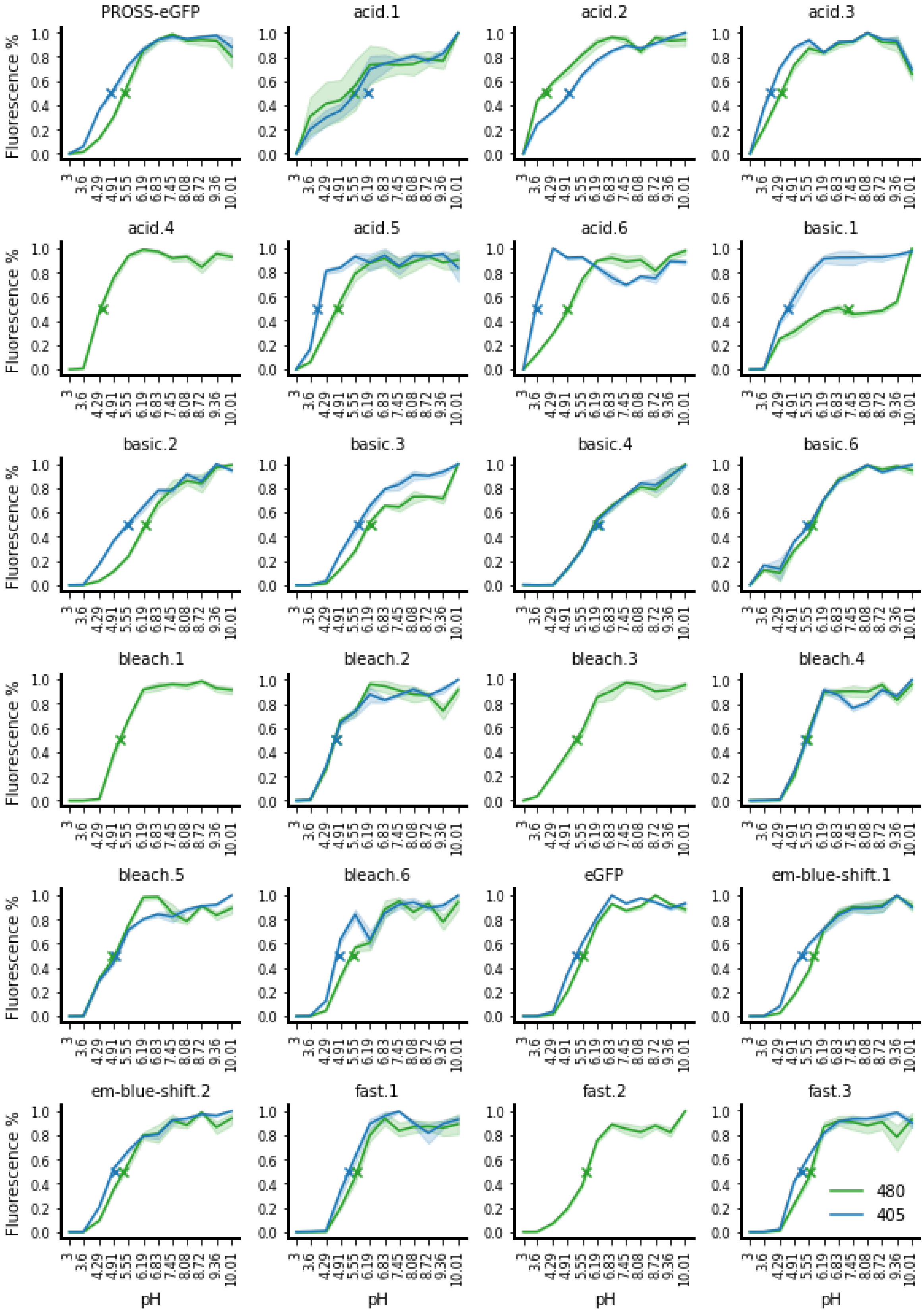
Supplementary Figure 14. Spectral properties of all tested designs. Excitation and emission spectra are shown in blue and orange colors, respectively.

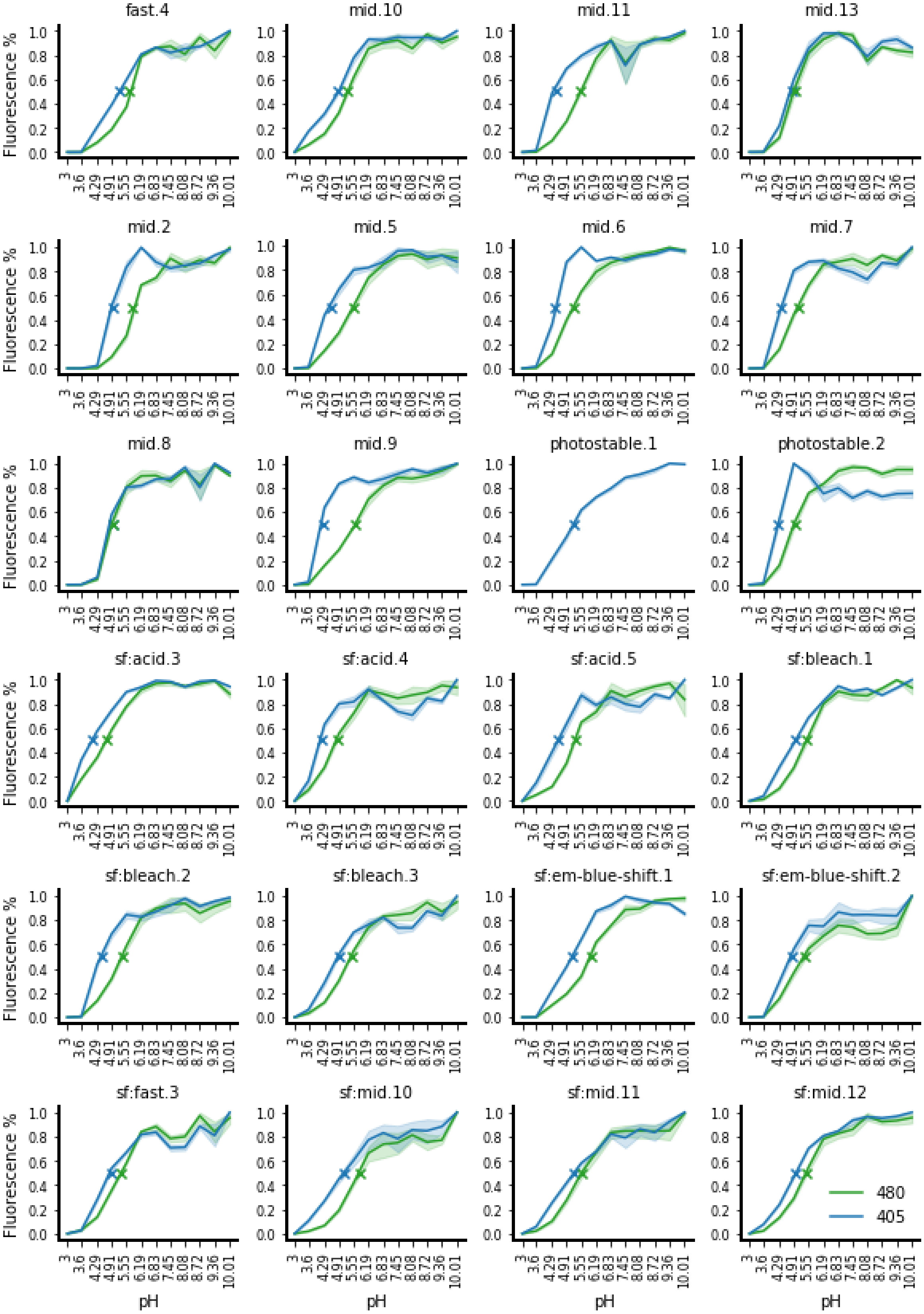


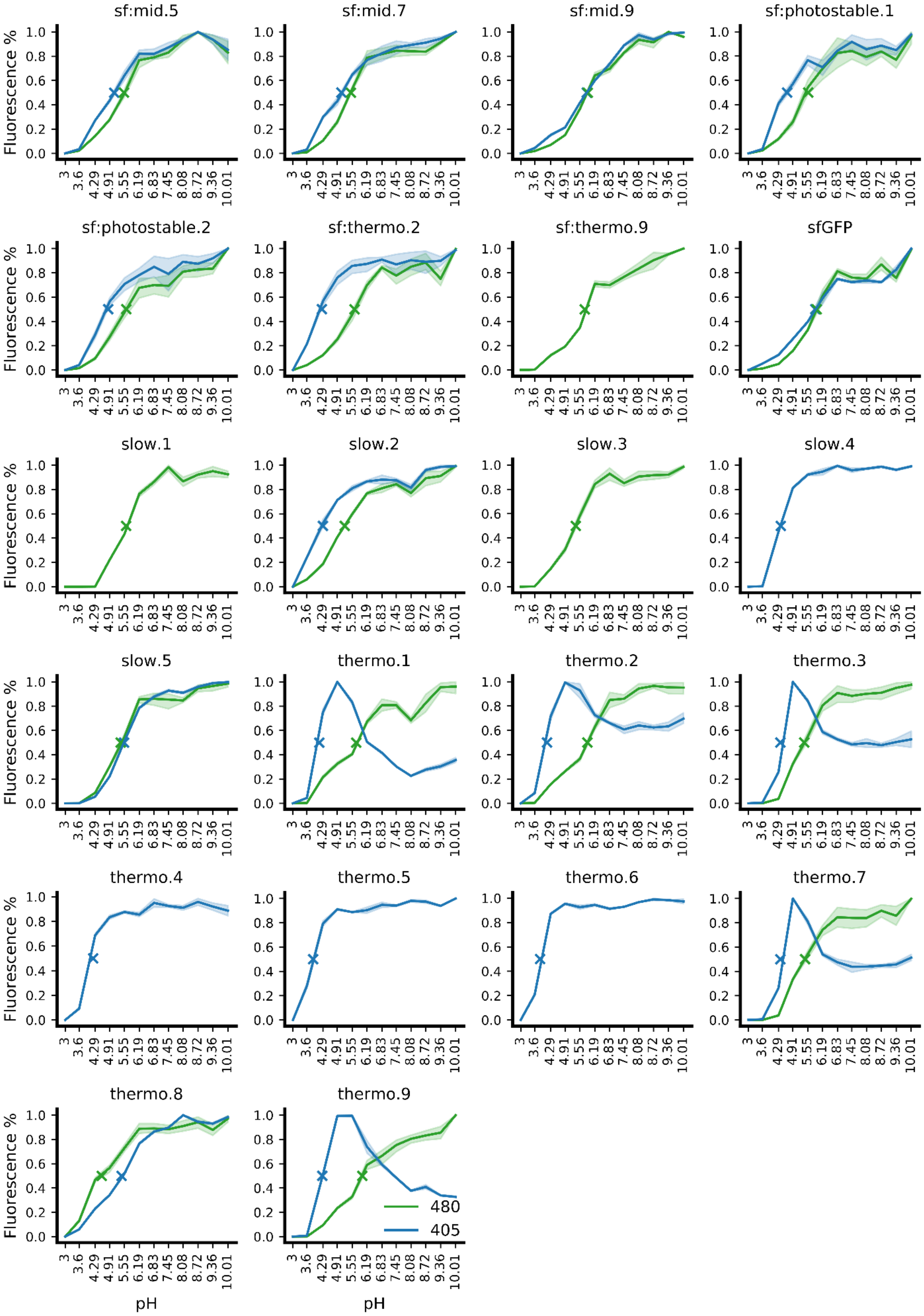
Supplementary Figure 15. Photobleaching measurement of all selected designs. Shows average and standard deviation of at least three independent measurements of photobleaching (Methods). The dashed line marks the eGFP. Error bars mark standard deviations.



Supplementary Figure 16. Fluorescence lifetime measurements of all tested designs. Bars depict the weighted average of bi-exponential fit to lifetime measurements. The dashed line marks the lifetime of eGFP. Error bars mark standard deviations.


PROSS-eGFP
acid.1
acid.2
acid.3
acid.4
acid.5
acid.6
basic.1
basic.2
basic.3
basic.4
basic.6
bleach.1
bleach.2
bleach.3
bleach.4
bleach.5
bleach.6
eGFP
em-blue-shift.1
em-blue-shift.2
fast.1
fast.2
fast.3
Fluorescence %
0.0
0.2
0.4
0.6
0.8
1.0
3
3.6
4.29
4.91
5.55
6.19
6.83
7.45
8.08
8.72
9.36
10.01
pH
480
405


fast.4
mid.10
mid.11
mid.13
mid.2
mid.5
mid.6
mid.7
mid.8
mid.9
photostable.1
photostable.2
sf:acid.3
sf:acid.4
sf:acid.5
sf:bleach.1
sf:bleach.2
sf:bleach.3
sf:em-blue-shift.1
sf:em-blue-shift.2
sf:fast.3
sf:mid.10
sf:mid.11
sf:mid.12
Fluorescence %
0.0
0.2
0.4
0.6
0.8
1.0
3
3.6
4.29
4.91
5.55
6.19
6.83
7.45
8.08
8.72
9.36
10.01
pH
480
405


sf:mid.5
sf:mid.7
sf:mid.9
sf:photostable.1
sf:photostable.2
sf:thermo.2
sf:thermo.9
sfGFP
slow.1
slow.2
slow.3
slow.4
slow.5
thermo.1
thermo.2
thermo.3
thermo.4
thermo.5
thermo.6
thermo.7
thermo.8
thermo.9
Fluorescence %
0.0
0.2
0.4
0.6
0.8
1.0
3
3.6
4.29
4.91
5.55
6.19
6.83
7.45
8.08
8.72
9.36
10.01
pH
480
405

Supplementary Figure 17. pH sensitivity profiles. The normalized fluorescence of each design is shown as a function of buffer pH. Green and blue lines refer to excitation at 480 and 405 nm, respectively. The pKa is the pH at which fluorescence is at 50% of the maximum, annotated by an "X".

Neighborhood center	Neighborhood members	Number of combinations tested	Total number of combinations
14	14: IV 16: VILMT 42: LIMV 44: LIMV 68: VACILMQST 119: LFMVY	1080	7200
16	14: IV 16: VILMT 18: LFI 44: LIMV 46: FY 64: L 68: VACILMQST 119: LFMVY 123: IFLMTV	9720	64800
18	16: VILMT 18: LFI 46: FY 64: L 108: TACEFILMPQSVY 123: IFLMTV	2340	2340
42	14: IV 42: LIMV 44: LIMV 68: VACILMQST 220: LITV 224: VIMSTY	1036	6912
44	14: IV 16: VILMT 42: LIMV 44: LIMV 46: FY 64: L 68: VACILMQST 220: LITV	1728	11520
46	16: VILMT 44: LIMV 46: FY 64: L 68: VACILMQST 123: IFLMTV 220: LITV	1296	8640
61	46: FY 61: VALMQT 64: L 66: YAFHLMNQSW 145: YACEFHIMQSTV 220: LITV	5760	5760
64	16: VILMT 18: LFI 44: LIMV 46: FY 61: VALMQT 64: L 123: IFLMTV 220: LITV	2592	17280
66	66: YAFHLMNQSW 145: YACEFHIMQSTV 150: VILMT 165: FVY 167: TACSV 220: LITV	5400	36000
68	14: IV 42: LIMV 44: LIMV 68: VACILMQST 112: VCILT 119: LFMVY 224: VIMSTY	6480	43200
69	69: QADEHILMNPSTVWY 72: SACFGINTV 150: VILMT 163: VP 165: FVY 185: NADMV 201: LFMQ	12150	81000
72	42: LIMV 72: SACFGINTV 201: LFMQ 224: VIMSTY	864	864
108	18: LFI 108: TACEFILMPQSVY 110: ACHLMY 123: IFLMTV	1404	1404
110	108: TACEFILMPQSVY 110: ACHLMY 112: VCILT 123: IFLMTV	2340	2340

112	68: VACILMQST 110: ACHLMY 112: VCILT 119: LFMVY 185: NADMV	1012	6750
119	14: IV 16: VILMT 68: VACILMQST 112: VCILT 119: LFMVY	2250	2250
123	16: VILMT 18: LFI 64: L 108: TACEFILMPQSVY 110: ACHLMY 112: VCILT 123: IFLMTV	5265	35100
145	61: VALMQT 66: YAFHLMNQSW 145: YACEFHIMQSTV 167: TACSV 181: HEFILTVY 220: LITV	17280	115200
150	66: YAFHLMNQSW 69: QADEHILMNPSTVWY 150: VILMT 163: VP 165: FVY 167: TACSV 201: LFMQ 224: VIMSTY	81000	540000
163	69: QADEHILMNPSTVWY 150: VILMT 163: VP 165: FVY 181: HEFILTVY 185: NADMV 201: LFMQ	10800	72000
165	69: QADEHILMNPSTVWY 150: VILMT 163: VP 165: FVY 167: TACSV 181: HEFILTVY	2700	18000
167	66: YAFHLMNQSW 145: YACEFHIMQSTV 165: FVY 167: TACSV 181: HEFILTVY	2160	14400
181	145: YACEFHIMQSTV 150: VILMT 165: FVY 167: TACSV 181: HEFILTVY	1080	7200
185	69: QADEHILMNPSTVWY 112: VCILT 163: VP 185: NADMV	750	750
201	69: QADEHILMNPSTVWY 72: SACFGINTV 150: VILMT 163: VP 201: LFMQ 224: VIMSTY	4860	32400
220	42: LIMV 44: LIMV 46: FY 61: VALMQT 64: L 66: YAFHLMNQSW 68: VACILMQST 145: YACEFHIMQSTV 220: LITV	124416	829440
224	42: LIMV 68: VACILMQST 69: QADEHILMNPSTVWY 72: SACFGINTV 201: LFMQ 224: VIMSTY	17496	116640

Supplementary Table 1. nohbonds library neighborhoods.

Neighborhood center	Neighborhood members	Number of combinations tested	Total number of combinations
42	145: FHMY 150: VCIMQ 167: ITV 205: SDENQT 222: EILMQ 224: VILMQ	1350	9000
44	42: LFIMQ 220: LMQ 222: EILMQ	75	75
61	42: LFIMQ 44: LEIMQ 61: VIMY 145: FHMY 205: SDENQT 222: EILMQ	1800	12000
62	42: LFIMQ 44: LEIMQ 69: QCILM 203: TCFHNS 205: SDENQT 220: LMQ 224: VILMQ	10125	67500
69	42: LFIMQ 69: QCILM 203: TCFHNS 205: SDENQT 222: EILMQ	4500	4500
92	42: LFIMQ 92: YHN 94: QILM 96: R 121: NI 150: VCIMQ 163: V 165: FY 183: QE 185: NDI 203: TCFHNS 222: EILMQ	32400	216000
94	44: LEIMQ 69: QCILM 220: LMQ 222: EILMQ 224: VILMQ	1875	1875
96	61: VIMY 145: FHMY 148: HCENST 203: TCFHNS 220: LMQ 222: EILMQ 224: VILMQ	6480	43200
112	61: VIMY 62: TACS 145: FHMY 148: HCENST 150: VCIMQ 165: FY 181: HN 203: TCFHNS	6912	46080
121	61: VIMY 62: TACS 148: HCENST 165: FY 167: ITV 181: HN 203: TCFHNS 205: SDENQT 220: LMQ	18662	124416
145	61: VIMY 69: QCILM 96: R 145: FHMY 148: HCENST 165: FY 167: ITV 181: HN	5760	5760
148	62: TACS 145: FHMY 150: VCIMQ 165: FY 167: ITV 203: TCFHNS 205: SDENQT	2592	17280
150	62: TACS 145: FHMY 167: ITV 205: SDENQT 220: LMQ 222: EILMQ	4320	4320
163	62: TACS 69: QCILM 148: HCENST 163: V	2880	2880

	165: FY 183: QE 203: TCFHNS		
165	62: TACS 69: QCILM 94: QILM 163: V 165: FY 181: HN 183: QE	640	640
167	62: TACS 69: QCILM 96: R 148: HCENST 150: VCIMQ 163: V 167: ITV 181: HN 183: QE	1080	7200
181	62: TACS 96: R 163: V 165: FY 167: ITV 183: QE	48	48
183	69: QCILM 92: YHN 94: QILM 96: R 112: VILMNQ 163: V 183: QE	720	720
185	69: QCILM 92: YHN 94: QILM 96: R 150: VCIMQ 163: V 165: FY 181: HN 185: NDI	3600	3600
203	69: QCILM 92: YHN 96: R 112: VILMNQ 121: NI 183: QE 185: NDI	1080	1080
205	69: QCILM 94: QILM 112: VILMNQ 121: NI 183: QE 185: NDI	1440	1440
220	69: QCILM 94: QILM 96: R 150: VCIMQ 165: FY 181: HN 183: QE 185: NDI	2400	2400
222	92: YHN 94: QILM 112: VILMNQ	72	72
224	92: YHN 94: QILM 121: NI 185: NDI	72	72

Supplementary Table 2. hbonds library neighborhoods.

Position	Amino acids after filtering	# amino acids after filtering	Amino acids in library	# amino acids in EpiNNet enriched library[a]
14	IV	2		1
16	VILMT	5	VI	2
18	LFI	3		1
42	LIMV	4	LV	2
44	LIMV	4		1
46	FY	2		1

61	VALMQT	6	VAL	3
64	L	1		1
65[b]	TS	2	TS	
66	YAFHLMNQSW	10		1
68	VACILMQST	9	VMA	3
69	QADEHILMNPSTVWY	15	QPLA	4
72	SACFGINTV	9	SCTAV	5
108	TACEFILMPQSVY	13	TLEIV	5
110	ACHLMY	6		1
112	VCILT	5	VI	2
119	LFMVY	5		1
123	IFLMTV	6		1
145	YACEFHIMQSTV	12	YMSTAIFV	8
150	VILMT	5	VI	2
163	VP	2		1
165	FVY	3		1
167	TACSV	5	TV	2
181	HEFILTVY	8	HYLIFV	6
185	NADMV	5		1
201	LFMQ	4		1
220	LITV	4	LV	2
224	VIMSTY	6	VI	2
	Total number of all combinations	1x10^{19}		11,059,200

Supplementary Table 3. Amino acids after filtration and EpiNNet enrichment for the nohbonds library.

[a] each position has, at a minimum, a single amino acid, the PROSS-eGFP identity at that position.

[b] position 65 is part of the chromophore and was mutated to Ser without being modeled in the Rosetta calculations.

Position	Amino acids after filtering	# amino acids after filtering	Amino acids in library	# amino acids in EpiNNet enriched library[a]
42	LFIMQ	5	LI	2
44	LEIMQ	5	LIM	3
61	VIMY	4		1
62	TACS	4		1
65[b]	TS	2	TS	2
69	QCILM	5	QLM	3
92	YHN	3		1
94	QILM	4	QILM	4
96	R	1		1
112	VILMNQ	6	VIL	3
121	NI	2	NI	2
145	FHMY	4		1
148	HCENST	6	HST	3
150	VCIMQ	5		1
163	V	1		1
165	FY	2		1
167	ITV	3	ITV	3
181	HN	2		1
183	QE	2		1
185	NDI	3	NI	2
203	TCFHNS	6	THFCW	5
205	SDENQT	6	SDT	3
220	LMQ	3		1
222	EILMQ	5	ELQM	4
224	VILMQ	5	VI	2
Total number of all		$6x10^{12}$		933120

combinations

Supplementary Table 4. Amino acids after filtration and EpiNNet enrichment for the hbonds library.

[a] each position has, at a minimum, a single amino acid, the PROSS-eGFP identity at that position.

[b] position 65 is part of the chromophore and was mutated to Ser without being modeled in the Rosetta calculations.

Position	PROSS-eGFP amino acid	nohbonds	hbonds
16	V	VI	
42	L	LV	LI
44	L		LIM
61	V	VAL	
65	T	TS	TS
68	V	VMA	
69	Q	QPLA	QLM
72	S	SCTAV	
94	Q		QILM
108	T	TLEIV	
112	V	VI	VIL
121	N		NI
145	Y	YMSTAIFV	
148	H		HST
150	V	VI	
167	T	TV	TIV
181	H	HYLIFV	
185	N		NI
203	T		THFCW

205	S		SDT
220	L	LV	
222	E		ELQM
224	V	VI	VI

Supplementary Table 5. Sequence spaces of the two libraries.

		Experimental	
		Functional	Non-functional
Deep-sequencing analysis	Functional	45	1
	Non-functional	5	11

Supplementary Table 6. Predictive values for the deep-sequencing data analysis. 62 designs were individually selected directly from FACS sorts and tested for fluorescence. These were used to calibrate the thresholds for the deep-sequencing analysis.

		nohbonds			hbonds		
Position	Amino acid	Total	Functional	Functional Frequency	Total	Functional	Functional Frequenc y
L42	L	367063	11636	3%	20156	1891	9%
L42	V	226468	2606	1%			
L42	I				179	22	12%
L44	L	593531	14242	2%	19959	1883	9%
L44	M				318	26	8%
L44	I				58	4	7%
V61	A	261368	3422	1%			
V61	L	104551	1694	2%			
V61	V	227612	9126	4%	20335	1913	9%
T65	S	363217	8536	2%			
T65	T	230314	5706	2%	20335	1913	9%
V68	A	258462	2053	1%			
V68	M	127481	1544	1%			
V68	V	207588	10645	5%	20335	1913	9%
Q69	A	239915	1806	1%			
Q69	L	143256	1948	1%	6396	532	8%
Q69	P	82078	1013	1%			
Q69	Q	128282	9475	7%	7243	767	11%
Q69	M				6696	614	9%
S72	A	137573	5736	4%			
S72	S	101580	3080	3%	20335	1913	9%
S72	V	176068	1540	1%			
S72	T	62959	1487	2%			
S72	C	115351	2399	2%			

Q94	Q	593531	14242	2%	5911	730	12%
Q94	L				4996	438	9%
Q94	I				4613	367	8%
Q94	M				4815	378	8%
T108	T	148714	8769	6%	20335	1913	9%
T108	E	161665	795	0%			
T108	I	111234	1406	1%			
T108	L	74790	874	1%			
T108	V	97128	2398	2%			
V112	V	382532	12106	3%	8277	903	11%
V112	I	210999	2136	1%	5878	510	9%
V112	L				6180	500	8%
N121	N	593531	14242	2%	10593	1094	10%
N121	I				9742	819	8%
Y145	I	67163	1831	3%			
Y145	M	65798	5555	8%			
Y145	S	71924	770	1%			
Y145	Y	150237	2586	2%	20335	1913	9%
Y145	V	70897	1168	2%			
Y145	F	92946	1217	1%			
Y145	A	37083	237	1%			
Y145	T	37483	878	2%			
H148	H	593531	14242	2%	8131	869	11%
H148	S				6537	567	9%
H148	T				5667	477	8%
V150	V	345895	11767	3%	20335	1913	9%
V150	I	247636	2475	1%			
T167	V	384448	6782	2%	6740	645	10%
T167	T	209083	7460	4%	7591	735	10%

T167	I				6004	533	9%
H181	H	151700	9203	6%	20335	1913	9%
H181	V	133739	1750	1%			
H181	Y	59555	726	1%			
H181	L	57981	1424	2%			
H181	F	79104	365	0%			
H181	I	111452	774	1%			
N185	N	593531	14242	2%	10689	1132	11%
N185	I				9646	781	8%
T203	T	593531	14242	2%	4605	482	10%
T203	H				4200	569	14%
T203	W				3690	286	8%
T203	F				4149	294	7%
T203	C				3691	282	8%
S205	S	593531	14242	2%	7082	670	9%
S205	T				6099	600	10%
S205	D				7154	643	9%
L220	L	315182	8802	3%	20335	1913	9%
L220	V	278349	5440	2%			
E222	E	593531	14242	2%	19861	1891	10%
E222	Q				216	12	6%
E222	L				105	8	8%
E222	M				153	2	1%
V224	V	259722	4877	2%	18875	1795	10%
V224	I	333809	9365	3%	1460	118	8%

Supplementary Table 7. Mutations occurring functional sequences from both libraries.

Dataset	A	B	RMSE
RF	0.0253	0.269	0.0134
NGS	0.223	0.134	0.0203
avGFP	0.184	0.191	0.0428
amacGFP	0.766	0.0427	0.0243
cgreGFP	0.542	0.518	0.0233
ppluGFP	0.334	0.248	0.0199

Supplementary Table 8. Overall epistasis and deleteriousness. Fitted parameters for all six datasets as shown in Figure 2C and D.

	Precision	Recall	f1-score	support
GFP	0.8	0.69	0.74	1583
AmCyan	0.75	0.52	0.62	981
GFP/AmCyan	0.67	0.47	0.55	285
Non-Functional	0.97	0.99	0.98	34112
accuracy			0.96	36961
macro avg	0.8	0.67	0.72	36961
weighted avg	0.96	0.96	0.96	36961

Supplementary Table 9. Classification accuracy metrics for the random forest. Conducted using an independent test set.

L42	V68	Q69	S72	T108	V112	Y145	T167	H181	L220	V224	Functional class	Enrichment (log2)
V	A	A	T	E	V	Y	T	H	V	I	AmCyan	7.2
V	A	A	T	E	V	M	T	H	V	I	AmCyan	7.6
V	A	A	T	E	V	F	T	H	V	I	AmCyan	5.2
V	A	A	T	E	V	Y	V	H	V	I	AmCyan	7.8
V	A	A	T	E	V	Y	T	H	L	I	AmCyan	6.2
V	A	A	T	E	V	I	T	H	V	I	AmCyan	5.2
V	A	A	T	E	V	Y	T	L	V	I	AmCyan	5.2
V	A	A	T	E	V	M	V	L	V	I	AmCyan	5.2
V	A	A	T	E	I	Y	T	H	V	I	GFP	1.3

Supplementary Table 10. Enrichment values for the highly connected cluster.

Source	Functional class	# designs tested	# functional designs
Deep-sequencing data	GFP (488/530nm)	10	6 (60%)
	AmCyan (405/525nm)	10	8 (80%)
	GFP & AmCyan	4	3 (75%)
	Total	24	17 (71%)
	Max number of mutations	6	0
	Designs of special interest	3	1 (33%)
Random forest predictions	GFP (488/530nm)	5	4 (80%)
	AmCyan (405/525nm)	5	4 (80%)
	GFP & AmCyan	4	4 (100%)
	Undetermined	5	4 (80%)
	Total	19	15 (79%)
Sorted for brightness	GFP (488/530nm)		4

Sorted for spectral shifts	GFP (488/530nm)	3
	AmCyan (405/525nm)	10

Supplementary Table 11. Individually expressed and tested designs.

Supplementary Table 12. Biophysical characterization of the individually tested designs.

Bibliography

1. Goldenzweig, A. & Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **87**, 105–129 (2018).
2. Shoichet, B. K., Baase, W. a., Kuroki, R. & Matthews, B. W. A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences* **92**, 452–456 (1995).
3. Somermeyer, L. G. *et al.* Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* vol. 11 Preprint at https://doi.org/10.7554/elife.75842 (2022).
4. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
5. Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T. & Ben-Tal, N. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* **58**, 610–617 (2005).
6. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).

7. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **4**, (2008).

8. Wilding, M., Hong, N., Spence, M., Buckle, A. M. & Jackson, C. J. Protein engineering: the potential of remote mutations. *Biochem. Soc. Trans.* **47**, 701–711 (2019).

9. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).

10. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).

11. Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **19**, 1817–1819 (2010).

12. Zhao, Y., Zhang, W., Zhao, Y., Campbell, R. E. & Harrison, D. J. A single-phase flow microfluidic cell sorter for multiparameter screening to assist the directed evolution of Ca2+ sensors. *Lab Chip* **19**, 3880–3887 (2019).

13. Ai, Henderson & Remington. Directed evolution of a monomeric, bright and photostable version of Clavularia cyan fluorescent protein: structural characterization and applications in fluorescence …. *Biochem. Biophys. Res. Commun.*

14. Platisa, J., Vasan, G., Yang, A. & Pieribone, V. A. Directed Evolution of Key Residues in Fluorescent Protein Inverses the Polarity of Voltage Sensitivity in the Genetically Encoded Indicator ArcLight. *ACS Chem. Neurosci.* **8**, 513–523 (2017).

15. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).

16. Domingo, J., Baeza-Centurion, P. & Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).

17. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).

18. Weinreich, D. M., Watson, R. A. & Chao, L. PERSPECTIVE: SIGN EPISTASIS AND GENETIC COSTRAINT ON EVOLUTIONARY TRAJECTORIES. *Evolution* **59**, 1165–1174 (2005).

19. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).

20. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **25**, 1260–1272 (2016).

21. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).

22. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2**, e00631 (2013).

23. Dellus-Gur, E. *et al.* Negative Epistasis and Evolvability in TEM-1 β-Lactamase—The Thin Line between an Enzyme's Conformational Freedom and Disorder. *J. Mol. Biol.* **427**, 2396–2409 (2015).

24. Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **72**, 178–186.e5 (2018).

25. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat.*

Rev. Mol. Cell Biol. **20**, 681–697 (2019).

26. Khersonsky, O. & Fleishman, S. J. What Have We Learned from Design of Function in Large Proteins? *BioDesign Research* **2022**, 1–11 (2022).
27. Pakhomov, A. A. & Martynov, V. I. GFP family: structural insights into spectral tuning. *Chem. Biol.* **15**, 755–764 (2008).
28. Rodriguez, E. A. *et al.* The Growing and Glowing Toolbox of Fluorescent and Photoactive Proteins. *Trends Biochem. Sci.* **42**, 111–129 (2017).
29. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications* vol. 10 Preprint at https://doi.org/10.1038/s41467-019-12130-8 (2019).
30. Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33–38 (1996).
31. Unger-Angel, L. *et al.* Protein recognition by bivalent, ‘turn-on’ fluorescent molecular probes. *Chemical Science* vol. 6 5419–5425 Preprint at https://doi.org/10.1039/c5sc01038a (2015).
32. Ai, H.-W., Shaner, N. C., Cheng, Z., Tsien, R. Y. & Campbell, R. E. Exploration of new chromophore structures leads to the identification of improved blue fluorescent proteins. *Biochemistry* **46**, 5904–5910 (2007).
33. Bandyopadhyay, B. *et al.* Local energetic frustration affects the dependence of green fluorescent protein folding on the chaperonin GroEL. *J. Biol. Chem.* **292**, 20583–20591 (2017).
34. Weinstein, J., Khersonsky, O. & Fleishman, S. J. Practically useful protein-design methods combining phylogenetic and atomistic calculations. *Curr. Opin. Struct.*

Biol. **63**, 58–64 (2020).

35. Lambert, T. J. FPbase: a community-editable fluorescent protein database. *Nat. Methods* **16**, 277–278 (2019).
36. Lipsh-Sokolik, R. *et al.* Combinatorial assembly and design of enzymes. *Science* **379**, 195–201 (2023).
37. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3**, e3647 (2008).
38. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5869–5874 (2006).
39. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).
40. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
41. Borg, I. & Groenen, P. J. F. *Modern Multidimensional Scaling: Theory and Applications*. (Springer Science & Business Media, 2005).
42. Mccandlish, D. M. Visualizing fitness landscapes. *Evolution* **65**, 1544–1558 (2011).
43. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).
44. Frenzel, E., Legebeke, J., van Stralen, A., van Kranenburg, R. & Kuipers, O. P. In vivo selection of sfGFP variants with improved and reliable functionality in industrially important thermophilic bacteria. *Biotechnol. Biofuels* **11**, 8 (2018).

45. Henche, A.-L., Koerdt, A., Ghosh, A. & Albers, S.-V. Influence of cell surface structures on crenarchaeal biofilm formation using a thermostable green fluorescent protein. *Environ. Microbiol.* **14**, 779–793 (2012).

46. Banerjee, S. *et al.* Mispacking and the Fitness Landscape of the Green Fluorescent Protein Chromophore Milieu. *Biochemistry* vol. 56 736–747 Preprint at https://doi.org/10.1021/acs.biochem.6b00800 (2017).

47. Cotlet, M., Goodwin, P. M., Waldo, G. S. & Werner, J. H. A Comparison of the Fluorescence Dynamics of Single Molecules of a Green Fluorescent Protein: One- versus Two-Photon Excitation. *ChemPhysChem* vol. 7 250–260 Preprint at https://doi.org/10.1002/cphc.200500247 (2006).

48. Campbell, B. C., Petsko, G. A. & Liu, C. F. Crystal Structure of Green Fluorescent Protein Clover and Design of Clover-Based Redox Sensors. *Structure* **26**, 225–237.e3 (2018).

49. Lac, A., Le Lam, A. & Heit, B. Optimizing Long-Term Live Cell Imaging. *Methods Mol. Biol.* **2440**, 57–73 (2022).

50. Snapp, E. L., Altan, N. & Lippincott-Schwartz, J. Measuring protein mobility by photobleaching GFP chimeras in living cells. *Curr. Protoc. Cell Biol.* **Chapter 21**, Unit 21.1 (2003).

51. Russ, W. P. *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).

52. Freschlin, C. R., Fahlberg, S. A. & Romero, P. A. Machine learning to navigate fitness landscapes for protein engineering. *Curr. Opin. Biotechnol.* **75**, 102713 (2022).

53. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).

54. Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).

55. Trudeau, D. L. & Tawfik, D. S. Protein engineers turned evolutionists—the quest for the optimal starting point. *Current Opinion in Biotechnology* vol. 60 46–52 Preprint at https://doi.org/10.1016/j.copbio.2018.12.002 (2019).

56. Goldenzweig, A. *et al.* Automated Structure-and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 1–10 (2016).

57. Weinstein, J. J., Goldenzweig, A., Hoch, S.-Y. & Fleishman, S. J. PROSS 2: a new server for the design of stable and highly expressed protein variants. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1071.

58. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

59. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

60. Altschul, S. F., Gertz, E. M., Agarwala, R., Schäffer, A. A. & Yu, Y.-K. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.* **37**, 815–824 (2009).

61. Fleishman, S. J. *et al.* RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One* **6**, e20161 (2011).

62. Ke, Meng, Finley & Wang. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*

63. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

64. Pedregosa *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*

65. Blecher-Gonen, R. *et al.* High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat. Protoc.* **8**, 539–554 (2013).

66. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Research* vol. 21 487–493 Preprint at https://doi.org/10.1101/gr.113985.110 (2011).

67. Frith, M. C., Wan, R. & Horton, P. Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res.* **38**, e100 (2010).

68. Huynh, K. & Partch, C. L. Analysis of protein stability and ligand interactions by thermal shift assay. *Curr. Protoc. Protein Sci.* **79**, 28.9.1–28.9.14 (2015).

69. Cranfill, P. J. *et al.* Quantitative assessment of fluorescent proteins. *Nat. Methods* **13**, 557–562 (2016).

70. Fery-Forgues, S. & Lavabre, D. Are Fluorescence Quantum Yields So Tricky to Measure? A Demonstration Using Familiar Stationery Products. *J. Chem. Educ.* **76**, 1260 (1999).