# Decoding Stability and Epistasis in Human Myoglobin by Deep Mutational Scanning and Codon-level Machine Learning

Christoph Küng[1, 2] , Olena Protsenko[1,2], Rosario Vanella[1, 2,] , Michael A. Nash[1, 2, *]

[1]Institute of Physical Chemistry, Department of Chemistry, University of Basel, 4058 Basel, Switzerland
[2] Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland
* Correspondence to: michael.nash@unibas.ch

**Abstract**

Understanding the linkage between protein sequence and phenotypic expression level is crucial in biotechnology. Machine learning algorithms trained with deep mutational scanning (DMS) data have significant potential to improve this understanding and accelerate protein engineering campaigns. However, most machine learning (ML) approaches in this domain do not directly address effects of synonymous codons or positional epistasis on predicted expression levels. Here we used yeast surface display, deep mutational scanning, and next-generation DNA sequencing to quantify the expression fitness landscape of human myoglobin and train ML models to predict epistasis of double codon mutants. When fed with near comprehensive single mutant DMS data, our algorithm computed expression fitness values for double codon mutants using ML-predicted epistasis as an intermediate parameter. We next deployed this predictive model to screen > $3 \cdot 10^6$ unseen double codon mutants *in silico* and experimentally tested highly ranked candidate sequences, finding 14 of 16 with significantly enhanced expression levels. Our experimental DMS dataset combined with codon level epistasis-based ML constitutes an effective method for bootstrapping fitness predictions of high order mutational variants using experimental data from variants of lower order.

In protein engineering, understanding the intricate relationships between sequence, structure and function is paramount for identifying variants with enhanced properties. A common approach uses laboratory directed evolution (1,2), however, success rates are low and the variant discovery process is stochastic. An alternative approach seeks to shift screening burden towards *in silico* methods using ML models that predict protein phenotype from sequence. However, phenotype prediction for point missense mutations at a single position remains challenging, and the challenge intensifies for mutations at multiple positions. Epistatic effects, the phenomenon whereby the fitness impact of one mutation depends on the presence of another, further adds layers of complexity. Hence, quantifying epistatic effects and discovering hidden synergies between mutations is an open challenge that once solved will improve our ability to develop models for predicting protein phenotype.

Synonymous codon substitutions (i.e. silent mutations) are in many cases not silent, but can subliminally influence protein phenotype. Changes in codon choice within a gene may lead to discontinuous translation speeds that influence co-translational folding (3). Furthermore, mRNA structure and cellular stability determined by the transcript sequence are affected by codon choices. Codon optimization is routinely performed for optimizing expression in recombinant hosts, however, no generally agreed upon algorithm exists (4,5). Based on the importance of codon choice in protein synthesis, ML-based approaches to phenotype prediction should ideally consider codon-level sequence information.

The cost and scarcity of high-quality large scale labeled biological data currently hamper our ability to develop and train ML models for protein engineering (6,7). One promising approach to this problem is offered by DMS, a technique that uses massively parallelized methods to assay the effects of large numbers of genetic mutations (8–10). Integration of DMS datasets into machine learning frameworks has potential to significantly improve the predictive power of ML models and help bridge the gap between low numbers of experimental data points and large numbers of computational predictions (11-12).

Prior work in the field has combined DMS and next-generation sequencing (NGS) and trained models to predict fitness of unseen *in silico* sequences (13,14). For example, NGS and DMS were combined to study epistasis using GB1 variant libraries containing single and double mutations (15). Weinstein and colleagues made use of ML to discover beneficial mutations in GFP and avoid unfavorable epistatic effects (16). Recently, a neural network was used to successfully predict highly expressed and stable variants of human myoglobin utilizing evolutionary and structural information (17). ML models trained on codon level sequence information have also been applied to tune expression levels of protein sequences (18). ML-based codon optimization was also achieved by treating encoding triplets in codon boxes (4). The effects of codon usage on protein expression were further shown by studying the dependence of codon identity on backbone torsion angles of the corresponding amino acids (19).

Here, we build on this previous work on codons, DMS, NGS and ML and report a combined experimental-computational pipeline to predict expression levels of double codon mutant sequences. We used deep mutational scanning data derived from a yeast surface displayed protein library to train a variety of supervised ML models and predict epistasis of double codon mutants. We then used experimental fitness values of single mutants combined with ML predictions to arithmetically compute a double mutant fitness value. The combination of a near comprehensive NNK single codon mutant variant library featuring a small fraction of

combinatorial double mutations equipped us with a rich labeled dataset to predict epistatic effects with high confidence.

We used our DMS/ML workflow to study human myoglobin (hMb), a relatively small, stable 17 kDa heme protein that served as an ideal model for our study. We generated DMS data quantifying the expression levels of yeast displayed single and double mutant hMb variants and used them to train models to predict double mutant epistasis and further make fitness predictions. We further show that DMS display/expression scores correlate with thermostability of soluble variants, and discovered a highly stable disulfide bonded hMb variant with a denaturation temperature >12 °C higher than WT.

## Results

### DMS data generation and validation

To generate training data for our epistasis learning algorithm, we used nicking strand mutagenesis and constructed a pooled gene library comprising 10,467 hMb variants that were expressed and displayed as Aga2p fusion proteins on the surface of EBY100 yeasts (20). Our library comprised 4,403 single amino acid substitutions (89.3% of all possible), 5,076 double mutations (0.1% of all possible) and small fractions of higher order mutants. We measured the expression level of each variant by quantitative high-throughput fluorescence activated cell sorting (FACS) combined with NGS (**Figure 1)**. Yeast cells displaying hMb variants were stained with a fluorescent antibody in proportion to hMb expression level. The pooled library was then sorted into four bins (**Figure 1B**) and the gene sequences of variants in each bin were determined by NGS. The frequencies of variant reads in the bins were transformed into expression fitness scores using a custom computational pipeline employing tools from (21) and (22) (see methods). Fitness scores for single mutations ($n_{Codon_variants}$= 3,934) and double mutations ($n_{Codon_variants}$ = 3,650) were then used to train the ML models (**Figure 1C**).

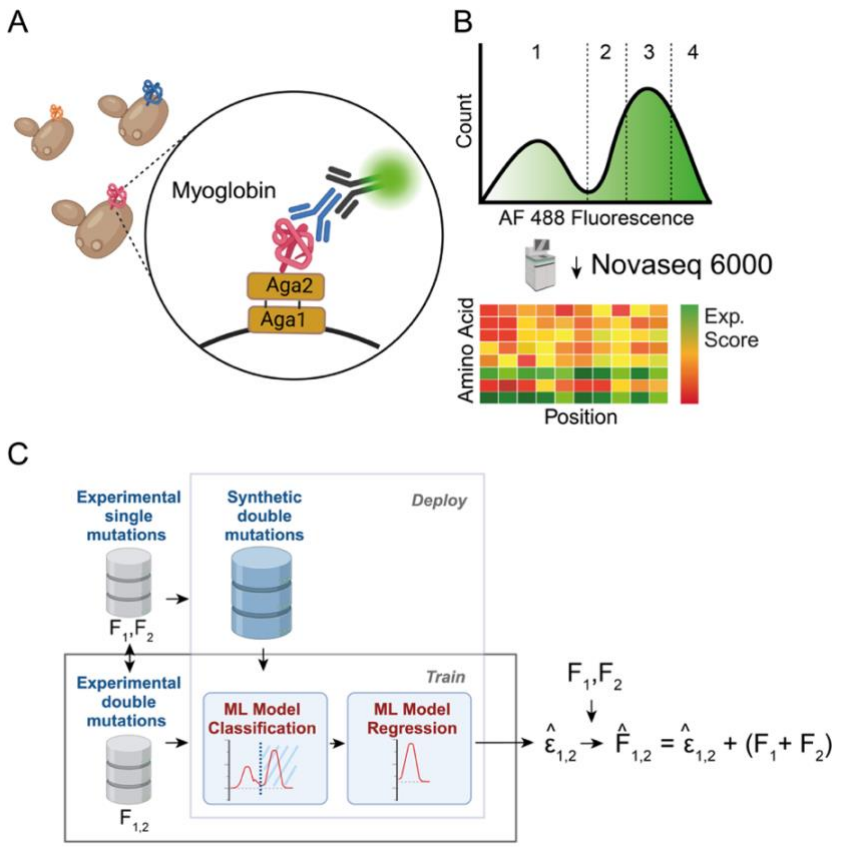


**Figure 1)** Scheme of the experimental and computational workflow. A) A myoglobin variant library is displayed on the surface of yeast cells and fluorescently labeled. B) The labeled library is sorted into four bins with known median fluorescence intensities. After sequencing the variants in each bin, we convert read counts into expression fitness scores to obtain the fitness landscape. The data nearly comprehensively cover fitness values for single NNK codon mutants ($F_1$, $F_2$) as well as a small fraction of double ($F_{1,2}$) codon mutants. C) An ML model consisting of a classifier and a regressor is trained in two stages. The classifier is trained to predict whether a variant lands in the low or high expression peak of the bimodal expression distribution. The regressor is trained to predict epistasis of double codon mutants ($\hat{\varepsilon}_{1,2}$). Once trained, the algorithm uses $\hat{\varepsilon}_{1,2}$ as an intermediate parameter and predicts double mutant fitness $\hat{F}_{1,2}$ of any variant for which the corresponding single mutant expression scores are known.

To assess reproducibility of experimental DMS fitness scores, we performed two independent biological replicates and evaluated their correlation (**Figure 2A**). These data showed strong agreement between replicates ($r$ = 0.90, Pearson's correlation coefficient p-value < 0.0001) with higher confidence for variants that were covered by a larger number of sorted cells (**Figure 2A**, color scale). As further validation, we analyzed the distribution of expression fitness values of nonsense (stop codon) and synonymous mutations (**Figure 2B**), and found that nonsense mutations (**Figure 2B**, red) exhibited low expression fitness scores (-0.52 ± 0.13, n = 126). Synonymous gene variants (**Figure 2B**, blue) that encoded WT hMb using alternate codons exhibited expression fitness values of 0.03 ± 0.07 (n = 109) on average, indicating the relevant range of different fitness values for WT. We investigated fluctuations in WT expression caused by synonymous codon usage and observed significant differences in expression levels depending on sequence position (**Figure S1).** Additionally, codon level influence on experimental expression level of missense mutations was analyzed by plotting fitness scores for the same amino acid substitutions encoded by different codons in a scatter plot (**Figure S2)**. These analyses (**Figure S1** and **S2**) together demonstrate how synonymous codons can influence expression levels in an unpredictable way.

Missense mutations encoding single substitutions generated a bimodal expression fitness histogram (**Figure 2B,** orange) across a broad range of values (fitness range: - 0.60 to 0.26). Most single missense mutations (67.15%) exhibited negative fitness values, indicating expression levels lower than WT hMb. To further validate expression fitness scores derived from DMS, we selected 16 variants covering a range of positive and negative fitness values in the DMS dataset. We independently synthesized their genes and tested expression fitness ($F_{exp}$) in monogenic cultures. The observed fluorescence intensity distribution following antibody staining was then used to calculate a single variant expression score (**Figure 2C**) using the same procedures as for the *en masse* DMS experiments. We found that monogenic expression scores were strongly correlated with those measured by pooled DMS ($r$ = 0.94, $p$ <0.0001). Taken together, these controls demonstrate that fitness scores determined by DMS report with high fidelity on the expression levels of thousands of variants in parallel.

We note that there is evidence that yeast surface display expression level is generally correlated with protein folding stability (23–26). Thermodynamically destabilizing mutations yield lower detectable expression levels because less stably folded domains are susceptible to yeast quality control mechanisms which hinder the secretion of unstable variants. Although mRNA stability, protease susceptibility and other factors may also influence surface display expression levels, we consider the estimated fitness scores as relevant to thermodynamic folding stability (see below).

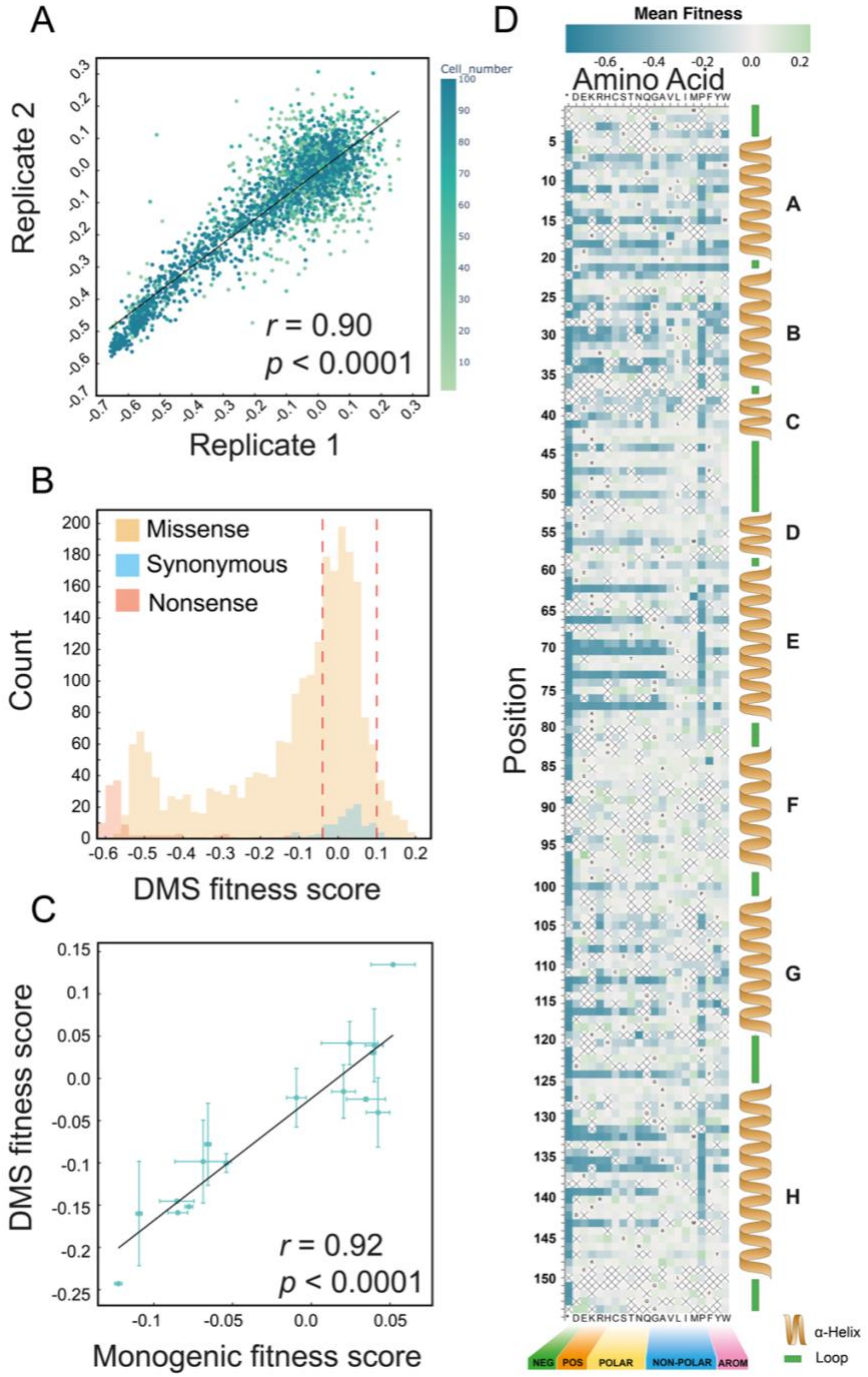



**Figure 2)** DMS of yeast displayed human myoglobin. A) Linear correlation of fitness scores from two biological replicates. The color gradient represents the number of cells carrying a particular hMb variant found in both replicates. B) Distribution of fitness scores for nonsense (red), missense (orange) and synonymous single mutations (blue). Red dashed lines indicate the width of the fitness range for synonymous mutations. C) Correlation of computationally determined fitness score (DMS) with monogenetic scores, experimentally determined for 16 single variants. Y error bars correspond to standard deviation (Std) of grouped amino identities, X error bars correspond to Std of biological replicates. D) Heatmap showing expression fitness scores of single mutants in the dataset. Amino acids along the X-axis are ordered by biochemical properties. Black letters within the plot indicate WT residue identity at the respective position. X symbols indicate missing variants or variants with too few reads to pass the quality threshold.

We next visualized the fitness values in the form of a heatmap (**Figure 2A**). We noticed conspicuous dark horizontal stripes with semi-regular periodicity within specific regions (residues 7-41, 62-77, 104-120 and 128-147), indicating that residues intolerant for mutations were found with regular spacings. After aligning these patterns with secondary structure features, we found that these corresponded to residues along the sides of alpha helices that faced inwards towards the hydrophobic core of the protein. By fitting a sinusoidal function to the experimental fitness scores for residues 70-79 (Helix E), we obtained a periodicity of 3.59 residues, which matched the periodicity of typical protein α-helices and further confirmed the origins of the periodicity of low fitness values (**Figure S3**). Another observation that emerged from the heatmap was the intolerance for proline mutations within helices, which manifested as a vertical dark stripe covering large ranges of residues. In the heatmap (**Figure 2A**), the amino acid substitutions are ordered by biochemical properties along the x-axis. This depiction further helps to illustrate an intolerance for polar side chains at residues facing the core of the protein. The higher than average tolerance of helix F to mutations is indicative of ligand (heme) binding site. Helix F is not part of the hydrophobic core, but was shown to exhibit collective motion driving ligand entry and exit (27). In the expression score it is tolerant for mutations indicating a stability-activity trade off for hMb binding oxygen *via* the heme group.

**Myoglobin is more tolerant of hydrophobic mutations**

We analyzed the effects of each endpoint amino acid (regardless of position) on fitness scores for all missense mutations. The average fitness of all missense mutants was -0.13, and we scored endpoint amino acid fitness as a percent change from this global average. This allowed assessing the influence of side chain biochemical properties on expression fitness (**Figure S4**). Apart from stop codons, the strongest negative impact was observed for mutations to proline (Avg. score: -123%, n = 122), a well known helix-breaking substitution. Charged amino acids (D, E, K, R and H) were found to severly negatively influence fitness scores (Avg. score: -16%, n = 580). The myoglobin tertiary structure is mostly defined by α-helices built around a strongly hydrophobic core, and we found the least disruptive mutations to be hydrophobic residues (A, V, L , I and M), (Avg. score: +68%, n = 639). The aromatic hydrophobic side chain of phenylalanine also had higher than average fitness (Avg. score: +51%, n = 116).

**Discovery of R32C disulfide bonded myoglobin with enhanced thermostability**

We found that cysteine substitutions were less detrimental to expression on average (Avg. score: +41%, n = 133). One cysteine mutation in particular (R32C) emerged with a fitness score among the highest in the entire dataset (R32C fitness = 0.13). Mapping the position of this mutation onto the 3D structure of myoglobin revealed that the original arginine side chain situated in helix B faced a native cysteine (Cys111) in the adjacent helix G. We hypothesized that the R32C mutation had introduced a new disulfide bond that increased protein stability and thereby expression level. To study this further, we expressed and purified R32C myoglobin from an oxidizing *E. coli* strain (Origami 2). Comparative SDS-PAGE analysis of WT and R32C myoglobin under reducing and non-reducing conditions supported the presence of the disulfide bond (**Figure S5 A**). To further evaluate the presence of this disulfide bond, we analyzed purified samples of WT and R32C variants with mass spectrometry, comparing the R32C variant under non-reducing and reducing conditions (**Figure S5 B**). The increase in detected mass of 2 Daltons for the reduced sample further confirmed the formation of the cysteine-cysteine bond. We used nano differential scanning fluorimetry (NanoDSF) to measure the denaturation temperature ($T_m$), and found a significant increase of +12.5 °C for the disulfide variant (R32C $T_m$= 83.8 °C; WT $T_m$= 71.3 °C, **Figure S6**). Disulfide bond engineering as well as cross-linking of non-canonical amino acids have both been reported for modulating

myoglobin properties (28), however, to our knowledge the R32C myoglobin variant discovered by our DMS screen has not been described in literature.

**Sequence conservation and solvent accessible surface area correlate with expression fitness**
We next analyzed correlations between sequence conservation and per position average expression fitness scores. We used ConSurf (29) to identify homologous sequences and determine conserved positions in myoglobin. The ConSurf tool used HMMER (30) to collect 1,425 myoglobin homologues from the UNIPROT database, of which 1,360 sequences passed similarity and coverage thresholds. Out of those, 300 sequences with the lowest E-value (matches with highest significance) were used to determine per position conservation scores on a scale from 1 to 9. We plotted the per position average fitness of all mutants in the DMS dataset vs. per position sequence conservation and found a significant negative correlation (**Figure S7 A**, ($r$ = -0.92, $p$ <0.0001), indicating more conserved positions were less tolerant for mutation on average and resulted in lower expression. It is important to stress that sequence conservation in myoglobin is driven not only by expression stability, but also by its function as an oxygen reservoir in muscle tissue. Positions 46, 98, 99 and 154 surrounding the heme binding site are crucial for myoglobin function and highly conserved ($CONS_{Score}$ = 9), however, they are highly tolerant for mutations in our expression screen, exhibiting neutral fitness or even positive fitness in the case of position 99. These sites contributed to substantial variability of the negative trend (**Figure S7 A**). These data highlight the ability of DMS to shed light onto activity-stability trade-offs (26).

Many conserved residues were found in the hydrophobic core of hMb, which prompted us to investigate correlations between expression fitness scores and solvent accessible surface area (SASA). SASA scores were extracted using Pymol from an AlphaFold model of human myoglobin (AF-P02144-F1) (31). We assigned every residue a score between 0 and 100, corresponding to the minimum and maximum solvent exposure per position in myoglobin, respectively. The analysis revealed a positive correlation ($r$ = 0.61, $p$ <0.0001, **Figure S7 B**) with a prominent outlier being Asp21 ($SASA_{Score}$ = 47), which joins helix A and B (**Figure S7 C**). We found this position does not accept any other amino acid, including negatively charged glutamate. Based on the observed correlations of expression fitness with sequence conservation of the mutated position and SASA, we decided to encode both of these parameters into the data structure for training ML models to predict epistatic effects (see below).

In addition, we were interested in how well yeast display expression level measured by DMS could be correlated with thermodynamic stability of soluble versions of the same protein sequences. We predicted the mean per position $\Delta\Delta G_{folding}$ of myoglobin (PDB: 3RGK, n = 149) using the FoldX stability prediction algorithm (39). Comparing the predicted values to the DMS expression fitness scores, we found strong negative correlation when averaged per position (**Figure 3A**, $r$ = -0.56, $p$ < 0.0001) but also for single individual mutations (**Figure 3B**, $r$ = -0.43, $p$ < 0.0001), again validating the experimental setup serving as a proxy for thermodynamic stability.

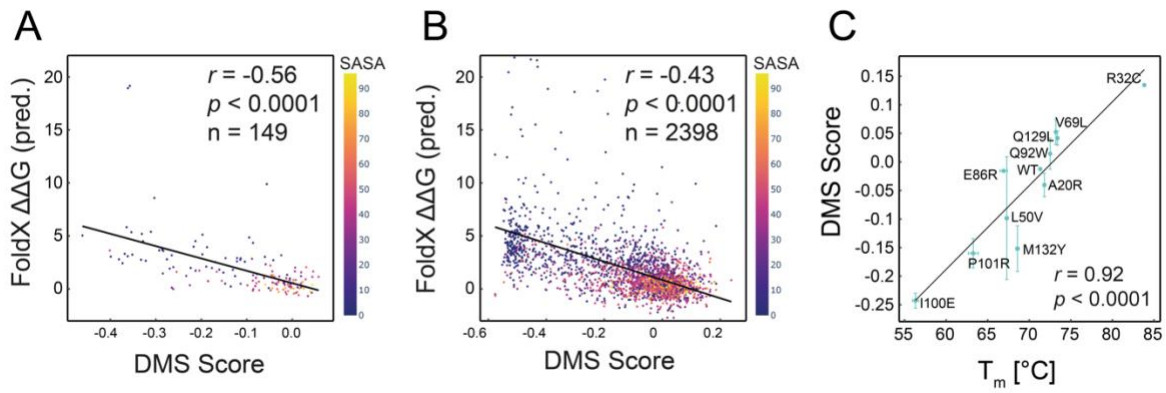



**Figure 3)** DMS scores correlate with thermodynamic stability. A) Scatterplot of FoldX $\Delta\Delta G_{folding}$ against DMS scores, grouped by amino acid residue position. Color coding presents solvent accessible surface area (SASA). B) FoldX $\Delta\Delta G_{folding}$ vs. DMS scores for individual variants. C) Linear correlation of DMS fitness score and melting temperature $T_m$ for 10 single variants along with wild type.

To further evaluate this connection, we expressed and purified 10 of the single mutant variants used for monogenetic DMS score validation along with the WT sequence as soluble proteins in *E. coli*, and assayed thermostability by NanoDSF. We found strong correlation between $T_m$ values measured by NanoDSF and DMS expression fitness scores (**Figure 3C**, $r$ = 0.92, $p$ < 0.0001). Together these data demonstrate that on average across the mutational ensemble, the DMS expression scores from yeast surface display are predictive of thermodynamic folding stability for hMb.

**Machine learning models to predict epistasis**

We next sought to leverage our large experimental dataset by training ML models to predict myoglobin variants with enhanced expression, with a focus on double mutants (i.e. two amino acid substitutions). Although this learning task is somewhat limited, we left the positions of the mutations within the cDNA sequence of myoglobin unrestricted, allowing for mutations across the entire length of the gene. Furthermore, we broke down our DMS dataset into codon level mutations instead of simply considering amino acid changes.

The scheme of our ML algorithm is presented in **Figure 1C**. We began with fitness scores of double mutants (considered at a codon level) as the target parameter to predict. Direct fitness prediction is a common task in protein engineering by ML (13,32–34), however, in our input data we observed no strong correlation between double mutant fitness ($F_{1,2}$) and respective single mutant fitness values ($F_1$ & $F_2$) or their sum ($F_1$+$F_2$) (**Figure S8**), indicating significant levels of epistasis. The task of direct $F_{1,2}$ fitness prediction for variants with mutation sites spanning the entire myoglobin cDNA sequence was apparently too complex given our sparse double mutant training data, and models trained for direct $F_{1,2}$ prediction performed poorly.

We then began investigating the property of mutational epistasis in double mutants, where epistasis of a double mutant is defined as $\varepsilon_{1,2} = F_{1,2} - (F_1 + F_2)$. We developed ML models with the pairwise epistasis $\hat{\varepsilon}_{1,2}$ as the target variable, which indirectly provided us with a way to predict double mutant fitness ($\hat{F}_{1,2}$) *in silico* for unseen variants by combining ML-predicted epistasis ($\hat{\varepsilon}_{1,2}$) with experimental fitness scores $F_1$ and $F_2$ of the corresponding single mutants according to,

$$\hat{F}_{1,2} = \hat{\varepsilon}_{1,2} + (F_1 + F_2) \qquad \text{(Eq. 1)}$$

For encoding our training data of double mutant fitness, we included SASA and conservation scores of the mutated positions as mentioned above, as well as the three-dimensional (3D) distance between mutated residues as measured in the protein structure, with an expectation that 3D-distance would be related to epistasis (35). Mutated codons ($Mut_1$, $Mut_2$) were implemented as categorical features in the input dataset and were transformed using a one-hot encoding method. In (34) it has been shown that this type of encoding achieved the best accuracy when compared with others under a range of regression models.

The ML models were therefore trained to predict $\hat{\varepsilon}_{1,2}$ according to Supplementary Eq. 4 which describes the feature set. The final dataset consisted of 3,650 data points corresponding to double codon mutants.

To address the bimodality of expression fitness scores (**Figure 2C**), we also made use of an upstream classifier. We split the input dataset into low ($F_{1,2} < -0.2$) and high ($F_{1,2} > -0.2$) expression groups, and trained a random forest classification model, achieving accuracy = 0.81, precision = 0.86 and F1 score = 0.85. This upstream classification of variants into low or high expressing pools eliminated bimodality in the training data, which facilitated use of the regression model in a second step.

**ML model evaluation**

We next evaluated the ML regression models for epistasis prediction (35–37). All models were trained on 90% of the input dataset, then tested and evaluated on a 10% test subset. Each mutant in the input dataset was introduced once (excluding duplication). Taking this into account, such a ratio made it possible to train models on a greater variety of mutants and therefore increase the accuracy of the models. The evaluation results for $\hat{\varepsilon}_{1,2}$ prediction on hMb variants with double codon mutations are presented in **Table 1**. The evaluation metrics were the coefficient of determination ($R^2$ ) on test and train datasets (**Figure 4A**), Pearson's ($r$) and Spearman's ($\rho$) correlation coefficients (**Figure 4B**), mean absolute error (MAE), and negative mean squared error (nMSE). All presented models performed well (**Table 1**) in terms of $R^2_{train}$ with values between 0.86-0.96. We found neural networks had larger differences in performance between train and test sets, for example, for the Fully Connected Neural Network, we found $R^2_{train}$ = 0.90 and $R^2_{test}$= 0.60 with Pearson's $r$ = 0.85. The mean absolute error was nearly the same for all models (MAE = 0.10) with the exception of the Convolutional Neural Network (MAE = 0.06). Overall the Random Forest model performed best with $R^2_{test}$ = 0.79 and Pearson's $r$ = 0.92, therefore we chose this as the model to take forward and utilize. Correlation plots of predicted *vs.* measured epistasis ($\varepsilon_{1,2}$ ) and fitness ($F_{1,2}$) for double codon mutants for different ML models are presented in **Figure S9**.

We analyzed the learning curves of the Random Forest model by plotting the negative mean squared error *vs.* the size of the training subset (**Figure 4C**). The learning curves revealed that the Random Forest model predicted double mutant epistasis even after training on relatively small input datasets (<1,000 training examples) with accuracy improvement decelerating after this stage. Nonetheless, model accuracy was still improving after exhausting all our training data. **Figure 4D** shows epistasis values predicted by the fully trained Random Forest model for double codon mutants vs. experimentally measured values (both on test and train subsets). In **Figure 4E**, we plotted double mutant fitness values that were calculated by combining the model prediction of epistasis with the experimental values of single-mutant fitness according to **Eq. 1**.

Compared to epistasis prediction, double mutant fitness prediction ($\hat{F}_{1,2}$) according to **Eq. 1** was less accurate (**Figure 4E)**, with Pearson's $r$ = 0.38, Spearman's $\rho$ = 0.35, and MAE = 0.09. We attributed this significant loss in model accuracy and increase in error when changing from epistasis prediction ($\hat{\varepsilon}_{1,2}$) to fitness prediction ($\hat{F}_{1,2}$) to the different data scales of $\hat{\varepsilon}_{1,2}$ and $\hat{F}_{1,2}$. Predicted epistasis values ($\hat{\varepsilon}_{1,2}$ ) were in a range [-0.47, 0.92] (**Figure 4D**), while predicted fitness scores ($\hat{F}_{1,2}$) were in a narrower range of [-0.18, 0.42] (**Figure 4E**). The epistasis model's MAE of 0.1 therefore represented a smaller percent error compared to the magnitude of epistasis. However, when $\hat{\varepsilon}_{1,2}$ with a give error magnitude was used to predict $\hat{F}_{1,2}$ by **Eq. 1**, the error as a percent of the fitness value was significantly larger.

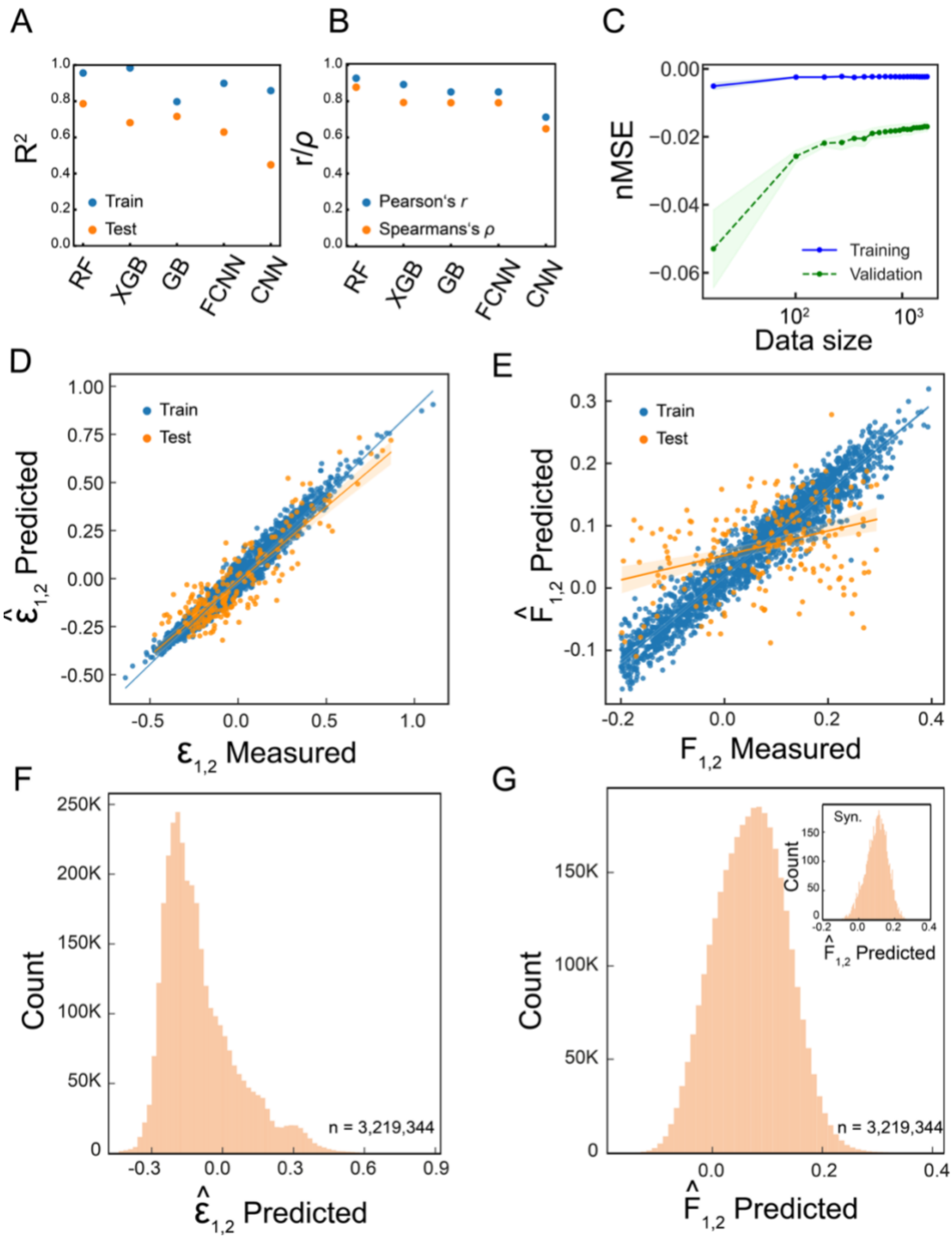



**Figure 4)** Evaluation and testing of ML regression models for epistasis prediction of double codon mutants. A) Coefficient of determination for the train and test sets ($R^2_{train}$, $R^2_{test}$) for each model. Model types included: Random forest (RF); XGBoosting (XGB); Gradient boosting

(GB); Fully connected neural network (FCNN); and Convolutional neural network (CNN). B) Pearson's (*r*) and Spearman's correlation coefficients ($\rho$) for the test set. C) Learning curves for the Random Forest model (nMSE = negative mean squared error). D) Correlation plot of predicted *vs.* measured epistasis ($\hat{\varepsilon}_{1,2}$ ) for double codon mutants. E) Correlation plot of predicted *vs.* measured fitness ($F_{1,2}$) for double codon mutants. F) Distribution of predicted epistasis values for > $3{\cdot}10^6$ *in silico* screened variants. G) Distribution of fitness values for > $3{\cdot}10^6$ *in silico* screened variants. For *in silico* screening in panels E and F, each synthetic sequence at the codon level was first passed through the classifier. Variants that were predicted to be in the high expressing pool were passed to the Random Forest regression epistasis model and $\hat{F}_{1,2}$ was calculated from $\hat{\varepsilon}_{1,2}$ according to **Eq. 1**.

| Model | $R^2_{train}$ | $R^2_{test}$ | Pearson's *r* / Spearman's $\rho$ | MAE |
|---|---|---|---|---|
| Random Forest | 0.96 | 0.79 | 0.92 / 0.87 | 0.10 |
| XGBoosting | 0.98 | 0.68 | 0.89 / 0.79 | 0.10 |
| Gradient Boosting | 0.80 | 0.72 | 0.85 / 0.79 | 0.10 |
| Fully connected neural network | 0.90 | 0.65 | 0.85 / 0.79 | 0.11 |
| Convolutional neural network | 0.86 | 0.49 | 0.71 / 0.65 | 0.06 |

**Table 1**. Evaluation of ML models for predicting epistasis parameters of myoglobin variants with double codon mutations. Five models were trained and evaluated in terms of coefficient of determination on the train ($R^2_{train}$) and test ($R^2_{test}$) datasets, Pearson's correlation coefficient (*r*), Spearman's correlation coefficient ($\rho$), and mean absolute error (MAE).

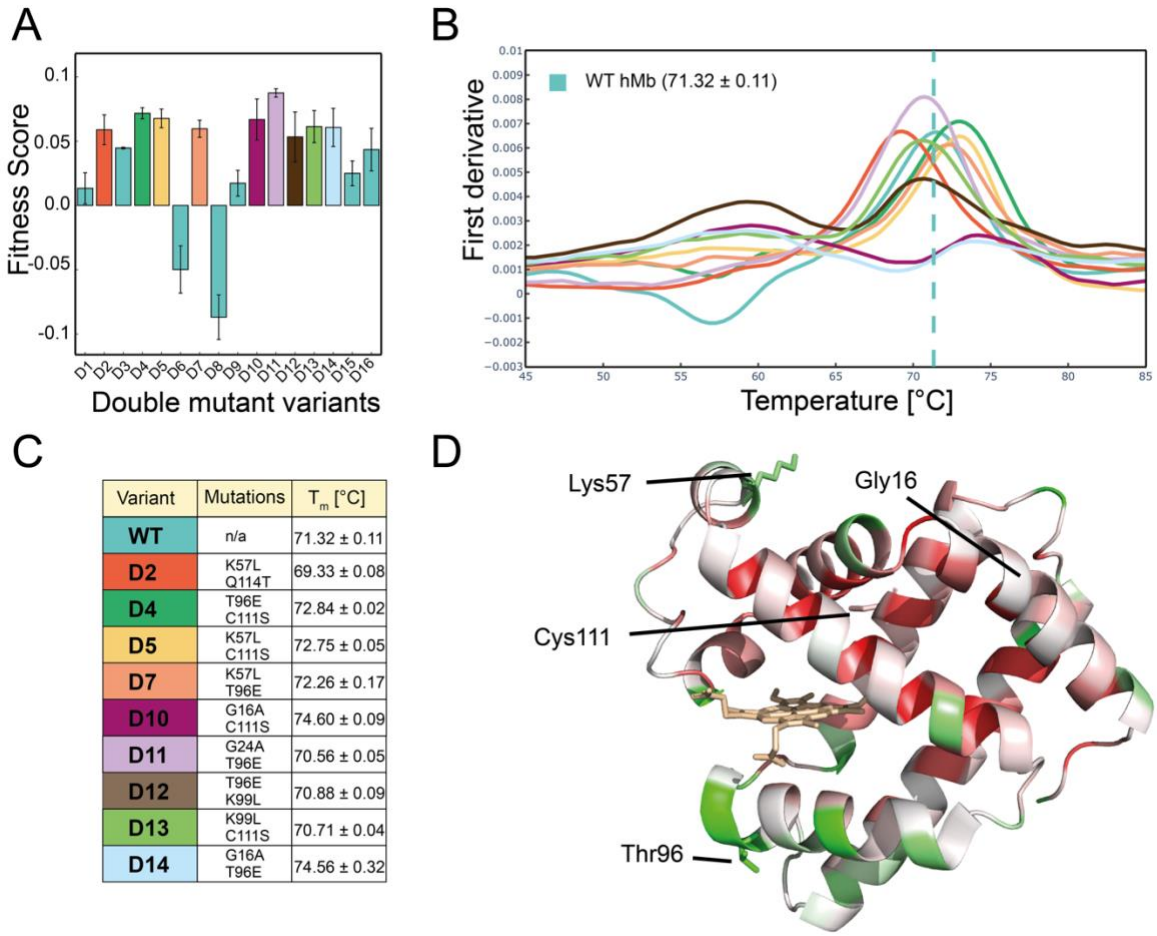
The simulated *in silico* dataset consisted of the unseen variants of double mutations derived from single mutants for which experimental fitness data was available. After data preprocessing, the *in silico* simulated dataset consisted of 3,219,344 variants. We deployed the hybrid classification/regression model to screen *in silico* synthetic myoglobin sequences containing exactly two codon substitutions (i.e., codon-level double mutants) (**Figure 4F, G**). We generated *in silico* codon mutations at two positions in myoglobin, and used the upstream classifier to classify the variants as either high or low expressing. For variants categorized as high expressing, we used the Random Forest regression model to predict the value of epistasis ($\hat{\varepsilon}_{1,2}$) and finally used **Eq. 1** combined with measured experimental fitness data ($F_1$ and $F_2$, respectively) to generate the double mutant fitness prediction, $\hat{F}_{1,2}$. Fitness scores generated in this manner fell in a range $\hat{F}_{1,2} \in$ [ -0.18, 0.41] with mean $\underline{x}$=0.07. We ranked the *in silico* variant list by $\hat{F}_{1,2}$ values and analyzed the frequency of amino acids in the top 5% of variants. The most common end point amino acids were L, S, A, R, and T (**Figure S10**). We also analyzed synonymous mutations for *in silico* simulated combinations. We calculated the difference between the maximal and minimal predicted fitness values $\Delta\ \hat{F}_{1,2} = \hat{F}_{1,2\ max} - \hat{F}_{1,2\ min}$ for each synonymous combination and found differences as high as 0.25.

**Experimental validation of predicted high fitness double codon mutants**

For experimental validation, we extracted the top 5% of simulated double codon mutant variants with the highest predicted fitness scores $\hat{F}_{1,2}$. We then took 16 of these for gene synthesis, cloning, transformation and testing in monogenic culture (38). We found the 16 top ranked variants showed very limited mutational diversity, with many of the same few mutations at the same positions overrepresented (57, 96, 99, 111 and 114). Therefore, we chose a sampling of variants that were highly ranked but also covered several other sequence

positions. The low numbers of variants available to us at this stage highlights a major challenge in ML-based protein engineering. Although there is much discussion about challenges in generating training data, it is also challenging to test large numbers of variants with predicted high fitness values derived from *in silico* screening. In our case, we settled on testing 16 variants that were predicted and highly ranked by the algorithm.

Following the same procedure as described for the validation of single mutants scores, the expression scores for the highly ranked double mutants were determined by comparison with WT (**Figure 5A**). We used the same conditions to express the protein variants for the monogenic assays, and adapted the quantitative FACS gating to mimic the DMS experiment (see methods). Out of 16 variants tested in monogenic format, we found that 14 (87.5%) were more highly expressed than WT. From those 14, 5 sequences were moderately more expressed ($0 < F_{exp} < 0.05$), while another 9 were expressed at substantially higher levels ($F_{exp} > 0.05$).




| Variant | Mutations | $T_m$ [°C] |
|---|---|---|
| WT | n/a | 71.32 ± 0.11 |
| D2 | K57L Q114T | 69.33 ± 0.08 |
| D4 | T96E C111S | 72.84 ± 0.02 |
| D5 | K57L C111S | 72.75 ± 0.05 |
| D7 | K57L T96E | 72.26 ± 0.17 |
| D10 | G16A C111S | 74.60 ± 0.09 |
| D11 | G24A T96E | 70.56 ± 0.05 |
| D12 | T96E K99L | 70.88 ± 0.09 |
| D13 | K99L C111S | 70.71 ± 0.04 |
| D14 | G16A T96E | 74.56 ± 0.32 |

**Figure 5)** Evaluation of predicted high fitness double mutants A) Experimental monogenetic fitness scores determined on yeast cell surface for 16 double mutants found among the top predicted expressors. Error bars correspond to standard deviation of biological triplicates. B) NanoDSF thermal denaturation curves of soluble proteins (first derivatives of fluorescence emission ratio (350 nm/330 nm) of the 9 best expressors found in yeast. Curves represent the mean of three replicates. C) Mutations and melting temperatures of the top double mutants expressors. D) Per position averaged stability fitness score mapped on the hMb 3D structure. WT residues of the top positions (Gly16, Lys57, Thr96, Cys111) implicated for fitness enhancement are labeled and shown as sticks.

After verifying the correlation of thermostability and DMS scores for single mutants (**Figure 3C**), we tested the top 9 ($F_{exp}$ > 0.05) expressing double codon variants as soluble proteins purified from *E. coli*. All variants were well expressed with a strong reddish color, suggesting heme binding. The variants were purified and yielded bands at an expected size by SDS-PAGE (**Figure S11**). NanoDSF showed that most variants exhibited melting temperatures similar to WT (71.32 ± 0.11 °C) (**Figure 5C**). The top 9 of these double codon mutant variants with high ML-predicted expression fitness in the yeast display assay were therefore found to have higher-than-WT fitness in a monogenic assay format on the yeast surface, and WT-like $T_m$ when expressed and purified as soluble proteins from *E. coli.*

Certain variants were found to exhibit a two-peak thermal denaturation profile with an intermediate state. For example, variants D10 and D14 shared a Gly16Ala mutation and both exhibited a two-peak thermal denaturation profile (**Figure 5B**). Similarly, variants D12 and D13 shared a Lys99Leu substitution and exhibited thermal denaturation profiles more complex than WT. Considering the higher unfolding peak, variants D10 and D14 were completely unfolded at 74.60 ± 0.09 °C and 74.56 ± 0.32 °C, respectively, making them the most stable variants expressed in *E. coli.*

Three other mutations found in variants D4, D5 and D7 conferred thermal stability higher than WT ($T_m$ = 72.84 ± 0.02 °C, 72.75 ± 0.05 °C and 72.26 ± 0.17 °C, respectively). These three variants contain combinations of three mutations: Thr96Glu, Cys111Ser and Lys57Leu. Combinations with Cys111Ser seem to be the most beneficial, as those combinations score high on both yeast surface display and as soluble proteins. In combination with the other two mutations in D11 and D13 however, the positive effect was lost. The accompanying mutations Gly24Ala and Lys99Leu, respectively, both are high expressors as single mutant variants (**Figure 2D**). Since the same is true for Cys111Ser, we see negative epistasis for bacterial expressed soluble proteins.

Enhanced stability in yeast display (i.e. increased expression level) could be related to O-glycosylation primarily at serine and threonine residues. Coincidentally, all the selected double mutants tested as soluble proteins either contained a WT Ser/Thr that was removed, or a newly substituted Ser/Thr. Therefore, we speculate that these mutations resulted in changes to protein glycosylation. When expressed in bacteria, the beneficial glycosylation change was lost due to a lack of these posttranslational modifications in *E. coli*. In previous work expressing enzymes by yeast surface display, we found increased stability due to changes in glycosylation patterns, similar to here (40).

To test this further, we displayed D5, D11, WT, and an empty cassette carrying only the Aga2 carrier protein on the yeast surface and purified the Aga2-fusions by reducing the disulfide bonds in between Aga2 and Aga1. After His-tag purification, elution fractions were loaded on a 15 % SDS page and compared to their deglycosylated counterparts (**Figure S12**). In fact, all myoglobin variants are heavily glycosylated on the yeast as seen by smears at higher molecular weight. After treatment with EndoH, most smears are focused to one clear band, suggesting N-glycosylation sites. However, the remaining smears could indicate additional O-glycosylation. The variants selected for this experiment were either adding (D5, Cys111Ser) or removing (D11, Thr96Glu) a potential O-glycosylated residue.

**Discussion**

In this report, we demonstrate merger of DMS, NGS and ML to predict expression level phenotype of double hMb mutants at the codon level. Our comprehensive DMS-derived fitness landscape of single point mutations of hMb showed structural trends and tolerance for mutations in specific regions spread across the entire sequence length. The rich dataset allowed discovery of crucial sites for stability, characterization of alpha helix pitch from mutational stability data, and identification of sequence stretches with increased mutational tolerance. We further discovered a highly stabilized interhelical *de novo* disulfide bonded variant of hMb. Together, these findings highlight the power of DMS in understanding of hMb stability.

We further implemented an ML algorithm trained on the DMS dataset. Our codon-level model was trained on epistasis of thousands of double mutants, and was deployed to calculate *in silico* double mutant fitness values for millions of unseen variants. The predicted highly expressing double mutant variants were tested experimentally, with 14/16 found to be more highly expressed than WT. Our generalizable workflow and codon-level epistasis-based predictive algorithm for protein fitness scores can be broadly used in the future to assist in the construction of gene libraries that are enriched in high fitness variants. Such an approach has potential to accelerate protein engineering and molecular discovery campaigns with broad applications in industry and biomedicine.

**Author contributions**

CK: Conceptualization, methodology, investigation, formal analysis, writing (original draft); OP: Conceptualization, methodology, investigation, formal analysis, writing (original draft); RV: Conceptualization, methodology, writing (review & editing); MAN: Conceptualization, methodology, writing (review & editing), funding acquisition.

**Conflict of interest**

The authors have no conflicts of interest to disclose.

**Acknowledgements**

This work was supported by the University of Basel, ETH Zurich and the Swiss National Science Foundation (200021_191962). The authors further thank Vanni Doffini, Sean Boult and Jaime Fernandez de Santaella for helpful discussions.

**Data availability**

The data necessary to reproduce the presented results, including raw sequencing files, scripts, custom computer codes as well as other raw data associated with figures, are available at: DOI: 10.5281/zenodo.10658344. Additionally, an interactive colorblind accessible heatmap is provided in the repository.

## References


1. Zeymer C, Hilvert D. Directed Evolution of Protein Catalysts. Annu Rev Biochem. 2018 Jun 20;87:131–57.
2. Arnold FH. Directed Evolution: Bringing New Chemistry to Life. Angew Chem Int Ed Engl. 2018 Apr 9;57(16):4143–8.
3. Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. Cell Communication and Signaling. 2020 Sep 9;18(1):145.
4. Fu H, Liang Y, Zhong X, Pan Z, Huang L, Zhang H, et al. Codon optimization with deep learning to enhance protein expression. Sci Rep. 2020 Oct 19;10(1):17617.
5. Xu Y, Liu K, Han Y, Xing Y, Zhang Y, Yang Q, et al. Codon usage bias regulates gene expression and protein conformation in yeast expression system P. pastoris. Microbial Cell Factories. 2021 Apr 26;20(1):91.
6. Kouba P, Kohout P, Haddadi F, Bushuiev A, Samusevich R, Sedlar J, et al. Machine Learning-Guided Protein Engineering. ACS Catal. 2023 Nov 3;13(21):13863–95.
7. Aghazadeh A, Nisonoff H, Ocal O, Brookes DH, Huang Y, Koyluoglu OO, et al. Epistatic Net allows the sparse spectral regularization of deep neural networks for inferring fitness functions. Nat Commun. 2021 Sep 1;12(1):5225.
8. Deep mutational scanning: a new style of protein science | Nature Methods [Internet]. [cited 2023 Dec 27]. Available from: https://www.nature.com/articles/nmeth.3027
9. Reich L "Luther," Dutta S, Keating AE. Generating High-Accuracy Peptide-Binding Data in High Throughput with Yeast Surface Display and SORTCERY. In: Stoddard BL, editor. Computational Design of Ligand Binding Proteins [Internet]. New York, NY: Springer; 2016 [cited 2024 Feb 13]. p. 233–47. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-3569-7_14
10. Höllerer S, Desczyk C, Muro RF, Jeschek M. From sequence to function and back – High-throughput sequence-function mapping in synthetic biology. Current Opinion in Systems Biology. 2024 Mar 1;37:100499.
11. Vanella R, Kovacevic G, Doffini V, Santaella JF de, Nash MA. High-throughput screening, next generation sequencing and machine learning: advanced methods in enzyme engineering. Chem Commun. 2022 Feb 17;58(15):2455–67.
12. Nikolados EM, Oyarzún DA. Deep learning for optimization of protein expression. Current Opinion in Biotechnology. 2023 Jun 1;81:102941.
13. Fahlberg SA, Freschlin CR, Heinzelman P, Romero PA. Neural network extrapolation to distant regions of the protein fitness landscape [Internet]. Synthetic Biology; 2023 Nov [cited 2024 Jan 3]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.11.08.566287
14. Freschlin CR, Fahlberg SA, Romero PA. Machine learning to navigate fitness landscapes for protein engineering. Current Opinion in Biotechnology. 2022 Jun 1;75:102713.
15. Olson CA, Wu NC, Sun R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. Current Biology. 2014 Nov 17;24(22):2643–51.
16. Weinstein JY, Martí-Gómez C, Lipsh-Sokolik R, Hoch SY, Liebermann D, Nevo R, et al. Designed active-site library reveals thousands of functional GFP variants. Nat Commun. 2023 May 20;14(1):2890.
17. Sumida KH, Núñez-Franco R, Kalvet I, Pellock SJ, Wicky BIM, Milles LF, et al. Improving Protein Expression, Stability, and Function with ProteinMPNN. J Am Chem Soc [Internet]. 2024 Jan 9 [cited 2024 Jan 22]; Available from: https://doi.org/10.1021/jacs.3c10941
18. Constant DA, Gutierrez JM, Sastry AV, Viazzo R, Smith NR, Hossain J, et al. Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression [Internet]. bioRxiv; 2023 [cited 2023 Dec 27]. p. 2023.02.11.528149. Available from: https://www.biorxiv.org/content/10.1101/2023.02.11.528149v1
19. Rosenberg AA, Marx A, Bronstein AM. Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. Nat Commun. 2022 May 20;13(1):2815.
20. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA. Plasmid-based

one-pot saturation mutagenesis. Nat Methods. 2016 Nov;13(11):928–30.
21. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018 Sep 15;34(18):3094–100.
22. BBMap: A Fast, Accurate, Splice-Aware Aligner (Conference) | OSTI.GOV [Internet]. [cited 2023 Oct 27]. Available from: https://www.osti.gov/biblio/1241166
23. Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. Proceedings of the National Academy of Sciences. 2017 Feb 28;114(9):2265–70.
24. Shusta EV, Holler PD, Kieke MC, Kranz DM, Wittrup KD. Directed evolution of a stable scaffold for T-cell receptor engineering. Nat Biotechnol. 2000 Jul;18(7):754–9.
25. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell. 2020 Sep 3;182(5):1295-1310.e20.
26. Rosario Vanella, Christoph Küng, Alexandre A. Schoepfer, Vanni Doffini, Jin Ren, Michael A. Nash. Understanding Activity-Stability Tradeoffs in Biocatalysts by Enzyme Proximity Sequencing. bioRxiv. 2023 Jan 1;2023.02.24.529916.
27. Begun A, Molochkov A, Niemi AJ. Protein tertiary structure and the myoglobin phase diagram. Sci Rep. 2019 Jul 25;9(1):10819.
28. Iannuzzelli JA, Bacik JP, Moore EJ, Shen Z, Irving EM, Vargas DA, et al. Tuning Enzyme Thermostability via Computationally Guided Covalent Stapling and Structural Basis of Enhanced Stabilization. Biochemistry. 2022 Jun 7;61(11):1041–54.
29. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016 Jul 8;44(Web Server issue):W344–50.
30. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions | Nucleic Acids Research | Oxford Academic [Internet]. [cited 2023 Dec 29]. Available from: https://academic.oup.com/nar/article/41/12/e121/1025950?login=false
31. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research. 2022 Jan 7;50(D1):D439–44.
32. Siedhoff NE, Schwaneberg U, Davari MD. Chapter Twelve - Machine learning-assisted enzyme engineering. In: Tawfik DS, editor. Methods in Enzymology [Internet]. Academic Press; 2020 [cited 2024 Jan 3]. p. 281–315. (Enzyme Engineering and Evolution: General Methods; vol. 643). Available from: https://www.sciencedirect.com/science/article/pii/S0076687920302275
33. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences. 2019 Apr 30;116(18):8852–8.
34. Nikolados EM, Wongprommoon A, Aodha OM, Cambray G, Oyarzún DA. Accuracy and data efficiency in deep learning models of protein expression. Nat Commun. 2022 Dec 15;13(1):7755.
35. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, et al. Inferring protein 3D structure from deep mutation scans. Nat Genet. 2019 Jul;51(7):1170–6.
36. Domingo J, Diss G, Lehner B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature. 2018 Jun;558(7708):117–21.
37. Sailer ZR, Harms MJ. High-order epistasis shapes evolutionary trajectories. PLOS Computational Biology. 2017 May 15;13(5):e1005541.
38. Gietz RD, Woods RA. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol. 2002;350:87–96.
39. The FoldX web server: an online force field - PubMed [Internet]. [cited 2023 Dec 28]. Available from: https://pubmed.ncbi.nlm.nih.gov/15980494/
40. Vanella R, Ta DT, Nash MA. Enzyme-mediated hydrogel encapsulation of single cells for high-throughput screening and directed evolution of oxidoreductases. Biotechnol Bioeng. 2019 Aug;116(8):1878–86.