Discovery of Electron Hole-hopping Redox Mutations in Myoglobin by Deep Mutational Learning

Christoph Küng[1, 2], Alperen Dalkıran[3], Rosario Vanella[1, 2], Diego A. Oyarzún[3, 4], 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, 4056 Basel, Switzerland.
[3] School of Informatics, University of Edinburgh, 10 Crichton St, Edinburgh, EH8 9AB, UK.
[4] School of Biological Sciences, University of Edinburgh, Max Born Crescent, Edinburgh, EH9 3BF, UK.

* Correspondence to: michael.nash@unibas.ch

Abstract

In addition to storing molecular oxygen, myoglobin catalyzes peroxidase-like reactions involving high valency iron(IV)-oxo species that support oxidation of a range of substrates at an open active site. Until now, it was unclear whether long-range electron transfer via hole-hopping could contribute to myoglobin's catalytic cycle. Here we used enzyme proximity sequencing (EP-Seq) to measure the peroxidase activity levels of >6,000 human myoglobin variants. The resulting fitness landscape reveals how aromatic substitutions, in particular surface-exposed tryptophans, can enhance peroxidase activity. Using protein language models in tandem with feedforward neural networks, we trained an accurate fitness predictor on the experimental dataset, and applied it to evaluate >4M double mutant variants. The predictions suggested a beneficial role for hole-hopping mutations in improving peroxidase activity. We experimentally tested 20 high scoring variants in a yeast display assay, all of which outperformed wild type myoglobin. Three selected variants were also tested in soluble format and similarly showed improved performance. A focused combinatorial library yielded a top double tryptophan variant (Q92W/F107W) with 4.9-fold higher catalytic efficiency than wild type. These results show that hole-hopping pathways can be identified and engineered through deep mutational learning, with broad implications for biocatalyst and redox enzyme design.

Keywords

Myoglobin peroxidase-activity; deep mutational scanning; hole-hopping electron transfer; protein engineering; proximity labelling; fitness landscape; yeast surface display.

Introduction

Myoglobin is a small, globular heme protein found predominantly in the cardiac and skeletal muscle of vertebrates, comprising a single polypeptide chain folded into eight α-helices[1]. Best known for its role in oxygen storage and transport, myoglobin buffers oxygen levels during high metabolic demand.[2] In addition to this classical function, myoglobin has more recently attracted interest for its catalytic activity[3,4]. This arises from its iron-containing protoporphyrin IX cofactor, which can reversibly cycle oxidation states to catalyze redox reactions. In the presence of hydrogen peroxide (H_2O_2), the iron center can be oxidized to form high-valent iron(IV)-oxo species referred to as Compound I/II. These reactive intermediates can catalyze one-electron oxidations on a broad range of organic substrates facilitated by an open and accessible active site geometry.

The peroxidase activity of myoglobin is normally suppressed in myocytes due to the reducing environment of the cell. However, there is evidence that this activity can occur *in vivo* under pathological conditions[5–7]. For example, in rhabdomyolysis or upon reperfusion of ischemic tissue, increased levels of reactive oxidative species (ROS) create conditions that support heme-mediated oxidation.[8–10] Myoglobin peroxidase activity may contribute to ROS detoxification in these settings, but it can also lead to oxidative damage[9,11]. When endogenous antioxidants like ascorbate or glutathione are depleted, myoglobin can oxidize lipids and damage proteins and DNA. Recognizing this expanded catalytic potential, researchers have sought to engineer myoglobin for a range of applications, including dye decolorization, and antibiotic degradation.[12–16]

Multiple studies point to tyrosine and tryptophan substitutions as important for engineering electron transfer pathways in peroxidases. These aromatic residues either directly enhance peroxidase activity, or increase cofactor reduction rates by electron donors. Their evolutionary appearance has been linked to the oxygenation of Earth's atmosphere, suggesting an adaptive response to oxidative stress[17–20]. Prior work by Gray and Winkler has emphasized both the catalytic and protective roles of such residues in natural enzymes, showing that they can extend redox activity beyond the active site and help shuttle oxidative equivalents through protein scaffolds.[21–23] Introducing these residues into a stable single-domain protein like myoglobin offers a powerful platform for dissecting protein-based radical chemistry and guiding rational enzyme design.

To systematically understand how mutations could influence the peroxidase-like activity of myoglobin, we sought out a high-throughput method that could provide suitable data for variant discovery with machine learning (ML). Deep mutational scanning (DMS) is one such powerful approach that enables massively parallel analysis of protein sequence-function pairings.[24–26] However, building high-throughput platforms to assay enzymatic activity remains challenging.[27] Linking genotype to enzymatic phenotype requires the ability to compartmentalize enzymatic reactions, as well as distinguish between signal generation due to improved catalytic properties from that attributable to higher enzyme abundance (i.e. expression level). Although droplet-based systems with colorimetric readouts[28–30] and survival-based selections[31–33] can be used to enrich functional variants, these approaches suffer from infrastructure and biochemical limitations, and frequently confound improved enzyme activity with increased protein expression levels.

To address this challenge, we recently developed enzyme proximity sequencing (EP-Seq), a high-throughput method based on yeast display that enables parallel assessment of both protein expression levels and enzymatic activity (**Figure 1**)[26,34,35]. This platform uses phenoxy radical-based labelling chemistry to couple enzymatic activity to yeast cell fluorescence. Highly active variants can then be separated from inactive variants by fluorescence-activated cell sorting (FACS). In prior work, this strategy relied on exogenous horseradish peroxidase for radical labeling, however in the present study, we used the intrinsic peroxidase activity of surface-displayed myoglobin[34]. Leveraging our high-throughput dataset, we then applied machine learning to model the mutational fitness landscape of human myoglobin and predict peroxidase activity of all double mutants.

Recent advances have established ML as a powerful tool for protein variant prediction[37]. The fusion of deep mutational scanning with machine learning, termed deep mutational learning (DML), offers an efficient method to gain deeper understanding in sequence-function relationships[38,39]. Unlike many ML approaches that yield opaque predictions, our DML framework facilitated interpretable hypothesis generation by recommending substitutions with residues that enable hole hopping. This connection, learned from the training data, illustrates how integrating DMS with ML can yield mechanistic insights rather than serving solely as an uninterpretable prediction tool.

This analysis led to the identification of highly active variants which were validated experimentally. Importantly, we further demonstrated that activity trends observed in the yeast-display format were recapitulated in the corresponding soluble enzyme versions, underscoring the generalizability of our approach.

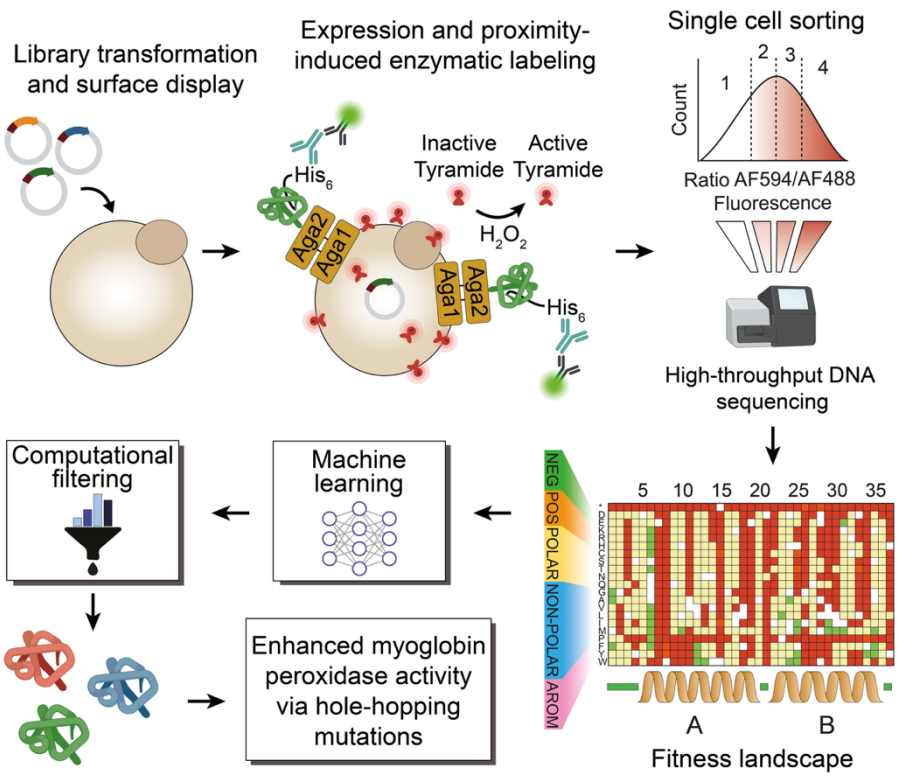

Figure 1) Experimental scheme. A barcoded library is transformed in EBY100 (*S. Cerevisiae*) strain and encoded protein variants are displayed on the cell surface. After immunostaining quantify expression, the peroxidase activity of myoglobin is used for enzyme-mediated radical proximity labelling *via*

reaction with the peroxidase substrate tyramide-Alexa Fluor 594. This labels the single cells in proportion to the peroxidase activity level of the myoglobin variants display on its surface, allowing library screening in a one-pot reaction. The 2-color labeled library is then sorted into four bins by fluorescent activated cell sorting (FACS) based on the expression-normalized peroxidase activity level. Variant distribution amongst bins is evaluated by next generation sequencing and converted into fitness scores. This largescale fitness landscape is then used to train a variety of machine learning algorithms to predict peroxidase activity from sequence. After model training, *in silico* screening and computational filtering of the highest ranked hits is undertaken to arrive at a final subset which is characterized biochemically in a soluble format. What emerges from this pipeline are highly ranked variants that introduce hole hopping (i.e. Tyr and Trp) mutations onto surface accessible sites of myoglobin, providing pathways for oxidative equivalents to access the active site and improve catalytic activity.

Results

System setup, library construction and barcoding

We first optimized the functional display of WT hMb fused to the C-terminus of the Aga2 yeast anchor protein. We successfully detected protein display by staining the His_6-tag at the C-terminus with a primary and secondary Alexa Fluor 488-conjugated fluorescent antibody by flow cytometry[25]. We next validated selective tyramide-based cell labeling for expressed, catalytically active variants, using negative controls either omitting H_2O_2 or using cells transformed with empty cassette plasmids (**Figure S1**). To generate the site saturated library of the coding region of hMb, we employed primers with nested NNK codons and barcoded each variant with a unique 15 nucleotide barcode (see Methods)[40].

After sequencing the sorted bins, we mapped the reads via a look up table to their corresponding variants, and converted the distribution of the variants among the bins to activity fitness scores. Applying confidence filters to remove variants with low sorting or sequencing coverage, we ended up with a dataset consisting of expression-normalized activity scores for 6,115 variants bearing single and multiple mutations.

Deep mutational scanning elucidates stability-activity trade offs

The results of the deep mutational scanning (DMS) experiment to quantify peroxidase activity were processed by computational filtering to consider only the single-site mutants (**Figure 2**). To assess reproducibility, we calculated activity scores for two biological replicates of the library, and observed a Pearson's r of 0.85 (n = 2,661; $p < 0.0001$), indicating strong agreement between replicates (**Figure 2A**). Data points with higher cell coverage are shown in darker shades, and reflect the improved correlation for variants observed more frequently in the dataset.

We categorized the variants based on mutation type, and visualized the fitness scores in a histogram (**Figure 2B**). By definition, the WT sequence is assigned a score of 0. Synonymous mutations encoding the same amino acid sequence cluster at 0.00 ± 0.06 (n = 102, 3.83% of the single mutants). Nonsense

mutations encoding stop codons (n = 123, 4.62%) exhibit the lowest activity scores of -0.43 ± 0.03, consistent with truncation of the protein chain. The largest and most informative group consists of the missense mutations (n =2,436; 91.54%), describing all single mutant variants with activity scores ranging from -0.47 to 0.3. As observed in other mutational scans, most amino acid substitutions are deleterious[26,41,42]. In our dataset, ~88% of missense mutations reduce peroxidase activity. This fraction is higher than the ~67% of missense mutations that decrease myoglobin expression levels[25]. This higher sensitivity to mutation likely indicates that enzymatic activity imposes stricter constraints than simple folding, making the peroxidase phenotype more vulnerable to mutational disruption than folding stability alone.

To validate the results of the DMS experiment, we conducted control assays for individual variants. We selected random sequences as well as variants with high activity scores in order to cover the full range of activity. Each variant was expressed individually on the yeast surface, and peroxidase activity was measured using the same tyramide proximity labeling protocol that was applied in the pooled screen. As in the *en masse* experiment, we calculated activity scores by quantifying expression-normalized fluorescence shifts (see Methods). The resulting values showed strong agreement with the DMS dataset ($r = 0.96$, $n = 20$, $p < 0.0001$; **Figure 2C**).

Figure 2D presents a heatmap of activity scores for all single-point mutants, where activity is normalized by expression level. Variants with reduced folding stability compared to wild type are shown with a black border. We estimated this stability threshold based on the distribution of synonymous mutants and defined destabilizing variants as those with expression scores < -0.06 (see supplementary info and ref [25]). Variants without a border are either stably expressed or were not present in the expression dataset.

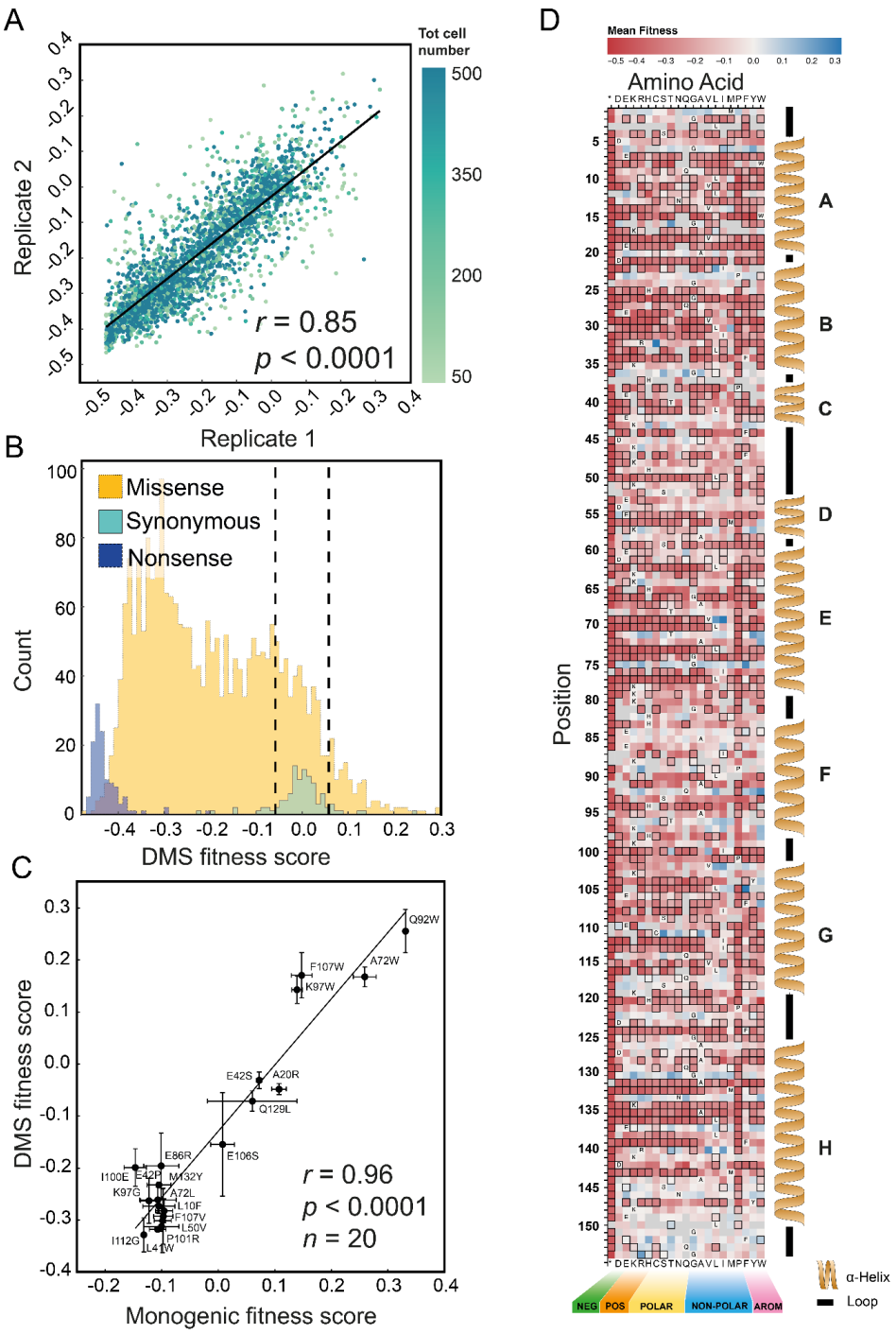

Figure 2) Deep mutational scanning of myoglobin peroxidase activity. **A)** Correlation of activity scores between two biological replicates. Color shading indicates total cell count per variant as indicated in the color bar. **B)** Distribution of activity scores grouped by mutation type. Missense mutations are shown in yellow, synonymous mutations in teal, and nonsense mutations in blue. Dashed lines indicate ±1 standard deviation from the mean of the synonymous codon scores. **C)** Validation of the DMS fitness scores using individually expressed monogenic variants. **D)** Heatmap of expression-normalized peroxidase activity scores for all single mutants. Amino acids are grouped by chemical class. Black borders around the squares indicate variants previously shown to have reduced stability based on a separate expression DMS screen.[25]

In addition to the destabilizing mutations discussed earlier (**Figure 2D,** shown in red with black borders), we also identified variants that are stably expressed on the yeast cell surface but lose their catalytic activity. Examples include mutations at critical positions such as the proximal and distal heme-coordinating histidines (His^{94} and His^{65}, respectively). Other functionally sensitive sites include residues within helix F, a region known to support heme binding but is not required for correct folding of the apoprotein. Helix F is structurally flexible and disrupted in apo-myoglobin[43], which makes it more tolerant to mutation compared to other helices. However, our data show that mutations in this region can severely impair peroxidase activity. This points to an activity-stability tradeoff, and underscores the importance of heme positioning and axial coordination in maintaining catalytic function, even in mutants that fold and express efficiently.

Identification of novel and highly active variants with machine learning

We next sought to discover myoglobin variants with improved peroxidase activity by combining deep mutational data with machine learning. Using the activity DMS dataset, we trained a supervised regression model to predict peroxidase activity from protein sequences encoded using pre-trained protein language models (**Figure 3A**). We evaluated both ESM and ProtTrans embeddings given their strong performance across a variety of protein prediction tasks[44,45]. The supervised regression layer was based on a deep feedforward neural network trained on the DMS activity scores using the protein embeddings as input features. To increase the confidence of model predictions, the training data excluded variants with large variation in DMS scores across replicates (**Figure 2A** and Methods). We also excluded higher-order mutants (4x, 5x) that were poorly covered in the original DMS screen. This filtering step led to a training set with N=4,769 myoglobin variants (**Figure 3B**). After hyperparameter optimization, both ESM and ProtTrans models performed well on held-out test data, and both outperformed models based on one-hot amino acid encoding (**Figure 3C, left)**. We next queried both models with a set of N=20 variants for which we had independently measured monogenic activity scores (**Table S6**; Methods). Comparison of predicted and measured values indicated that ESM embeddings provided better out-of-distribution accuracy. Based on this result, we selected the ESM embeddings for the final computational screen (**Figure 3C, right**).

Given that the training data was highly enriched in double mutants (46.8%) and included only a small fraction of triple mutants (8.7%), we restricted the prediction screen to double mutants that were not present in the training set. This increased reliability of model predictions. Moreover, a two-dimensional projection of both the training and query sequences showed good overlap, which further supports the robustness of the predictions despite the relatively limited coverage of the training data (**Figure 3D**). We embedded a total of N=4,250,505 double mutants using ESM and queried an ensemble of 20 regressors, each trained using five-fold cross-validation and four random seeds for weight initialization. The predicted activity scores for these unseen double mutants followed a distribution similar to that of the training data (**Figure 3E)**. For experimental validation, we selected 20 from a set of 65 consensus hits that scored in the top 0.2% (N=10,000 variants) across the model ensemble (**Table S7**). All 20 of these tested variants showed improved peroxidase activity over the wild type. This perfect success rate demonstrates the effectiveness of our machine learning-guided approach for identifying highly active myoglobin variants.

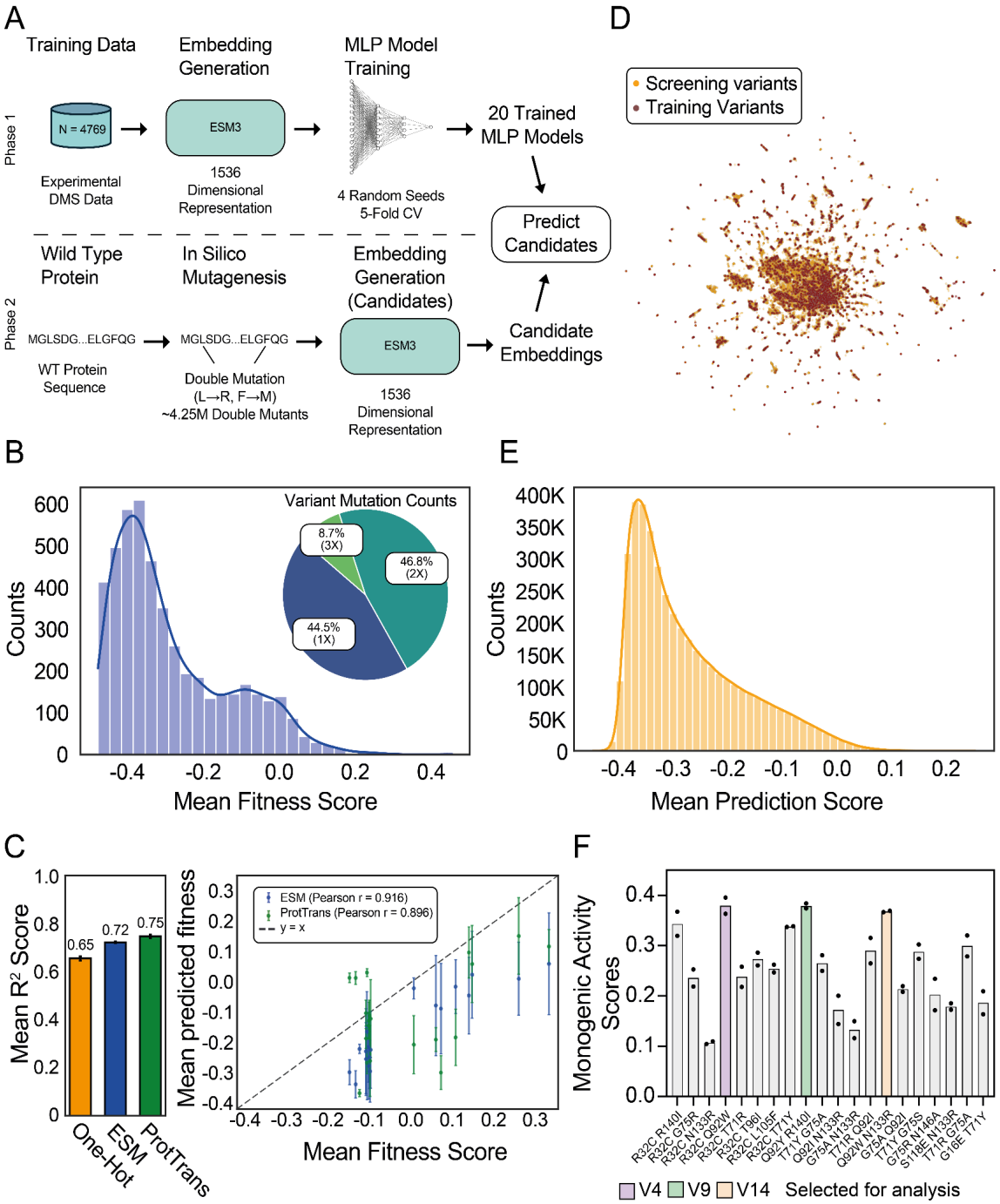

Figure 3) Machine learning workflow, model performance, and experimental validation of predicted high-activity double mutants. **(A)** Overview of our sequence-to-function prediction pipeline. Phase 1: Activity DMS data were used to train a feedforward neural network (multilayer perceptron) regressor using ESM-3 embeddings as input features. Phase 2: An *in silico* library of ~4.25 million double mutants was embedded using ESM-3 and scored using an ensemble of 20 models. **(B)** Characteristics of the training data (N = 4,769 unique variants) after filtering out low confidence variants and higher order mutants. The histogram shows the distribution of experimental DMS scores. The pie chart indicates the proportion of single, double, and triple mutants in the training data. **(C)** Evaluation of model

performance. Left: predictive performance on a held-out test set of two MLP models trained on ESM-3 and ProtTrans embeddings, and a 1D convolutional neural network model trained on amino acid one-hot encodings. Bars are mean R^2 scores between predicted and ground truth fitness scores, computed across predictions from models trained in 5-fold cross-validation; error bars denote one standard deviation across folds. Right: Comparison of model predictions and experimentally measured monogenic activity scores for 20 variants not used in training (**Table S6**). **(D)** Two-dimensional UMAP projection of ESM-3 embeddings of training myoglobin variants and the double mutants screened with the pretrained model. A random sample of 50,000 variants from the full screening library is shown in orange, highlighting extensive overlap in coverage between both libraries. **(E)** Distribution of consensus predicted fitness for the ~4.25 million double mutants. The histogram shows the mean predicted fitness scores across all 20 trained ESM-3 MLP model instances. **(F)** Experimental validation of 20 high-confidence candidate variants selected from the top 0.2% of predicted variants. All tested variants showed activity above wild type.

Analysis of best performing ML predicted sequences as soluble enzymes

After showing that the top machine learning-predicted variants all exhibited higher peroxidase activity than WT myoglobin when displayed on the yeast surface, we further characterized the top three candidates as purified soluble enzymes. These variants, referred to as Var4, Var9, and Var14, were selected based on their top ranked monogenic activity scores and were expressed in *E. coli* along with WT myoglobin. The mutational compositions of these variants are shown in **Figure 4A**, and their positions are mapped onto the 3D structure in **Figure 4D**. Mutations were modeled using AlphaFold and visualized in PyMol as sticks. We obtained the purified protein and could verify by means of reducing and non-reducing SDS page that the R32C mutation, known from our prior study[25], forms a disulfide bond in the ML-predicted double mutant (**Figure 4B)**. The non-reduced form of the protein migrates faster due to the intramolecular disulfide with C111, while addition of β-mercaptoethanol eliminates this change in electrophoretic mobility.

To test peroxidase activity for the soluble variants, we adapted the tyramide labeling assay to a soluble format. Employing the same reaction mixture as with yeast displaying proteins, we stained uninduced yeast cells carrying an empty cassette plasmid by administering soluble hMb variant enzymes. The endpoint fluorescence data (**Figure 4C**) confirmed that all three double mutants produced higher signal than wild type, reproducing the high activity ranking observed in the yeast-displayed format. Negative controls lacking any enzyme showed minimal background fluorescence.

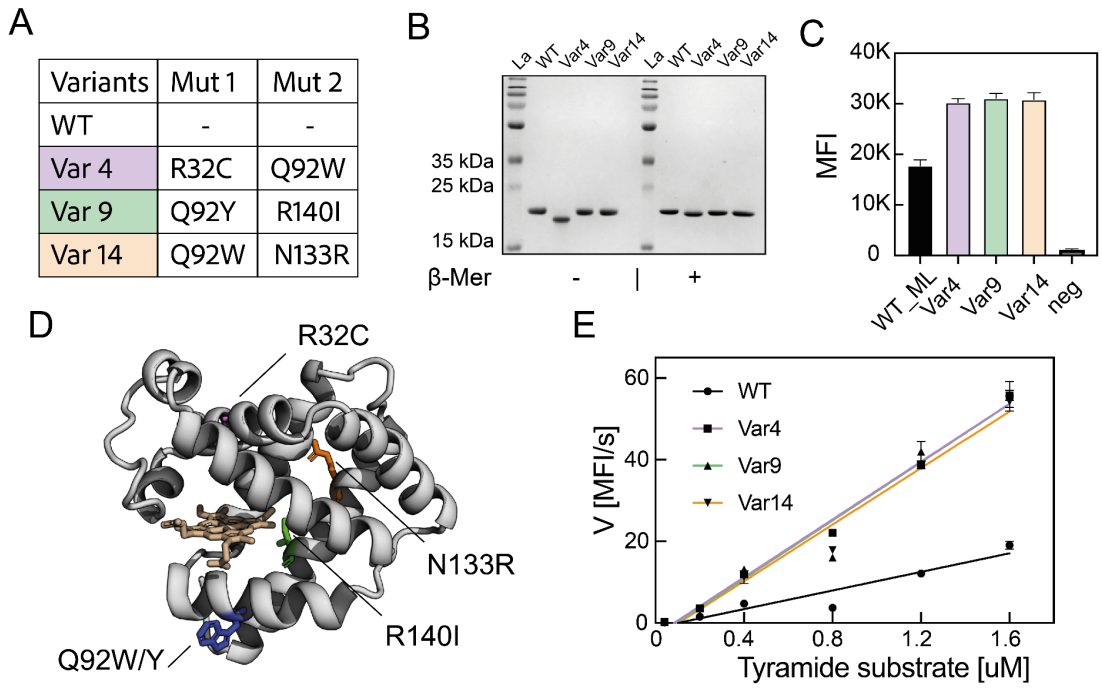

Figure 4) Soluble validation of ML predicted variants. **A)** Selected variants and featured mutations. **B)** 15% SDS-PAGE gel of purified soluble variants. Samples were boiled under non reducing (-) and reducing (+) conditions. **C)** Endpoint MFI of uninduced cells stained by purified double mutants with tyramide AF594. Error bars correspond to STD of triplicates and negative control contains no myoglobin in reaction mix. **D)** Structure showing positions of double mutants, according to color code introduced before. Position Q92 which is mutated in multiple variants is shown here only as Trp and in blue. **E)** Michaelis Menten analysis of soluble enzymes with tyramide. Uninduced yeast cells carrying an empty cassette were stained at different substrate concentrations and endpoint fluorescence assayed at different time points in order to get reaction velocities for variants along with WT.

To gain a deeper understanding of the kinetics of these improved variants, we performed a Michaelis-Menten analysis using the same endpoint-based tyramide labeling assay. Reaction velocities were measured at varying substrate concentrations by stopping the reaction at different timepoints and fitting linear regressions to the endpoint mean fluorescence intensities (MFIs) (see methods). **Figure 4E** shows the resulting velocity curves for WT and the three selected double mutants. Due to the high cost of tyramide, we were unable to reach substrate saturation. Therefore, a linear approximation of the Michaelis-Menten model was used in the low-substrate regime to estimate catalytic efficiency. All three double mutants showed slopes more than three times steeper than WT, indicating high catalytic rates at the concentrations tested, including under the standard library screening concentration which corresponds to 1.2 uM.

In all three of these improved variants (Var4, Var9, and Var14), residue Q92 is mutated to either tyrosine or tryptophan, suggesting that the introduction of a redox-active residue at this surface site contributes to enhanced activity. The R32C mutation found in Var4 likely serves a stabilizing role that supports acquisition of secondary activity-enhancing mutations. This stabilizing effect of R32C was presumably the reason it was frequently found among top ML-ranked double variants. To further

explore the catalytic behavior of the selected variants, we expanded our kinetic analysis to include alternative peroxidase substrates, specifically reactive blue 19 and guaiacol. The results are presented in **Figure S2** and discussed in the Supplementary Note **SN1**.

Strategically placed Trp wires boost activity towards bulky substrates

Among the machine learning-predicted variants discussed above, we observed a clear enrichment of mutations that introduce tyrosine or tryptophan. Beyond the frequently occurring R32C disulfide variant[25] found in many of the top double mutants (**Table S7**), four of the five most active candidates contain either a tyrosine at position 71 or a tyrosine or tryptophan at position 92. This trend is also observable across the whole DMS dataset, where single-point variants containing these substitutions often show higher peroxidase activity. This can be seen in the heatmap in the two rightmost columns in **Figure 2D**. Conversely, mutations that remove native tyrosine residues tend to show very deleterious effects on activity (**Figure S3**). For the two native tryptophan residues we cannot estimate their role in activity since both are elementary for stability and do not tolerate substitution.
When comparing the effects of mutations on protein stability and catalytic activity, we find that aromatic amino acids are particularly important for maintaining function. In many cases, mutations that are tolerated in the expression screen exhibit reduced activity when an aromatic side chain is removed. Wild type tyrosine, histidine and phenylalanine residues tolerate some substitutions, but these almost always come at the expense of catalytic efficiency.

These findings support a broader role for redox-active aromatic residues in modulating myoglobin peroxidase activity. It has been reported that many oxidoreductases possess clusters or chains of tyrosine and/or tryptophan residues that serve as hole hopping relay stations, providing alternative electron transfer pathways to mitigate oxidative stress and preserve function.[21,46] Similarly, dye-decoloring peroxidases use surface exposed aromatic residues as stepping stones for oxidation of bulky substrates that cannot pass through the heme access tunnel.[47,48] Such work has inspired others to use surface tryptophans and tyrosines in myoglobins to enhance dye decoloring peroxidative activity in myoglobins.[12,13]

Consistent with this rationale, our data show increased peroxidase activity for variants introducing tyrosine or tryptophan at surface-accessible positions. In many cases, these substitutions are interchangeable, supporting the idea that their redox potential rather than specific side-chain orientations are beneficial to catalysis (e.g., positions 71, 92, 137, 146 or 152). We visualized this trend in **Figure 5A** by mapping all single tryptophan substitution scores onto the protein surface. Variants that improve activity are shown in blue, and deleterious substitutions are shown in red. Positions where tryptophan residues increase the peroxidase activity are placed around the heme active site on the surface of the protein. This spatial arrangement suggests that these engineered residues facilitate electron transfer for substrates like tyramide, which are too bulky to access the heme directly. We note that the native tyrosine and tryptophan residues are not surface exposed (SASA W8: 4, W15: 2, Y103: 18, Y146:0, where 0 is buried and 100 equals fully exposed). These observations support the idea that enhanced labeling efficiency arises from introduction of redox relays close to the substrate interface.

To examine whether the activity-enhancing effects of individual tryptophan mutations could be combined synergistically, we designed a small combinatorial library targeting the four positions with the highest fitness scores in the single-mutant DMS dataset, which were A72W, Q92W, K97W and F107W. Notably, substitution A72W, which yields one of the most active variants in the library, was the only substitution at that position 72 that retained or improved peroxidase activity, while all others were neutral or deleterious. Furthermore, since most mutations at A72 were neither strongly destabilizing nor beneficial in terms of expression stability[25], this position seems to be functionally specific to peroxidase activity.

The selected variants were modeled using AlphaFold and visualized in PyMol[49] (**Figure 5B**). Residues were color coded according to the legend in **Figure 5C**, where native tyrosine and tryptophan residues are shown in cyan. We synthesized these combinatorial variants using site-directed mutagenesis and tested their peroxidase activity in the yeast surface display system (**Figure 5C**) using the same WT-normalized scoring method as described previously (see **Figure S4**). Consistent with the prior DMS data and ML predictions, all single TRP variants outperformed WT myoglobin (**Figure 5C**). The combinatorial mutants also exhibited improved activity relative to WT, although the effects were more variable. While most combinations did not show improvement upon introduction of additional TRP residues, the impact of F107W was dependent on which TRP mutation it was paired with. In combination with A72W, activity declined, whereas when combined with double mutant Q92W it was amongst the very best variants found in this study.

In addition to tyrosine residues, tryptophans possess electron-rich aromatic side chains that can under certain conditions undergo oxidative modification by peroxidase-generated radicals, especially under high peroxide or high substrate conditions[50]. To ensure that the improved labeling signal observed in our TRP-substituted mutants was not due to the introduction of these additional labeling sites, we performed the decoupled labeling assay with soluble myoglobin variants as described above. We purified WT, Q92W, F107W and Q92W/F107W variants as soluble proteins (**Figure 5D)** and used them to stain uninduced yeast cells containing only an empty cassette plasmid under otherwise identical reaction conditions. This decoupled labelling assay eliminated any artefacts due to the display format, and allowed direct attribution of labeling signal to peroxidase activity. The same performance trend (**Figure 5E**) observed in the display assay was found in this decoupled format, confirming that the enhanced signal was the result of increased enzymatic rate rather than TRP-modification on myoglobin itself.

As presented in **Figure 4** for the ML predicted variants, we quantified reaction velocities at different tyramide concentrations as well for the tryptophan mutants. As shown in **Figure 5F**, the double mutant Q92W/F107W exhibits a 4.9-fold increase in reaction velocity relative to the wild type at the substrate concentration used in the DMS assay. Interestingly, although the single mutants Q92W and F107W each individually enhance reaction velocity by factors of 3.9 and 2.4, respectively, their combined effect in the double mutant was not fully additive, indicating some negative epistatic effects.

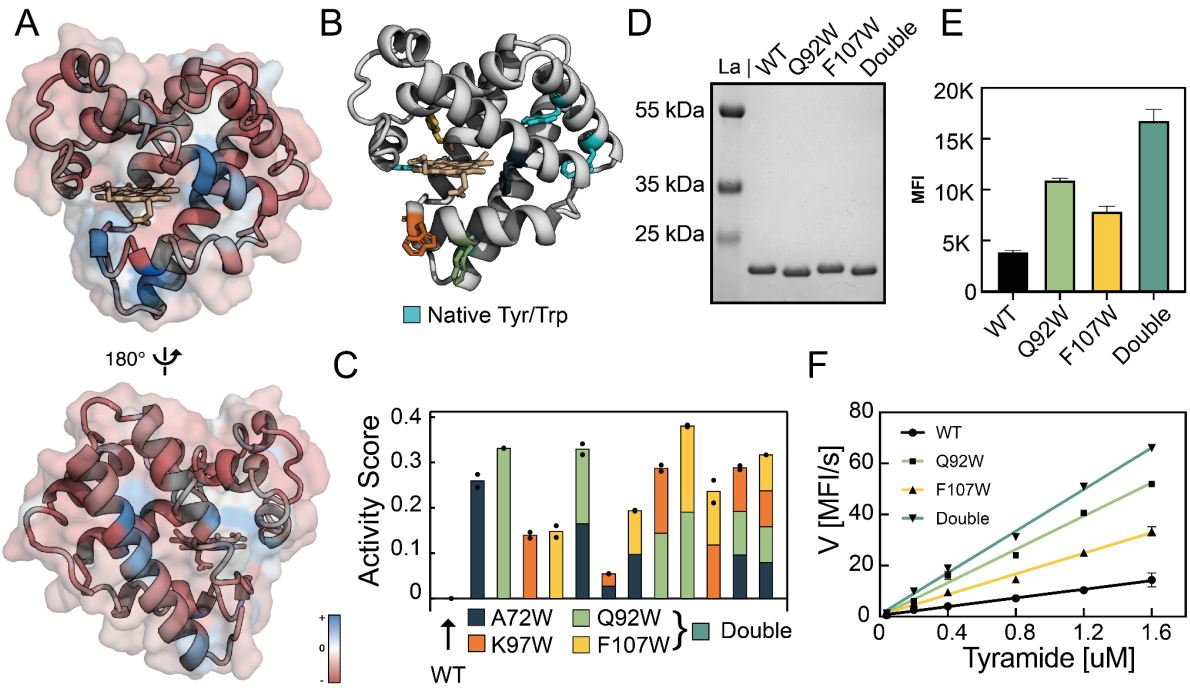

Figure 5) Tryptophan-mediated enhancement of peroxidase activity. **A)** Myoglobin 3D structure with residues colored by single mutant fitness scores for TRP substitutions. Positions without measurements are shown in white as cartoon loops. The heme cofactor is shown in white. **B)** Residues selected for combinatorial Trp-library are shown as sticks and colored according to the legend in panel C. **C)** Barplot of monogenic activity scores for individual TRP mutations and their combinations. **D)** SDS-PAGE gel analysis of purified protein variants. **E)** Decoupled cell labelling assay. Endpoint MFI values are shown for uninduced yeast cells containing an empty cassette plasmid, stained with soluble enzymes and the tyramide fluorophore. **F)** Reaction velocities for soluble enzyme variants measured using a time-dependent tyramide assay. Data represent the linear region used to approximate catalytic efficiency.

We further attempted to explain observed trends for mutants by analyzing the potential electron transfer pathways employing the published tools eMap and EHPath. eMap is a python based web application that predicts possible electron or hole transfer channels from pdb files based on graph theory. Shortest path algorithms are used to estimate shortest pathways from user-specified hole donor (here heme) to the surface of the protein, assigning scores to all pathways[51]. EHPath is another python module estimating and ranking mean residence times of a transferring charge along such hopping pathways[52]. The settings used are described in the method section and the results are presented in **Figure S5** and discussed in supplementary notes **SN2**. We were interested in studying the additivity of the double mutant Q92W/F107W in more detail and hence used purified variants to test the activity in Michaelis-Menten analysis. Equally to the machine learning predicted variants above, we switched to reactive blue 19 as substrate for this experiment. The results including the fits and extracted kinetic parameters are shown in **Figure S7**. We see that the trend observed in the monogenic tyramide scores as well as the decoupled cell labelling assay holds true also for Rb19 dye, with both

single mutants being more active than WT and the double mutant benefiting from additivity of composing mutants.

Next, we assayed how the mutations would act on peroxidase activity towards smaller substrates that might fit in the active site crevice. We performed a Michaelis-Menten assay using guaiacol for the selected variants along with WT myoglobin and saw that in fact the activity towards this smaller substrate is similar for all variants tested (**Figure 4E**). We used molecular docking to further validate our hypothesis, differentiating between bulkier substrates that benefit from a surface radical and smaller ones that do not. We find that for WT, Q92W and F107W the binding site for guaiacol is identical and directly adjacent to the heme cofactor **(Figure S7 F**). Contrary to the machine learning variants, here we study single mutations, allowing for direct allocation of observed effects. Considering that we found improvement towards guaiacol for the machine learning predicted variant 9 which contains the R140I and Q92Y mutations, we suspect that especially the R140I mutation is leading to enhanced guaiacol reactivity, while the tyrosine mutation, similar to the tryptophan substitutions are beneficial for bulky tyramide oxidation.

Lastly, we cross-referenced our deep mutational scanning data with variants reported in gnomAD to identify mutations in myoglobin observed in human clinical populations. We annotated these clinically observed variants with their stability and activity scores derived from our screening assays (**Table S8**). Although our peroxidase screening assay used a non-physiological substrate, these annotations may still provide useful insights into naturally occurring variants. Notably, several clinically observed variants show substitution of wild-type residue with tyrosine or tryptophan, suggesting potential impacts on enhancing oxidation of bulky substrates.

Conclusion

In this study we provide a comprehensive map of the peroxidase activity fitness landscape of human myoglobin, utilizing the high-throughput EP-Seq platform. We highlight regions of activity-stability trade-offs, and show global trends of amino acid groups as well as single mutations with enhanced activity. By leveraging the extensive labeled mutant library, we integrate high-throughput DMS with ML to successfully train a predictive model and identify novel double mutants with elevated peroxidase activity. Notably, all 20 ML-predicted sequences were substantially more active than wild type (WT).

The most promising sequences were expressed as soluble proteins to assess whether their improved activities observed on the yeast cell surface translated to the soluble format. When tested at the same substrate concentration used in the initial library screening, these variants exhibited over threefold higher catalytic efficiency compared to WT. Analysis of the machine learning dataset pointed toward the introduction of oxidizable amino acids such as tyrosine and tryptophan as a key driver of enhanced activity with the bulky tyramide substrate. We validated this hypothesis by constructing a small focused combinatorial library based on top-performing tryptophan mutations. Among these, surface-exposed residues such as Q92 and A72 substantially boosted activity, supporting the hypothesis that these residues serve as redox-active relay stations and facilitate long-range electron transfer. In particular, the combination of mutations F107W and Q92W increased tyramide oxidation by nearly fivefold in assays with soluble proteins.

These findings demonstrate that activity fitness scores derived from the EP-Seq deep mutation scanning platform reliably predict activity trends in soluble enzymes, providing a key validation of this platform. Beyond methodological relevance, the insights reported will find broad applications ranging from engineering bulky dye decoloring peroxidases for industrial use to providing knowledge about mutations that alter peroxidase activity of globins, including hemoglobin based oxygen carriers (HBOCs)[53]. Surface-exposed tyrosine residues have already been shown to accelerate reduction by physiological reductants such as ascorbate in hemoglobin[54]. The strategy developed here could guide future engineering efforts to identify mutations that tune redox activity for safe HBOCs. Finally, we establish a combined approach of high-throughput screening and machine learning to expedite enzyme engineering using yeast surface-displayed libraries, with results directly transferable to soluble enzymes.

Finally, this work underscores the utility of combining high-throughput experimental fitness landscapes with pretrained protein language models to drive hypothesis generation, prioritize variants, and ultimately expand the functional repertoire of enzymes. The ability to use ML predictions to successfully guide mutational searches across vast sequence space represents a generalizable framework for accelerating biocatalyst development.

Author contributions

C.K. and **M.A.N.** conceived the study and drafted the manuscript. **C.K.** carried out the practical work and computational analyses. **A.D.** carried out the machine learning and variant prediction. **R.V.** contributed to the conceptualization and optimization of the EP-Seq experimental and computational workflow. **D.A.O.** designed the machine learning work. **M.A.N.** secured funding and administered the project.

Acknowledgements

This work was supported by the University of Basel, ETH Zurich, and the Swiss National Science Foundation (200021_191962 and 10004516). AD and DAO were supported by a UKRI Engineering Biology Mission Award CYBER under BBSRC grant BB/Y007638/1.

Competing interests

The authors have no conflicts of interest to disclose.

Materials & Methods

Materials

Restriction enzymes, DNA ligase kit and calf intestinal phosphatase (CIP) were purchased from New England Biolabs (Beverly, MA). GelRed™ dye nucleic acid gel stain was purchased from Biotium. The primary and secondary antibodies as well as the tyramide AF488 were purchased from Thermo Fisher Scientific (Waltham, MA). Custom oligonucleotides were obtained from Microsynth (Balgach, CH). The DNA miniprep, gel purification and PCR purification kits were purchased from Thermo Fisher Scientific. Chemicals, if not indicated otherwise, are from Sigma Aldrich (St. Louis, MO).

Myoglobin gene cloning

The homo sapiens myoglobin (Mb) gene was obtained from TWIST Bioscience codon optimized for *Saccharomyces Cerevisiae*. The gene was cleaved in a double digest with restriction enzymes EcoRI and XhoI and cloned in a pYD1 expression vector. Sequence was confirmed by sanger sequencing and the plasmid featuring the myoglobin gene transformed in yeast strain EBY100 by a lithium acetate transformation protocol [55]. Positive colonies were selected on synthetic defined (SD) agar plates lacking tryptophan (-TRP) supplemented with glucose 2% (wt/vol) and ampicillin.

Construction of barcoded myoglobin mutant library

For mutagenesis of the human myoglobin wild type gene, we used a plasmid based one-pot mutagenesis protocol which was previously described [40]. Since higher efficiency was observed for smaller plasmids, we transferred the expression cassette from the pYD1 (5446 bp) to a smaller puc19 plasmid (3647 bp). The puc19 plasmid contains two BbvCI restriction sites up- and downstream of the expression cassette. The original protocol was slightly modified by adding an extra step after the first incubation with Nt.BbvCI enzyme and exonuclease. The reaction mixture was followed by incubation with 10 units of quick CIP phosphatase (New England Biolabs) at 37°C for 20 min, with subsequent incubation at 80°C for 20 minutes. With this step, the removal of 5’ phosphate from nicked and partially degraded strands of wild type gene is promoted and hence limits the reformation of closed wild type plasmids in the following amplification and ligation steps. Then, nested NNK codon containing primers were used for synthesis of the mutated strand targeting codon triplets 1 to 154 (full gene) of the myoglobin wild type gene. Primers were designed with the script create_primers.py from [40] whereas the length of the primers was adjusted to reach a similar melting temperature around 60 °C for all primers. Designed primers were ordered from Integrated DNA technologies in a 96 well plate format and mixed in equimolar ratios to the final concentration of 10 µM. With the mixed primers, the site saturation protocol was performed as described in [40]. After obtaining the plasmid based mutation library, molecular barcodes were introduced through PCR and hence linked to each of the amplified Myoglobin variants. PCR was accomplished by using primers P1 and P2 shown in **Table S1**, adding a 15 N barcode in front of the galactose promoter sequence. The amplified and barcoded library was cloned in a pYDKan backbone, by HiFi DNA Assembly (New England Biolabs) and using the

vector plasmid which was linearized by BbVCI/PmeI double digestion. After column purification (Zymo-Spin I, Zymoresearch), the HiFi assembly product was then transformed into electrocompetent XL1-Blue (Agilent) cells by electroporation (Gene Pulser, Biorad).

Successfully transformed cells were selected for in LB agar plates supplemented with (50 µg/ml) Kanamycin and the size of the library was estimated by counting colony forming units on plates with serially diluted fractions of the library. The total size of the barcoded library was estimated to be 300,000 variants. After incubation at 37°C for 20 hours, the colonies were scraped from the plates and resuspended in liquid LB Kan, further grown for 2 hours and then plasmids extracted by standard Miniprep method. 10 µg from the plasmid library pool was introduced into the *Saccharomyces cerevisiae* strain EBY100, using the lithium acetate transformation protocol[55]. To assess the efficiency of the transformation, successive dilutions of the transformed material were spread onto agar plates containing -Trp and 2% (wt/vol) glucose. Counting of the resulting colony forming units revealed a transformation efficiency around 2 million colonies. Following the transformation, the yeast mutant library was cultivated for 24 hours at a temperature of 30°C in a liquid medium lacking tryptophan (-Trp), supplemented with 2% (wt/vol) glucose. Subsequently, the culture was diluted to an OD_{600} of 1 and transferred to fresh -Trp medium with 2% (wt/vol) glucose to minimize the chance of obtaining multiple transformants. This culture was then allowed to grow for an additional 16 hours at 30°C. Finally, portions containing approximately 5 *10^7 cells were prepared and suspended in a solution of 25% (vol/vol) glycerol before being preserved at -80°C. Plasmid maps are provided at: https://zenodo.org/records/16781055.

Table S1) Primers used for linkage of molecular barcodes to the myoglobin mutant variants. NNK nested primers used for the generation of the library are provided at: https://zenodo.org/records/16781055

Barcoding Primer P1	CGATTTTGTTACATCTACACTGTTGTTATCAGATCAGCGGGTTTAAAC
Barcoding Primer P2	GCGCGGCCTTTTGCCCTGCAGGCCNNNNNNNNNNNNNNNAGGGAACAAAAGCTGG CTAGTACGG
PacBio 1	TTACGGTTCCTGGCGGCCGCG
PacBio 2	GTCGATTTTGTTACATCTACACTGTTGTTATCAGATCAGCGGG
Illumina_for1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGCCTTTTGCCCTGCAGGC
Illumina_for2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGCCTTTTGCCCTGCAGGC
Illumina_for3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGCCTTTTGCCCTGCAGGC

Illumina_for4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCATGGCCTTTTGCCCTGCAGGC
Illumina_rev1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGAGGAGAGTCTTCCTTCGGAGGG
Illumina_rev2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGGAGGAGAGTCTTCCTTCGGAGGG
Illumina_rev3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGGAGGAGAGTCTTCCTTCGGAGGG
Illumina_rev4	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATGGAGGAGAGTCTTCCTTCGGAGGG

In order to limit the size of the library and increase the confidence in barcode assignment from the PacBio-generated lookup table, we capped/bottlenecked the library to 100,000 cells. This was done on a FACS Sony SH800S Sorter by simply gating for viable single cells and sorting 100,000 events in growth media. This culture was grown for 20 h overnight and stored as glycerol stocks at -80°C. The capping was done in yeast cells, since better survival rate was observed when compared to *E.Coli.*

PacBio long read sequencing

In order to link the molecular barcodes to their respective hMb variant sequence, Pacbio long read sequencing was used. The capped library glycerol stock was thawed and used to inoculate 40 mL -TRP and 2% (wt/vol) glucose at an initial OD of 0.1 and allowed to grow for 24 hours. OD then was around 6.8 and 1 mL (6 samples) of the cultures were collected, spun down, and resuspended in 250 µl of Miniprep Resuspension buffer 1 (GeneJET, ThermoFisher). Then 4 µl of Zymolyase was added (Zymoresearch) and incubated for 2 h at 37°C with shaking at 1,000 rpm. Then the standard protocol as suggested in the kit was followed, whereas the final elution volume was 10 µl of prewarmed nuclease free sterile water. Then, the capped library was transformed into electrocompetent E.Coli by electroporation of 6 individual reactions. After outgrowth, amplified plasmids were purified by miniprep and 8 ug were digested by restriction enzymes Not1 and Pme1 for 16 hours. The fragment of interest was isolated by agarose gel electrophoresis and after gel purification of corresponding bands, DNA was cleaned by AMPure XP bead purification.

The purified DNA was then used for SMRT bell ligation and run on a Pacific Biosciences Sequel IIa sequencer with a movie time of 15 hours. Resulting sequences were filtered for quality score Phred 20 and higher. Additionally, sequences were filtered for a maximal length of 1800 bp (amplicon length is 1569). Then, the computational workflow already established and reported in [26] was used to extract the barcode and assign it to a respective mutation. We used Minimap2 [56] to map the long read sequences to a reference file with the wild type sequence of the expression vector PCR amplicon. Then, a C script was used to process the same file output of the mapped sequences using a custom library file (libscodon.h). The script extracts the 15 base pair molecular barcode and its quality along

with the mutational sequence for every read. The output of this step is a first version of the lookup table. Errors such as barcodes assigned to multiple mutants in this raw version of the LUT are then purged in a python script collapsing all the reads, and filtering them for their confidence. Barcodes with the highest number of reads linking the barcode to the mutation are kept and if two barcodes were read the same number of times, the one with the higher quality was kept in the LUT. Then, the LUT was filtered eliminating barcodes that have the wrong length and barcodes that were read less than twice.

Table S2) Computational workflow. Steps and respective commands to map the PacBio reads to the reference and extract the molecular barcode to generate the final look up table (LUT). The raw sequencing files as well as the computational scripts are provided at: https://zenodo.org/records/16781055.

Step	**Tool**	**Command**
Mapping of the PacBio read to the reference file.	Minimap2 (version 2.19)	**minimap2 --cs -ax map-hifi** Myo_ref.fa pacbio.fastq **>** aln.sam
Extracting barcode sequence and quality along with resp mutations and indels.	ppba C script libscodon.h library	**./ppba** Myo_ref.fa aln.sam **>** pre_lut.tsv
Filtering and cleaning of LUT	generate LUT.ipynb Python script	

Expression and labeling of myoglobin barcoded library

Aliquoted glycerol stocks of the library were thawed on ice and resuspended in liquid -TRP medium supplemented with 2% (wt/vol) glucose and Ampicillin 100 µg/ml at a starting OD of 0.1 in 50 mL. After preculture growth for 30 hours with shaking (180 rpm), appropriate volume of cells were spun down and resuspended in 100 mL at an OD 0.4 in -TRP liquid induction medium with 1.8% (wt/vol) galactose, 0.2 % glucose and Ampicillin, buffered with 100 mM sodium phosphate buffer pH 7. Cells were grown at 20°C for 48 hours with shaking at 180 rpm. Then, the culture was spun down, washed once, and resuspended in PBS BSA 0.1% (wt/vol). For cell staining, 4 $*10^6$ cells were collected in a round bottom 96 well plate and incubated in 100 ul of 1:500 dilution of primary Anti Histag Antibody (stock concentration of 1.2 mg mL^{-1}). Afterwards, cells were washed again with 200 ul of PBS BSA 0.1% (wt/vol) and incubated for 20 min in ice cold dilution (1:500, stock concentration 2 mg mL^{-1}) of

secondary goat anti-mouse antibody conjugated with Alexa Fluor 488. Then, cells were washed again twice with cold PBS BSA 0.1%.

Tyramide labeling

For the high throughput screening of the variant and wild type activity, the protein displaying cells were subjected to a single-cell tyramide fluorophore proximity labeling technique (EP-Seq)[26]. The cells, once expressing the library, were stained by primary and secondary antibody labeling, targeting the polyhistdine tag. Then, 3 million cells were added (3000 cells/ul) to 1 mL of 50 mM sodium phosphate buffer, pH 6.1, containing 1:333 dilution of Alexa fluor™ 594 Tyramid reagent (Thermo Fisher, B40957), 100 mM glucose and 62.5 nM Glucose Oxidase (from *Aspergillus niger,* Thermo fisher). The substrate concentration was chosen to yield enough signal to discriminate active from inactive variants, without saturating the yeast cell surface for the WT, with the idea that higher active variants will result in higher fluorescence after labelling reaction (see **Figure S1**) . The exact concentration of the tyramide dye is not reported by the manufacturer. We estimated the concentration by measuring UV-VIS spectrophotometry using an extinction coefficient of $\varepsilon_{TyrAF594}$ = 92'000. Therefore the 1:100 diluted stock solution corresponded to 3.935 uM (see **Figure S8**).

The cells were kept in suspension by shaking at 1000 rpm on a Thermomixer, while incubating for 60 min at 37°C. Then, the reaction was stopped by addition of ice cold 400 uL of PBS BSA (0.1%) supplemented with 0.05% Tween20 and subsequently washed twice by spinning cells to pellets at 13'000 g for 1 min and resuspending in 500 uL of the same PBS buffer. Then, cells were taken to fluorescent activated cell sorting (FACS) (**Figure S1**).

Fluorescent activated cell sorting for deep mutational scanning

The capped, expression-stained and tyramide labeled library was sorted on a BD FACS Aria. The library was loaded on the cell sorter and gates were applied as follows. Gate 1 (Bin (-)) included all non expressing and non-redshifting (inactive) cells. The rest of the cells were then projected on a histogram plot displaying the ratio of the signal of the yellow over blue channel, proxying the expression normalized activity shift. On this histogram we then added three gates to capture the remaining cells by equiproportionate binning (Bin 2(+), Bin 3(++) and Bin 4 (+++) (**Figure S4**) . The same library was grown twice in biological replicates and the sorting experiment was conducted with each of the libraries, sorting each time at least 2.5 Million cells, hence covering the library 25 times. Additionally, we sorted 1000 cells from a separately grown expression culture of wild type with a known barcode in every sort tube. This internal control allows for monitoring any bias in growth conditions in between tubes. The gated populations were sorted in -TRP media supplemented with 1 % BSA to facilitate pelleting of the cells. Cells were spun down and resuspended in 10 mL -TRP medium with 2% glucose and ampicillin. The cultures were let recover and grow for 48 hours at 30°C with shaking at 180 rpm. Then, aliquots of 5 *10^7 cells were stored as glycerol stocks at -80°C.

Illumina barcode sequencing

In order to prepare the sorted bins for high throughput sequencing, glycerol stocks (2 per bin) from both biological replicates were thawn for 5 minutes on ice and spun down at 13000 g for 1 min in a benchtop centrifuge. Then, the supernatant was removed and 250 ul of resuspension solution from bacterial miniprep kit (GeneJET Plasmid Miniprep Kit, Thermo Fisher) were added for resuspension of the pellet. Per tube, 4 ul of Zymolyase was added, and let incubate for two hours at 37 °C with shaking at 1000 rpm. After visual inspection for successful lysis of yeast cell walls (solution appears clear), 250 ul of lysis buffer and subsequently 350 ul of neutralization buffer was added, following the standard protocol for Plasmid miniprep. For better coverage of binned barcodes, two Zymoprep tubes per bin were combined in one miniprep column. After conducting the washing steps, final elution was done in 15 ul of nuclease free water. Those 15 ul then served as template for the PCR reaction, amplifying the 15 N barcode containing region in a standard 50 ul PCR reaction (NEBNext Ultra II Q5 Master Mix, New England Biolabs). 25 ul of Master Mix were added to 15 ul of template DNA with 5 ul of each primer (final conc 1 uM). For forward and reverse primers used to amplify individual binned genetic information, we used staggered primers (Ilumina_for/rev 1-4, **Table S1**), in order to provide different starting nucleotides during illumina sequencing in an else very similar amplicon. Primers were designed for use with the Illumina Nextera indexing library prep.

The amplicons for all bins were then loaded on a standard 1 % (wt/vol) agarose gel to verify correct size and then purified form cut gel using a GeneJET Gel Extraction Kit (Thermo Fisher). After cleaning and concentrating the DNA further using a DNA Clean and Concentrator-5 kit (Zymoresearch), we determined DNA concentration. In a second PCR (**Table S4**) using 1 ng of template DNA, the Nextera indexing sequence was added for all of the bins. Samples were purified with AMPure XP beads, pooled and sequenced through Novaseq 6000 Illumina sequencing.

Table S3) PCR conditions for barcode amplification

Step	**Temp. [°C]**	**Time [s]**	**Cycles**
Initial Denaturation	98	30	1
Denaturation	98	10	20
Annealing	72	30	
Extension	72	10	
Final extension	72	120	1

Table S4) PCR conditions for Illumina Nextera Indexing

Step	Temp. [°C]	Time [s]	Cycles
Initial Denaturation	98	30	1
Denaturation	98	30	8
Annealing	55	30	
Extension	72	30	
Final extension	72	120	1

A computational pipeline was established to process the obtained Illumina reads in order to receive the information about abundance of corresponding mutations in each bin. By using the BBMap algorithm and providing the illumina reference file containing a 15 N barcode, the sequences were mapped and alignment information stored in a sam file [57]. Barcodes were then extracted with the pib.c script and linked through their corresponding mutation by running the rib.c script which searches for the barcodes and assigns tags according to the description in **Table S5**. Finally, all barcodes present in the LUT were sorted and grouped by identity, revealing the total number of each variant in each bin.

Table S5) Pipeline for Illumina read processing. All bins were processed separately through the individual owing steps. The raw sequencing files as well as the computational scripts are provided at: https://zenodo.org/records/16781055.

Step	Tool	Command
Sequence alignment of illumina reads to reference file.	BBMap cite	bbmap.sh in=*.fq ref=reference.fa out=*.sam
Extracting barcode sequence and respective quality score for the illumina reads	pib.c script (process illumina barcode)	pib *.sam > *.fq
Align the extracted barcodes to the previously processed look up table and assign tags (0 = not found, 1 = found, 2 = Read quality below Q20, 3 = Barcode length deviates from 15).	rib.c script (read illumina barcode)	rib -t lut.ts -q 20 *.fq > *.tsv

Filtering for found variants, sort them alphabetically and group them by identity counting the number of barcodes per bin.		grep ^1 *.tsv|sort|uniq -c|sed -E 's/^ *//; s/ /\t/' > t1sct_*.tsv

In a jupyter notebook (ill_tag1_bins) , the grouped barcodes from all the bins were merged into one file, which was then used as the input file of the further analysis in the python script.

Fitness score quantification

The files containing the abundance info for all barcodes in all bins was loaded in a jupyter notebook, processing the data for the generation of a fitness score for all mutants. Raw read numbers were transformed into numbers of sorted cells (Cv) by applying following equation, where $(Ctot)$ is the total number of sorted cells per bin, (Rv) is the number of reads per variant and $(Rtot)$ is the total number of reads per bin.

$$Cv = Ctot\left(\frac{Rv}{Rtot}\right) \qquad (1)$$

Then, for every barcode, a fitness value was determined by calculating the expected value of the fluorescent intensity across all gated bins. The total cell number for all barcodes was multiplied by the weights given through the MFI in every bin, and hence a weighted mean was determined.

$$\beta = \frac{\sum_{i=1}^{gates} \omega_i \times \%c_i}{\sum_{i=1}^{gates} \%c_i} \qquad (2)$$

Finally, the stability fitness score was defined as the log2 of the ratio of weighted mean per variant over weighted mean of the wild type.

$$F_{Exp} = log2\left(\frac{\beta_{Var}}{\beta_{WT}}\right) \qquad (3)$$

For the biochemical analysis of the library (**Figure 2**), scores from barcodes coding for the same mutation were averaged and we only considered mutations that were covered by at least 50 cells among all bins.

Monogenic validation of fitness scores

In order to determine the activity score for monogenetic cultures of the variants, yeast competent cells were transformed with the respective plasmids and variants expressed on the surface as described above. Then, cells were stained with the same antibodies and protocol as used in the DMS experiment. The variants were then, along with WT, used in Tyramide labeling reactions as described above. After washing, cells were loaded on the flow cytometer and 10’000 events recorded. Since the DMS sorting was conducted in a ratio based manner in order to normalize for expression level, we

used a similar scoring strategy. We set 4 gates capturing all singlet cells, as it was done in the *en masse* experiments, capturing all non expressions and nonactive cells in one gate and dividing the remaining cells in three populations. The gates were set on the wild type population to closely match the percentages obtained for wild type sequences in the DMS experiment and left constant for the whole experiment (See **Figure S4**). For each variant, the percentage of cells falling in each gate was multiplied with the mean MFI for the same gate over all variants. We determined a weighted mean for every variant by using the percentage of cells falling in each gate as weight, similarly to the procedure used in generating the high-throughput DMS score (Equation 2, where MFI corresponds to the mean of the per gate overall variants). To set the calculated score in relation with the wild type, we report the final score as the log2 values of the ratio (Equation 3).

Pathways analysis with eMap and EHpath

We used the eMap server to find potential electron hopping pathways. For that we loaded the Alpha fold structures and used default settings for distance options (Center of mass), Surface definition (Residue depth) and amino acids taken into account (Tyrosine & Tryptophan)[51].

For further analysis of such antioxidant rescue pathways, we used the "EHPath" python script, estimating mean residence time for available pathways, which is defined as the inverse of the transfer rate. Hence do low numbers describe fast pathways, whereas bigger values amount to faster hopping rates.
The script is fed with coordinates for an electron donor (here the heme), an electron acceptor (defined surface accessible residue) and bridging amino acid residues which are extracted by the following python command, as described in the reference[52].

```
grep -E 'CYS|MET|TRP|TYR' POI.pdb > POI_bridging.pdb
```

The value for the requested cut off number of maximum nodes was set to 4, indicating that each pathway consists of maximal 5 nodes. We extracted residence times for the 5 quickest pathways and set the alpha value to 1, whereas alpha is defined as

$\boldsymbol{\lambda}_{adj} = 0.8 + \alpha(\boldsymbol{\lambda} - 0.8)$ (ineV)

yielding $\boldsymbol{\lambda}_{adj} = \boldsymbol{\lambda}$, with $\boldsymbol{\lambda}$ being the reorganization energy and $\boldsymbol{\lambda}_{adj}$ the adjusted reorganization energy.

Substrate docking with Swissdock Autodock VINA

The substrate docking experiments were conducted using the Swissdock Autodock VINA engines (https://www.swissdock.ch/) [58,59]. We submitted the tyramide fluor AF594 substrate as SMILES string:
CN1C2=C(C=C3C(OC4=CC5=[N+](C)C(C)(C)C=C(CS(O)(=O)=O)C5=CC4=C3C3=C(C=CC(=C3)C(=O)NCCC3=CC=C(O)C=C3)C([O-])=O)=C2)C(CS(O)(=O)=O)=CC1(C)C
and uploaded the pdb files.

We set the search box around the heme access site and set sampling exhaustivity to 10. The best ranked pose was extracted along with the calculated affinity.
For the docking studies with attracting cavities, also provided by Swisdock, we set the search space around the whole protein as a 70*70*70 Angstrom cube and set the RIC (Random initial conditions) to 5.

Kinetics Michaelis menten analysis

Catalytic constants for RB19 and Guaiacol were determined using a TECAN UV-VIS spectrophotometer plate reader. Reaction mixtures were set up in 96 well plates and contained 1.3 uM Mb, 1 mM H_2O_2 and the reducing substrate in a 50 mM sodium phosphate buffer (pH 6.1). Rb19 was titrated from 5-400 uM and Guaiacol from 0.01-6 mM. Then the reactions were tracked spectrophotometrically (Decrease in absorbance at 595 nm for RB19 or increase at 470 nm for Guaiacol. The slopes were transformed in substrate turnover by means of calibration curves and the initial velocities extracted. All reactions were completed in triplicates. The reaction velocities were fitted with a Michaelis-Menten curve in Graphpad Prism.

Michaelis-Menten analysis with tyramide decoupled assay

In order to measure michaelis menten kinetics for the actual screening substrate tyramide AF594, we design an experiment assaying the endpoint fluorescence of uninduced yeast cells containing an empty cassette. For that we prepared 96 deep well plates containing 0.27 uM (final concentration) purified soluble myoglobin in 50 mM sodium phosphate buffer (pH 6.1). Then we added the reaction mix containing the cells (20 mio/mL), 1 mM (final) H_2O_2 and the tyramide substrate titrated from 1 uL to 35 uL /ul. These volumina were taken from a stock of 1:100 diluted tyramide stock. The acquired stock contains 150 “slides” dissolved in 150 ul DMSO. The exact concentration cannot be stated since it is unknown. While in the DMS experiment and the decoupled cell labelling validation in **Figure 4C** the hydrogen peroxide was administered by GOx/glucose system to minimize oxidative stress for yeast cell survival, here we directly added it with the intention to lower complexity for the time sensitive experiment.
The 96 well plate was kept at 37°C with shaking at 850 rpm. Then, reactions were stopped by diluting the reaction mix adding 400 uL of PBS with 0.1 % BSA and 0.05% Tween20 supplemented with 10 mM ascorbate. The ascorbate acts as a highly concentrated reducing agent and together with the dilution effect stops the reactions reliably, as has been assayed in separate experiments. Time points were 1, 3 , 5 and 10 minutes. Then the cells were washed with 400 uL of the same PBS buffer and finally resuspended in 200 uL. Then the cell's fluorescence was assayed by means of an Attune NxT flow cytometer equipped with a plate autosampler. The median fluorescence intensity of singlet cell population was exported and used to fit the reaction speed for every concentration. Linear regression analysis was performed using GraphPad Prism software.

Machine learning models of variant activity

Dataset and Preprocessing. The initial library contained 6,115 variants with up to five mutations relative to the wild-type sequence. Variants with four or five mutations were excluded from further analysis. For the remaining variants, fitness scores for expression-normalized activity were obtained from two biological replicates. A filtering step was applied, removing variants that showed an absolute difference greater than 0.1 between their replicate fitness scores to ensure data robustness. The final fitness score for each variant was calculated as the arithmetic mean of its two replicate scores. This preprocessing resulted in a dataset of 4,769 unique variants, which were used for model development.

Sequence embeddings. Protein sequence embeddings were generated using two pre-trained models: ESM3 "Open-Small" and ProtTrans "T5-XL". For ESM3, each sequence was converted into a 1536-dimensional embedding vector, and for ProtTrans T5-XL, into a 1024-dimensional vector. In both cases, embeddings were obtained by taking the mean of the final hidden layer representations across all amino acid positions.

Model architectures and baselines. Several machine learning approaches were benchmarked for predicting variant fitness from sequence representations. These included Random Forest Regressor (RF) and Support Vector Regressor (SVR) using both one-hot encoded sequences and pre-trained protein language model embeddings (ProtTrans and ESM3). A Convolutional Neural Network (CNN) was also evaluated with one-hot encoded sequences (Figure 3C, left). The Multi-Layer Perceptron (MLP) architecture, when paired with ProtTrans embeddings or ESM3 embeddings, yielded the highest initial performance and was selected for further development. After comparing the performance of both embeddings on unseen variants (Figure 3C, right and Supplementary Figure S6), we selected the ESM3 model for the screening of unseen variants. The final MLP architecture consisted of an input layer with 1536 units for ESM3, followed by three hidden layers with 512, 256, and 128 units, respectively. Rectified Linear Unit (ReLU) was used as the activation function after each hidden layer. To prevent overfitting and stabilize training, Batch Normalization was applied after each hidden linear layer, and dropout with a probability of 0.6 was applied after each activation.

Model training and evaluation. Models were trained using the Pytorch library. The dataset of 4,769 variants was initially divided into a training set (80%) and a fixed held-out test set (20%) using a stratified continuous split based on the distribution of mean fitness scores. A 5-fold stratified cross-validation (CV) procedure was employed on the 80% training data to train and evaluate the ESM3 MLP model. To assess model robustness and account for variations due to random weight initialization and data shuffling, the entire script, including data splitting and 5-fold CV training, was executed four times, each with a distinct random seed. This resulted in a total of 20 trained MLP models (4 seeds × 5 folds).

For each of the 20 training instances, the MLP model was trained for 500 epochs using the AdamW optimizer (learning rate = 0.0001, L2 weight decay = 1e-5) with Mean Squared Error (MSE) as the loss function. A batch size of 32 was used, and gradient norms were clipped at a maximum value of 1.0. Hyperparameters such as learning rate, optimizer choice (AdamW vs. Adam, RAdam), dropout probability (0.2-0.7), activation functions (ReLU, LeakyReLU, SiLU), and the inclusion of gradient clipping were determined through preliminary experiments using ProtTrans embeddings before finalizing the ESM3 MLP training configuration. During each epoch of a CV fold, model performance

on the corresponding validation set was assessed using the coefficient of determination (R^2) between ground truth and predicted fitness. The model state (weights) from the epoch that yielded the highest validation R^2 score was saved as the best model for that specific fold and seed. This best model was subsequently evaluated on the fixed 20% test set (Figure 3, left).

Prediction on Double Mutant Library and Candidate Selection. An in silico library of all possible double amino acid substitutions was generated from the wild-type protein sequence. After removing variants already present in the training dataset, this library contained about 4.25 million unique double mutant sequences. ESM3 embeddings were generated for this entire library. Each of the 20 trained MLP models (from the 4 seeds × 5 folds) was then utilized to predict the fitness scores for all ~4.25 million double mutants.

To identify a set of variants with robust fitness predictions for experimental testing, we selected the top 10,000 sequences predicted by each model and identified 65 consensus sequences shared by all of the 20 prediction sets, which we deemed to be variants with high-confidence fitness predictions. From this set, 20 variants were selected for experimental validation (Figure 3F).

References

1. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature [Internet] 181:662–666. Available from: https://www.nature.com/articles/181662a0
2. Wittenberg JB (1970) Myoglobin-facilitated oxygen diffusion: role of myoglobin in oxygen entry into muscle. Physiol Rev 50:559–636.
3. Wan L, Twitchett MB, Eltis LD, Mauk AG, Smith M (1998) In vitro Evolution of Horse Heart Myoglobin to Increase Peroxidase Activity. Proceedings of the National Academy of Sciences of the United States of America [Internet] 95:12825–12831. Available from: https://www.jstor.org/stable/46153
4. Guo C, Chadwick RJ, Foulis A, Bedendi G, Lubskyy A, Rodriguez KJ, Pellizzoni MM, Milton RD, Beveridge R, Bruns N (2022) Peroxidase Activity of Myoglobin Variants Reconstituted with Artificial Cofactors. ChemBioChem [Internet] 23:e202200197. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cbic.202200197
5. Boutaud O, Roberts LJ (2011) Mechanism-Based Therapeutic Approaches to Rhabdomyolysis-Induced Renal Failure. Free Radic Biol Med [Internet] 51:1062–1067. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116013/
6. Holt S, Moore K (2000) Pathogenesis of Renal Failure in Rhabdomyolysis: The Role of Myoglobin. Experimental Nephrology [Internet] 8:72–76. Available from: https://doi.org/10.1159/000020651
7. Moore KP, Holt SG, Patel RP, Svistunenko DA, Zackert W, Goodier D, Reeder BJ, Clozel M, Anand R, Cooper CE, et al. (1998) A Causative Role for Redox Cycling of Myoglobin and Its Inhibition by Alkalinization in the Pathogenesis and Treatment of Rhabdomyolysis-induced Renal Failure *. Journal of Biological Chemistry [Internet] 273:31731–31737. Available from: https://www.jbc.org/article/S0021-9258(19)59006-8/abstract
8. Alayash AI, Patel RP, Cashon RE (2001) Redox reactions of hemoglobin and myoglobin: biological and toxicological implications. Antioxid Redox Signal 3:313–327.
9. Wilson MT, Reeder BJ (2021) The peroxidatic activities of Myoglobin and Hemoglobin, their pathological consequences and possible medical interventions. Mol Aspects Med:101045.
10. Vlasova I (2018) Peroxidase Activity of Human Hemoproteins: Keeping the Fire under Control. Molecules [Internet] 23:2561. Available from: http://www.mdpi.com/1420-3049/23/10/2561
11. Reeder BJ, Sharpe MA, Kay AD, Kerr M, Moore K, Wilson MT (2002) Toxicity of myoglobin and haemoglobin: oxidative stress in patients with rhabdomyolysis and subarachnoid haemorrhage. Biochemical Society Transactions [Internet] 30:745–748. Available from: https://doi.org/10.1042/bst0300745
12. Li L-L, Yuan H, Liao F, He B, Gao S-Q, Wen G-B, Tan X, Lin Y-W (2017) Rational design of artificial dye-decolorizing peroxidases using myoglobin by engineering Tyr/Trp in the heme center. Dalton Trans. [Internet] 46:11230–11238. Available from: https://pubs.rsc.org/en/content/articlelanding/2017/dt/c7dt02302b
13. Guo W-J, Xu J-K, Wu S-T, Gao S-Q, Wen G-B, Tan X, Lin Y-W (2022) Design and Engineering of an Efficient Peroxidase Using Myoglobin for Dye Decolorization and Lignin Bioconversion. International Journal of Molecular Sciences [Internet] 23:413. Available from: https://www.mdpi.com/1422-0067/23/1/413
14. Wu G-R, Sun L-J, Xu J-K, Gao S-Q, Tan X-S, Lin Y-W (2022) Efficient Degradation of Tetracycline Antibiotics by Engineered Myoglobin with High Peroxidase Activity. Molecules 27:8660.
15. Reeder BJ, Svistunenko DA, Cooper CE, Wilson MT (2012) Engineering Tyrosine-Based Electron Flow Pathways in Proteins: The Case of Aplysia Myoglobin. J. Am. Chem. Soc. [Internet] 134:7741–7749. Available from: https://doi.org/10.1021/ja211745g
16. Pott M, Hayashi T, Mori T, Mittl PRE, Green AP, Hilvert D (2018) A Noncanonical Proximal Heme Ligand Affords an Efficient Peroxidase in a Globin Fold. J. Am. Chem. Soc. [Internet] 140:1535–1543. Available from: https://doi.org/10.1021/jacs.7b12621

17. Ayuso-Fernández I, Emrich-Mills TZ, Haak J, Golten O, Hall KR, Schwaiger L, Moe TS, Stepnov AA, Ludwig R, Cutsail III GE, et al. (2024) Mutational dissection of a hole hopping route in a lytic polysaccharide monooxygenase (LPMO). Nat Commun [Internet] 15:3975. Available from: https://www.nature.com/articles/s41467-024-48245-w
18. Moosmann B (2021) Redox Biochemistry of the Genetic Code. Trends in Biochemical Sciences [Internet] 46:83–86. Available from: https://www.sciencedirect.com/science/article/pii/S0968000420302711
19. Granold M, Hajieva P, Toşa MI, Irimie F-D, Moosmann B (2018) Modern diversification of the amino acid repertoire driven by oxygen. Proceedings of the National Academy of Sciences [Internet] 115:41–46. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1717100115
20. Ravanfar R, Sheng Y, Gray HB, Winkler JR (2023) Tryptophan extends the life of cytochrome P450. Proceedings of the National Academy of Sciences [Internet] 120:e2317372120. Available from: https://www.pnas.org/doi/10.1073/pnas.2317372120
21. Gray HB, Winkler JR (2015) Hole hopping through tyrosine/tryptophan chains protects proteins from oxidative damage. Proceedings of the National Academy of Sciences [Internet] 112:10920–10925. Available from: https://www.pnas.org/doi/10.1073/pnas.1512704112
22. Winkler JR, Gray HB (2015) Could tyrosine and tryptophan serve multiple roles in biological redox processes? Philos Trans A Math Phys Eng Sci [Internet] 373:20140178. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4342971/
23. B. Gray H, R. Winkler J (2021) Functional and protective hole hopping in metalloenzymes. Chemical Science [Internet] 12:13988–14003. Available from: https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc04286f
24. Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. Nat Methods [Internet] 11:801–807. Available from: https://www.nature.com/articles/nmeth.3027
25. Küng C, Protsenko O, Vanella R, Nash MA (2025) Deep mutational scanning reveals a de novo disulfide bond and combinatorial mutations for engineering thermostable myoglobin. Protein Science [Internet] 34:e70112. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.70112
26. Vanella R, Küng C, Schoepfer AA, Doffini V, Ren J, Nash MA (2024) Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. Nat Commun [Internet] 15:1807. Available from: https://www.nature.com/articles/s41467-024-45630-3
27. Höllerer S, Desczyk C, Muro RF, Jeschek M (2024) From sequence to function and back – High-throughput sequence-function mapping in synthetic biology. Current Opinion in Systems Biology [Internet] 37:100499. Available from: https://www.sciencedirect.com/science/article/pii/S2452310023000562
28. Agresti JJ, Antipov E, Abate AR, Ahn K, Rowat AC, Baret J-C, Marquez M, Klibanov AM, Griffiths AD, Weitz DA (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. Proceedings of the National Academy of Sciences [Internet] 107:4004–4009. Available from: https://www.pnas.org/doi/full/10.1073/pnas.0910781107
29. Romero PA, Tran TM, Abate AR (2015) Dissecting enzyme function with microfluidic-based deep mutational scanning. Proceedings of the National Academy of Sciences [Internet] 112:7159–7164. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1422285112
30. Thomas N, Belanger D, Xu C, Lee H, Hirano K, Iwai K, Polic V, Nyberg KD, Hoff KG, Frenz L, et al. (2025) Engineering highly active nuclease enzymes with machine learning and high-throughput screening. Cell Systems [Internet] 16:101236. Available from: https://www.sciencedirect.com/science/article/pii/S2405471225000699
31. Stiffler MA, Hekstra DR, Ranganathan R (2015) Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. Cell [Internet] 160:882–892. Available from: https://www.sciencedirect.com/science/article/pii/S0092867415000781
32. Trinidad DD, Macdonald CB, Rosenberg OS, Fraser JS, Coyote-Maestas W (2024) Deep mutational scanning of EccD3 reveals the molecular basis of its essentiality in the mycobacterium ESX secretion

system. bioRxiv [Internet]:2024.08.23.609456. Available from:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11370616/
33. Jansen SC, Mayer C (2024) A Robust Growth-Based Selection Platform to Evolve an Enzyme via Dependency on Noncanonical Tyrosine Analogues. JACS Au [Internet] 4:1583–1590. Available from:
https://pubs.acs.org/doi/10.1021/jacsau.4c00070
34. Küng C, Vanella R, Nash MA (2023) Directed evolution of Rhodotorula gracilisD-amino acid oxidase using single-cell hydrogel encapsulation and ultrahigh-throughput screening. React. Chem. Eng. [Internet] 8:1960–1968. Available from:
https://pubs.rsc.org/en/content/articlelanding/2023/re/d3re00002h
35. Vanella R, Boult S, Kueng C, Nash M (2025) Decoding Substrate Specificity in a Promiscuous Biocatalyst by Enzyme Proximity Sequencing. :2025.07.10.664162. Available from:
https://www.biorxiv.org/content/10.1101/2025.07.10.664162v1
36. Hsu C, Nisonoff H, Fannjiang C, Listgarten J (2022) Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol [Internet] 40:1114–1122. Available from:
https://www.nature.com/articles/s41587-021-01146-5
37. Yang KK, Wu Z, Arnold FH (2019) Machine-learning-guided directed evolution for protein engineering. Nat Methods [Internet] 16:687–694. Available from:
https://www.nature.com/articles/s41592-019-0496-6
38. Frei L, Gao B, Han J, Taft JM, Irvine EB, Weber CR, Kumar RK, Eisinger BN, Ignatov A, Yang Z, et al. (2025) Deep mutational learning for the selection of therapeutic antibodies resistant to the evolution of Omicron variants of SARS-CoV-2. Nat. Biomed. Eng [Internet] 9:552–565. Available from:
https://www.nature.com/articles/s41551-025-01353-4
39. Taft JM, Weber CR, Gao B, Ehling RA, Han J, Frei L, Metcalfe SW, Overath MD, Yermanos A, Kelton W, et al. (2022) Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor-binding domain. Cell [Internet] 185:4008-4022.e14. Available from: https://www.sciencedirect.com/science/article/pii/S0092867422011199
40. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA (2016) Plasmid-based one-pot saturation mutagenesis. Nat Methods [Internet] 13:928–930. Available from:
http://www.nature.com/articles/nmeth.4029
41. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, et al. (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell [Internet] 182:1295-1310.e20. Available from: https://www.sciencedirect.com/science/article/pii/S0092867420310035
42. Li Y, Arcos S, Sabsay KR, te Velthuis AJW, Lauring AS (2023) Deep mutational scanning reveals the functional constraints and evolutionary potential of the influenza A virus PB1 protein. Journal of Virology [Internet] 97:e01329-23. Available from:
https://journals.asm.org/doi/full/10.1128/jvi.01329-23
43. Picotti P, Marabotti A, Negro A, Musi V, Spolaore B, Zambonin M, Fontana A (2004) Modulation of the structural integrity of helix F in apomyoglobin by single amino acid replacements. Protein Science [Internet] 13:1572–1585. Available from:
https://onlinelibrary.wiley.com/doi/abs/10.1110/ps.04635304
44. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al. (2022) ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence [Internet] 44:7112–7127. Available from: https://ieeexplore.ieee.org/document/9477085
45. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, et al. (2025) Simulating 500 million years of evolution with a language model. Science [Internet] 387:850–858. Available from: https://www.science.org/doi/10.1126/science.ads0018
46. Meng S, Li Z, Ji Y, Ruff AJ, Liu L, Davari MD, Schwaneberg U (2023) Introduction of aromatic amino acids in electron transfer pathways yielded improved catalytic performance of cytochrome P450s. Chinese Journal of Catalysis [Internet] 49:81–90. Available from:

https://www.sciencedirect.com/science/article/pii/S1872206723644456
47. Sáez-Jiménez V, Rencoret J, Rodríguez-Carvajal MA, Gutiérrez A, Ruiz-Dueñas FJ, Martínez AT (2016) Role of surface tryptophan for peroxidase oxidation of nonphenolic lignin. Biotechnology for Biofuels [Internet] 9:198. Available from: https://doi.org/10.1186/s13068-016-0615-x
48. Li L, Wang T, Chen T, Huang W, Zhang Y, Jia R, He C (2021) Revealing two important tryptophan residues with completely different roles in a dye-decolorizing peroxidase from Irpex lacteus F17. Biotechnol Biofuels [Internet] 14:128. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8165797/
49. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature [Internet] 596:583–589. Available from: https://www.nature.com/articles/s41586-021-03819-2
50. Bobrow MN, Harris TD, Shaughnessy KJ, Litt GJ (1989) Catalyzed reporter deposition, a novel method of signal amplification application to immunoassays. Journal of Immunological Methods [Internet] 125:279–285. Available from: https://www.sciencedirect.com/science/article/pii/002217598990104X
51. Anon eMap: A Web Application for Identifying and Visualizing Electron or Hole Hopping Pathways in Proteins | The Journal of Physical Chemistry B. Available from: https://pubs.acs.org/doi/10.1021/acs.jpcb.9b04816
52. Teo RD, Wang R, Smithwick ER, Migliore A, Therien MJ, Beratan DN (2019) Mapping hole hopping escape routes in proteins. Proceedings of the National Academy of Sciences [Internet] 116:15811–15816. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1906394116
53. Silkstone GGA, Silkstone RS, Wilson MT, Simons M, Bülow L, Kallberg K, Ratanasopa K, Ronda L, Mozzarelli A, Reeder BJ, et al. (2016) Engineering tyrosine electron transfer pathways decreases oxidative toxicity in hemoglobin: implications for blood substitute design. Biochem J [Internet] 473:3371–3383. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5095908/
54. Cooper CE, Simons M, Dyson A, Leiva Eriksson N, Silkstone GGA, Syrett N, Allen-Baume V, Bülow L, Ronda L, Mozzarelli A, et al. (2024) Taming hemoglobin chemistry—a new hemoglobin-based oxygen carrier engineered with both decreased rates of nitric oxide scavenging and lipid oxidation. Exp Mol Med [Internet] 56:2260–2270. Available from: https://www.nature.com/articles/s12276-024-01323-x
55. Gietz RD, Woods RA (2002) Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol 350:87–96.
56. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences Birol I, editor. Bioinformatics [Internet] 34:3094–3100. Available from: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778
57. Bushnell B BBMap: A Fast, Accurate, Splice-Aware Aligner. In: ; 2014. Available from: https://www.semanticscholar.org/paper/BBMap%3A-A-Fast%2C-Accurate%2C-Splice-Aware-Aligner-Bushnell/f64dd54444a724574deb7710888091350eebb2b9
58. Bugnon M, Röhrig UF, Goullieux M, Perez MAS, Daina A, Michielin O, Zoete V (2024) SwissDock 2024: major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. Nucleic Acids Research [Internet] 52:W324–W332. Available from: https://doi.org/10.1093/nar/gkae300
59. Eberhardt J, Santos-Martins D, Tillack AF, Forli S (2021) AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J. Chem. Inf. Model. [Internet] 61:3891–3898. Available from: https://doi.org/10.1021/acs.jcim.1c00203