Deep Mutational Scanning Reveals a De Novo Disulfide Bond and Combinatorial Mutations for Engineering Thermostable Myoglobin

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, 4056 Basel, Switzerland
* Correspondence to: michael.nash@unibas.ch

Abstract

Engineering protein stability is a critical challenge in biotechnology. Here, we used massively parallel deep mutational scanning (DMS) to comprehensively explore the mutational stability landscape of human myoglobin (hMb) and identify key mutations that enhance stability. Our DMS approach involved screening over 10,000 hMb variants by yeast surface display, single-cell sorting and high-throughput DNA sequencing. We show how surface display levels serve as a proxy for thermostability of soluble hMb variants, and report strong correlations between DMS-derived display levels and top-performing machine learning stability prediction algorithms. This approach led to the discovery of a variant with a *de novo* disulfide bond between residues R32C and C111, which increased thermostability by >12 °C compared to wild-type hMb. By combining single stabilizing mutations with R32C, we engineered combinatorial variants that exhibited predominantly additive effects on stability with minimal epistasis. The most stable combinatorial variant exhibited a denaturation temperature exceeding 89 °C, representing a >17 °C improvement over wild-type hMb. Our findings demonstrate the capabilities in DMS-assisted combinatorial protein engineering to guide the discovery of thermostable variants, and highlight the potential of massively parallel mutational analysis for the development of proteins for industrial and biomedical applications.

Keywords

Protein engineering; deep mutational scanning; high-throughput screening; protein stability; disulfide bond; yeast display

Introduction

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 [1–4], however, success rates are low and the discovery process is stochastic. Especially in campaigns for improving enzyme activity, limits are imposed by the natural trade-offs between stability and phenotypic activity [5]. This universal phenomenon often leads to the destabilization of proteins when introducing mutations that improve function. A strategy to overcome or attenuate this effect is to start with a highly stable variant that can tolerate destabilizing but functionally beneficial mutations.

It was shown, for example, through lattice simulations and experiments that highly stable sequences are more amenable to *in vitro* evolution and can achieve higher levels of functional improvement upon mutation[6]. This suggests a generalizable strategy for protein engineering involving the design of superior variant libraries that are biased to preferentially include stabilizing mutations[7]. In the past, approaches based on random mutagenesis[8] or ancestral sequence reconstruction[9] have been used for improving the thermostability of enzymes. A more recent approach used large-scale mutational data to create mutability landscapes and identify regional trends in mutational tolerance[10,11]. Such an approach allows to systematically screen mutational neighborhoods to find hotspots of stabilizing mutations, as well as sequence stretches where mutations are detrimental to stability. In this context, massively parallelized methods for assaying thousands of variants in parallel as is typically done in mutational scanning (DMS) experiments can unlock valuable information for informed library design[12,13].

In this work we use a DMS workflow to study human myoglobin (hMb), a 17 kDa globular heme protein that serves as an ideal model for our study highlighting structural features important for stability. In addition to its native function as an oxygen binding protein, hMb has been investigated as an engineering scaffold for the introduction of new catalytic activity. For instance, Hilvert and colleagues used directed evolution to improve the peroxidase activity of myoglobin[14]. Other campaigns transformed hMb into an artificial metalloenzyme capable of performing atom transfer radical cyclisation as well as keto reductase catalysis[15,16].

Here, we quantified expression levels of yeast surface displayed holo hMb variants by DMS. We show how expression levels can serve as a proxy for thermostability and reveal mutational effects on hMb stability in a massively parallel workflow. We note that in literature there is extensive evidence that yeast surface display expression level is generally correlated with protein folding stability[5,17–19]. Thermodynamically destabilizing mutations yield lower detectable expression levels in yeast surface display 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, and further validated this experimentally.

We found strong correlations between our DMS dataset and hMb evolutionary conservation scores, as well as stability prediction algorithms. This approach enabled discovery of highly stabilizing mutations in hMb, including a *de novo* disulfide bonded variant. By combining single mutants, we were able to further enhance stability, observing primarily additive effects with low epistasis. This method paves the way for more efficient protein engineering of hMb catalytic activity by providing a stable scaffold for further optimization.

Results

DMS data generation and validation

To generate the DMS dataset, 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 2,578 single amino acid substitutions (80% of all possible), 2,769 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, **Tables S1-5**).

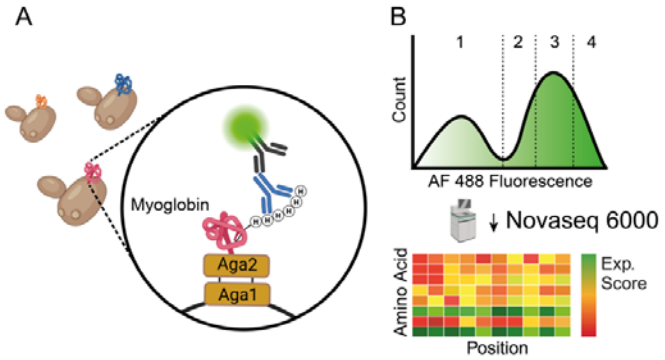

Figure 1. Scheme of the experimental and computational workflow. **A)** A myoglobin variant library is displayed on the surface of yeast cells and fluorescently labeled based on expression level of the hMb variant via C-terminal His_6-Tag. **B)** The labeled library is sorted into four bins with known median fluorescence intensities. After sequencing all variants in each bin by NGS, we convert read counts into expression fitness scores to obtain the fitness landscape.

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 fitness values for synonymous codon substitutions.

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 = 0.92, 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 hMb variants in parallel.

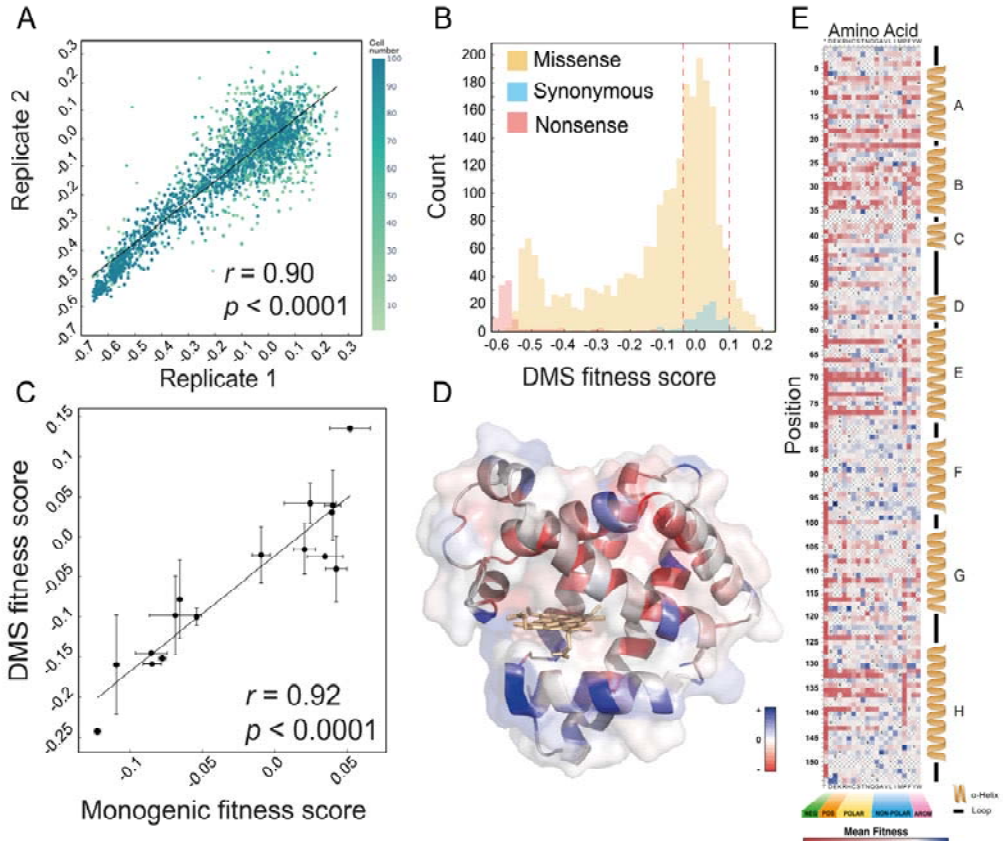

Figure 2. DMS of yeast displayed human myoglobin. **A)** Linear correlation of fitness scores from two biological replicates. The color gradient represents the number of cells carrying a particular hMb variant found in both replicates. **B)** Distribution of fitness scores for nonsense (red), missense (orange) and synonymous single mutations (blue). Red dashed lines indicate the width of the fitness range for synonymous mutations. **C)** Correlation of expression fitness scores determined by 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)** hMb 3D structure (PDB: 3RGK) colored by mean fitness score per position. **E)** Heatmap showing expression fitness scores of single substitution 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 per position average fitness score on the 3D structure of hMb (**Figure 2D**) and the single mutant fitness values in the form of a heatmap (**Figure 2E**). In the heatmap we noticed conspicuous dark red 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 or 3D structure, 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 S1**). 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 2E**), 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[23]. 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 S2**). Apart from stop codons, the strongest negative impact was observed for mutations to proline (Avg. score: -163%, n = 122), a well known helix-breaking substitution[24]. Charged amino acids (D, E, K, R and H) were found to severely negatively influence fitness scores (Avg. score: -38%, 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: +63%, n = 639). The aromatic hydrophobic side chain of phenylalanine also had higher than average fitness (Avg. score: +42%, n = 116). Many of the observed effects have been discussed in a meta-study worth mentioning analyzing the roles of amino acids among different proteins[25,26].

Discovery of R32C disulfide bonded myoglobin with enhanced thermostability

We found that cysteine substitutions were less detrimental to expression on average (Avg. score: +30%, 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 investigate 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 3A**). 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 again under reducing and non-reducing conditions (**Figure 3C**). The increase in detected mass of 2 Daltons for the reduced sample corresponding to the two hydrogen atoms 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 S3**). Disulfide bond engineering as well as cross-linking of non-canonical amino acids have both been reported for modulating myoglobin properties[27,28], however, to our knowledge the R32C myoglobin variant discovered by our DMS screen has not been described in literature.

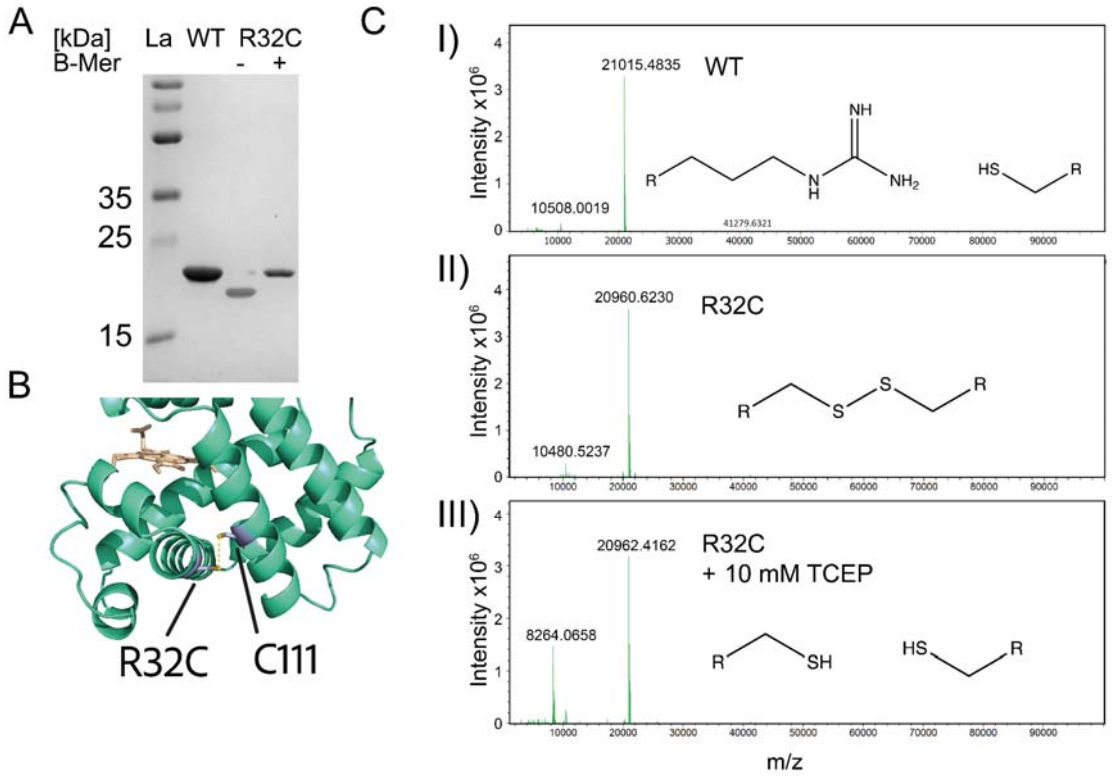

Figure 3. A) 15% SDS Page gels of purified proteins for WT and R32C. Lane 1) reference ladder; lane 2) WT hMb; lanes 3 & 4) R32C hMb variant boiled without (-) or with (+) 5% (vol./vol.) beta-mercaptoethanol reducing agent, respectively. The lower apparent molecular weight for the non-reduced sample is consistent with an intramolecular disulfide bond. **B)** Residues involved in disulfide bond formation, as mapped on protein structure (PDB: 3RGK) Cysteines are colored blue and thiols are shown in yellow. Heme cofactor is shown in wheat color. **C)** Mass spectrometry results. Deconvoluted peaks for purified WT **(I)** and R32C **(II)** protein variants. **III)** Same sample as B), incubated with 10 mM TCEP prior to desalting and MS analysis. Reduction of disulfide bond yielded a mass increase of ~2 Daltons.

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 hence excluding paralogues) 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 S4 A**, (r = -0.92, p <0.0001), indicating more conserved positions were less tolerant for mutation on average and resulted in lower expression levels. 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 skeletal and cardiac 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 positive fitness in the case of position 99. These sites contributed to substantial variability of the negative trend (**Figure S4 A**). These data highlight the ability of DMS to shed light onto activity-stability trade-offs[5].

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 S4 B**) with a prominent outlier being Asp21 ($SASA_{Score}$ = 47), which joins helix A and B (**Figure S4 C**). We found this position does not accept any other amino acid, including negatively charged glutamate as can well be observed in **Figure 2E**. 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[32]. Comparing the predicted values to the DMS expression fitness scores, we found strong negative correlation when averaged per position (**Figure 4A**, r = -0.56, p < 0.0001) but also for single individual mutations (**Figure 4B**, r = -0.43, p < 0.0001), again validating the experimental setup serving as a proxy for thermodynamic stability. We further performed the same analysis with a more recently developed stability prediction tool (ThermoMPNN)[33]. This deep neural network builds on the ProteinMPNN framework [34] and was trained on a mega-scale dataset[35] to predict change of stability in point mutations. It has been benchmarked against state-of-the-art methods and showed strong performance, especially when predicting stabilizing mutations. Again, we tested averaged and single mutation correlations with predicted $\Delta\Delta G_{folding}$ from ThermoMPNN and obtain Pearson correlation scores even higher than with FoldX (**Figure 4C**, r = -0.74, p < 0.0001) for averaged as well as for singles (**Figure 4D**, r = -0.71, p < 0.0001).

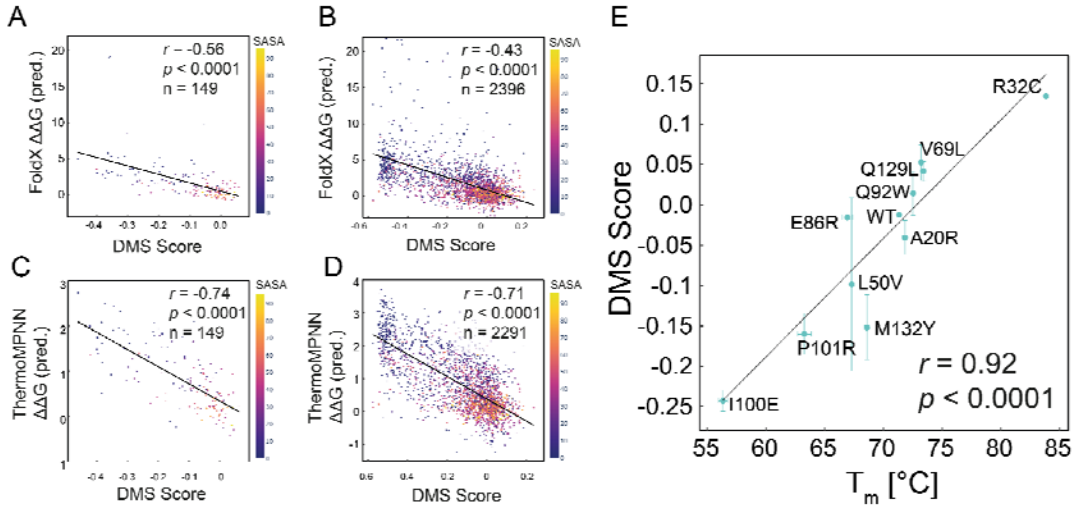

Figure 4. DMS scores correlate with computationally predicted and experimentally measured thermodynamic stability. **A)** Scatterplot of FoldX $\Delta\Delta G_{folding}$ vs. experimental DMS scores, grouped by amino acid residue position. Color coding represents solvent accessible surface area (SASA). **B)** FoldX $\Delta\Delta G_{folding}$ vs. DMS scores for individual variants. **C)** ThermoMPNN $\Delta\Delta G_{folding}$ vs. DMS scores, grouped by amino acid residue position. **D)** ThermoMPNN $\Delta\Delta G_{folding}$ vs DMS scores for individual variants. **E)** 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 nano differential scanning fluorimetry (NanoDSF). We found strong correlation between T_m values measured by NanoDSF and DMS expression fitness scores (**Figure 4C**, $r = 0.92$, $p < 0.0001$).

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. The strong connection between yeast expression level and protein stability has been reported in prior work and is further verified here by correlating DMS expression fitness scores with the thermostability of soluble hMb variants[19,36]. Similarly, it has been shown that in *E. coli*, expression levels of holo myoglobin highly depend on apomyoglobin stability[37].

DMS scores in hMb show low levels of epistasis

Next we investigated double (n = 2,769) and triple (n = 391) mutants present in the library. We processed them alongside the single mutants in the computational pipeline and filtered them for the same quality thresholds. We concluded that since most mutations are detrimental for expression level, the fraction of double mutants with lower expression than WT level would be even higher. In fact, we found that 83.90 % of the double mutants and 90.03 % of the triple mutant variants were expressed at lower levels than wild type (fitness score < 0) (**Figure 5A**). When comparing the DMS fitness scores of double ($F_{1,2}$) and triple ($F_{1,2,3}$) mutants with the sum fitness scores of the individual

single mutants (i.e. $F_1 + F_2$; $F_1 + F_2 + F_3$), we found that they correlated strongly ((**Figure 5B**, $r = 0.87$, $p < 0.0001$) for double mutants and (**Figure 5C**, $r = 0.80$, $p < 0.0001$) for triple mutants). From that observation, we concluded that in a simple protein like hMb, for low level combinations of up to three mutations, we do not observe significant epistasis. For the analysis for plots in **Figure 5**, we filtered the data for a minimal sum of $F_{1,2}$ and triple $F_{1,2,3}$ of -0.62, because our computational pathway, depending on the WT fitness allows for DMS fitness scores only to reach within -0.62 to 0.38. Due to the nature of the log2 operation on min-max normalized values, the data is limited to that range. While not a problem when analyzing single mutants or comparing scores, for the described analysis of linear additivity we need to take this into account, as computational scores $F_{1,2}$ and $F_{1,2,3}$ plateau at -0.62 while the sum of their individual mutations F_1, F_2 and F_3 can result in lower scores. We note that the number of variants in **Figure 5B** and **C** differ from **Figure 5A** because we could only analyze multiple mutants for which we have the corresponding single mutants passing the quality threshold.

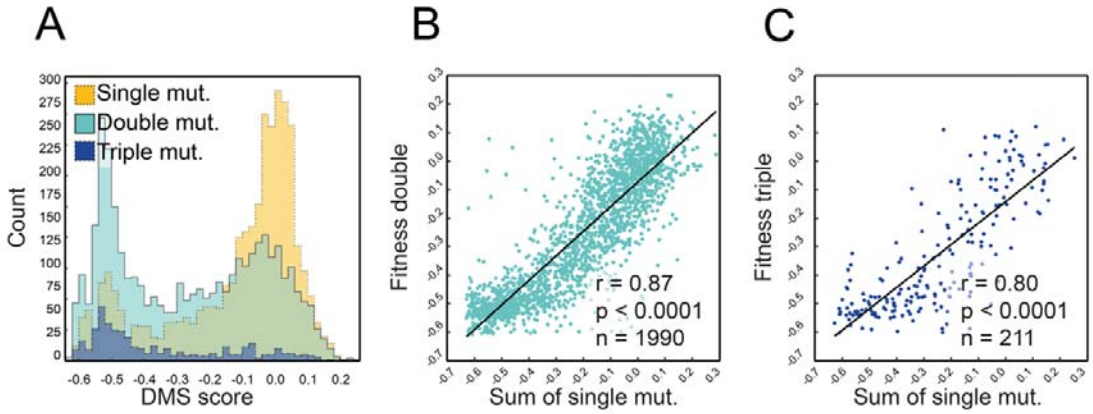

Figure 5. Combinatorial mutant analysis. **A)** Histograms of DMS fitness scores, grouped by number of mutations. Single mutants are shown in yellow, double mutants in cyan and triple mutants in dark blue. **B)** Scatterplot showing DMS scores of double mutants (for $F_{1,2} > -0.63$) vs. sum of corresponding single mutant DMS fitness scores ($F_1 + F_2$) ($r = 0.87$, $p < 0.0001$), indicating low epistasis among the set of double mutants. **C)** Scatterplot showing DMS scores of triple mutants vs. sum of corresponding single mutant DMS fitness scores ($F_1 + F_2 + F_3$) ($r = 0.8$, $p < 0.0001$).

In order to study further the effect of combining mutations, we generated a small library of 15 combinatorial variants derived from the 5 best expressing single mutations in the library (**Table S6**). Along with the R32C disulfide bond mutant, we created different sets of double, triple, quadruple and quintuple mutants, and measured their expression level on the yeast cell surface as monogenic cultures. The featured mutations are listed in **Figure 6A** and mapped on the 3D structure in **Figure 6B**. Endpoint residues shown are not obtained from crystal structures but are simply modeled in Pymol software to provide an overview. Mutant Cβ-carbons are all more than 7 Å away from each other and obtain all a score of over 0.1 in the DMS experiment. As expected, and discussed earlier, all highly stabilizing mutations lie on the protein surface, hence not interfering with the hydrophobic core. Monogenic fitness scores for the combinatorial library members were measured and are presented in **Figure 6C**, alongside the single R32C mutation. The color code from **Figure 6A** is used to indicate which mutations are contained in the combinatorial mutants presented in **Figure 6C** (color heights are not quantitative).

A

Variant	Mutations	DMS score
Mut 1	R32C	0.135
Mut 2	K148S	0.122
Mut 3	V102L	0.121
Mut 4	Q27A	0.118
Mut 5	K43S	0.107

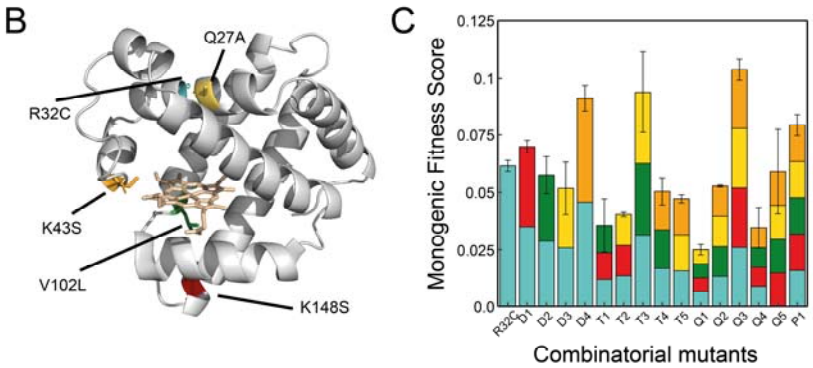

Figure 6. A) Table of highly expressing mutants selected for combinatorial analysis. **B)** 3D structure with selected mutations (PDB: 3RGK). Endpoint residues are displayed as sticks and color-coded according to table in A) **C)** Relative fitness of combinatorial mutants measured in monogenic culture. Different colors symbolize which mutations are contained within the combinatorial variants (sizes of colored sub-bars are not quantitative).

All variants show a higher expression level and hence higher thermostability than WT. Several of the combinatorial variants that we tested exhibited higher stability than the R32C disulfide bond reported above. Amongst the double mutants D1-D4, only D4 (K43S & R32C) presents a significantly higher expression level. The triple and quadruple mutants are generally less well expressed than R32C alone, with the exceptions of T3 and Q3, both of which featured a Q27A substitution. We found that the combination of L102V (green) and K148S (red) were detrimental for stability, with the three lowest expressed variants containing those mutations. Taking into account the 3D structure in **Figure 6B,** one could assume that the close 3D distance between the two exchanged residues in neighboring helices might negatively impact the otherwise stabilizing effects of the mutations individually.

Variant P1 contains a combination of all the chosen mutations, and provides an example of how additivity no longer holds for higher order mutants. The variant Q5 contains all the chosen mutations except the R32C disulfide bond. Interestingly we observed an expression level similar to the R32C mutation alone. This again shows the massive impact on stability of this mutant alone, which was further visualized in a correlation matrix uncovering the individual contributions of each mutation (**Figure S5**). We note that that these mutations have not been tested for their compatibility with the physiological function of molecular oxygen binding.

Next, we investigated the effect of multiple mutations on thermostability of soluble proteins. We purified the best 3 (D4, T3 and Q3) variants as determined by the on-yeast expression measurements, and measured their denaturation temperatures via NanoDSF (**Figure S6**). Based on non-reducing SDS page gels, we observed successful disulfide bond formation and high purity for all variants. We found increased stability with measured T_m values of 84.85, 89.18 and 84.47 °C for D4, T3 and Q3, respectively. Therefore, in accordance with the yeast surface display expression level measurements, we found higher thermostability for all three combinatorial mutants. However, while for variants D4 and Q3, we detected an increase of 13.51 and 13.13 °C over wild type, respectively, we could measure a massive increase of T_m for variant T3, with a melting temperature of more than 17 °C higher than WT.

Compared to the single R32C variant, variants D4 and Q3 are hence only marginally improved. Other recent work from the protein design field presented stability enhancing mutations for hMb using ProteinMPNN[38]. Interestingly, in that work the most stable variant described also features mutation Q27A, along with a mutation of position V102 to an isoleucine, which is closely related to the leucine introduced in our variant T3[39]. However, an important difference is that the ProteinMPNN neural network predicted sequences with ~41-55% sequence identity while in our case only 3 positions in hMb have been mutated. The variants reported here therefore share much higher sequence identity with parental hMb (98% similarity) while achieving comparable stability enhancements of more than 15 °C.

Discussion

In this report, we used comprehensive DMS by yeast surface display to identify stabilizing mutations in hMb. We demonstrate correlation of yeast display expression level with thermostability of hMb variants when expressed and tested as soluble proteins. This finding was further supported through predictive computational folding stability models like FoldX and ThermoMPNN. 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, allowing for discovery of crucial sites for stability. We demonstrated how alpha helix pitch could be measured from mutational stability data, and further discovered a highly stabilized interhelical *de novo* disulfide bonded variant of hMb.

We studied the effects of combining single mutations and showed that mutations contributed to expression and thermostability stability in a mostly additive manner for small numbers of mutations, however these linear trends deviated for higher variants containing more than three mutations. Making use of this knowledge, we were able to design a new-to-nature hMb variant, with significantly improved thermostability (T_m = 89.18°C), representing >17 °C improvement in T_m from wild type. Together, these findings highlight the power of DMS in understanding protein stability and accelerating protein engineering campaigns with broad applications in industry and biomedicine.

Author contributions

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

Supplementary Materials

Supplementary materials are available for this article and include detailed materials and methods, Figures S1-S6, and Table S6.

Competing interest statement

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: https://zenodo.org/records/10658344. Additionally, an interactive colorblind accessible heatmap is provided in the repository.

References

1. Zeymer C, Hilvert D (2018) Directed Evolution of Protein Catalysts. Annu Rev Biochem 87:131–157.

2. Arnold FH (2018) Directed Evolution: Bringing New Chemistry to Life. Angew Chem Int Ed Engl [Internet] 57:4143–4148. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5901037/

3. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol [Internet] 10:866–876. Available from: https://www.nature.com/articles/nrm2805

4. 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

5. 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

6. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. Proceedings of the National Academy of Sciences [Internet] 103:5869–5874. Available from: https://www.pnas.org/doi/full/10.1073/pnas.0510098103

7. Teufl M, Zajc CU, Traxlmayr MW (2022) Engineering Strategies to Overcome the Stability–Function Trade-Off in Proteins. ACS Synth. Biol. [Internet] 11:1030–1039. Available from: https://doi.org/10.1021/acssynbio.1c00512

8. Giver L, Gershenson A, Freskgard P-O, Arnold FH (1998) Directed evolution of a thermostable esterase. Proceedings of the National Academy of Sciences [Internet] 95:12809–12813. Available from: https://www.pnas.org/doi/10.1073/pnas.95.22.12809

9. Cole MF, Gaucher EA (2011) Utilizing natural diversity to evolve protein function: applications towards thermostability. Current Opinion in Chemical Biology [Internet] 15:399–406. Available from: https://www.sciencedirect.com/science/article/pii/S1367593111000408

10. Guo C, Ni Y, Biewenga L, Pijning T, Thunnissen AWH, Poelarends GJ (2021) Using Mutability Landscapes To Guide Enzyme Thermostabilization. Chembiochem [Internet] 22:170–175. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7821111/

11. Frappier V, Keating AE (2021) Data-driven computational protein design. Curr Opin Struct Biol 69:63–69.

12. 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

13. Vanella R, Kovacevic G, Doffini V, Santaella JF de, Nash MA (2022) High-throughput screening, next generation sequencing and machine learning: advanced methods in enzyme engineering. Chem. Commun. [Internet] 58:2455–2467. Available from: https://pubs.rsc.org/en/content/articlelanding/2022/cc/d1cc04635g

14. 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

15. Lubskyy A, Guo C, J. Chadwick R, Petri-Fink A, Bruns N, M. Pellizzoni M (2022) Engineered myoglobin as a catalyst for atom transfer radical cyclisation. Chemical Communications [Internet] 58:10989–10992. Available from: https://pubs.rsc.org/en/content/articlelanding/2022/cc/d2cc03227a

16. Zhang X, Chen D, Stropp J, Tachibana R, Zou Z, Klose D, Ward TR (2024) Repurposing myoglobin into an abiological asymmetric ketoreductase. Chem [Internet] 10:2577–2589. Available from: https://www.cell.com/chem/abstract/S2451-9294(24)00289-4

17. Klesmith JR, Bacik J-P, Wrenbeck EE, Michalczyk R, Whitehead TA (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. Proceedings of the National Academy of Sciences [Internet] 114:2265–2270. Available from: https://www.pnas.org/doi/10.1073/pnas.1614437114

18. Shusta EV, Holler PD, Kieke MC, Kranz DM, Wittrup KD (2000) Directed evolution of a stable scaffold for T-cell receptor engineering. Nat Biotechnol [Internet] 18:754–759. Available from: https://www.nature.com/articles/nbt0700_754

19. 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

20. 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

21. 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

22. Bushnell B BBMap: A Fast, Accurate, Splice-Aware Aligner.

23. Begun A, Molochkov A, Niemi AJ (2019) Protein tertiary structure and the myoglobin phase diagram. Sci Rep [Internet] 9:10819. Available from: https://www.nature.com/articles/s41598-019-47317-y

24. Imai K, Mitaku S (2005) Mechanisms of secondary structure breakers in soluble proteins. Biophysics (Nagoya-shi) 1:55–65.

25. Dunham AS, Beltrao P (2021) Exploring amino acid functions in a deep mutational landscape. Molecular Systems Biology [Internet] 17:e10305. Available from: https://www.embopress.org/doi/full/10.15252/msb.202110305

26. Gray VE, Hause RJ, Fowler DM (2017) Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. Genetics [Internet] 207:53–61. Available from: https://doi.org/10.1534/genetics.117.300064

27. Iannuzzelli JA, Bacik J-P, Moore EJ, Shen Z, Irving EM, Vargas DA, Khare SD, Ando N, Fasan R (2022) Tuning Enzyme Thermostability via Computationally Guided Covalent Stapling and Structural Basis of Enhanced Stabilization. Biochemistry [Internet] 61:1041–1054. Available from: https://doi.org/10.1021/acs.biochem.2c00033

28. Moore EJ, Zorine D, Hansen WA, Khare SD, Fasan R (2017) Enzyme stabilization via computationally guided protein stapling. Proceedings of the National Academy of Sciences [Internet]

114:12472–12477. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1708907114

29. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res [Internet] 44:W344–W350. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987940/

30. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res [Internet] 41:e121. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3695513/

31. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research [Internet] 50:D439–D444. Available from: https://doi.org/10.1093/nar/gkab1061

32. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. Nucleic Acids Research [Internet] 33:W382–W388. Available from: https://doi.org/10.1093/nar/gki387

33. Dieckhaus H, Brocidiacono M, Randolph NZ, Kuhlman B (2024) Transfer learning to leverage larger datasets for improved prediction of protein stability changes. Proceedings of the National Academy of Sciences [Internet] 121:e2314853121. Available from: https://www.pnas.org/doi/full/10.1073/pnas.2314853121

34. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, et al. (2022) Robust deep learning–based protein sequence design using ProteinMPNN. Science [Internet] 378:49–56. Available from: https://www.science.org/doi/10.1126/science.add2187

35. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, Mangan NM, Ovchinnikov S, Rocklin GJ (2023) Mega-scale experimental analysis of protein folding stability in biology and design. Nature [Internet] 620:434–444. Available from: https://www.nature.com/articles/s41586-023-06328-6

36. Hackel BJ, Ackerman ME, Howland SW, Wittrup KD (2010) Stability and CDR Composition Biases Enrich Binder Functionality Landscapes. Journal of Molecular Biology [Internet] 401:84–96. Available from: https://www.sciencedirect.com/science/article/pii/S0022283610006042

37. Samuel PP, Smith LP, Phillips GN, Olson JS (2015) Apoglobin Stability Is the Major Factor Governing both Cell-free and in Vivo Expression of Holomyoglobin *♦. Journal of Biological Chemistry [Internet] 290:23479–23495. Available from: https://www.jbc.org/article/S0021-9258(20)44692-7/abstract

38. Sumida KH, Núñez-Franco R, Kalvet I, Pellock SJ, Wicky BIM, Milles LF, Dauparas J, Wang J, Kipnis Y, Jameson N, et al. (2024) Improving Protein Expression, Stability, and Function with ProteinMPNN. J. Am. Chem. Soc. [Internet] 146:2054–2061. Available from: https://doi.org/10.1021/jacs.3c10941

39. Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. Science [Internet] 185:862–864. Available from: https://www.science.org/doi/10.1126/science.185.4154.862