JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Improving Protein Expression, Stability, and Function with ProteinMPNN

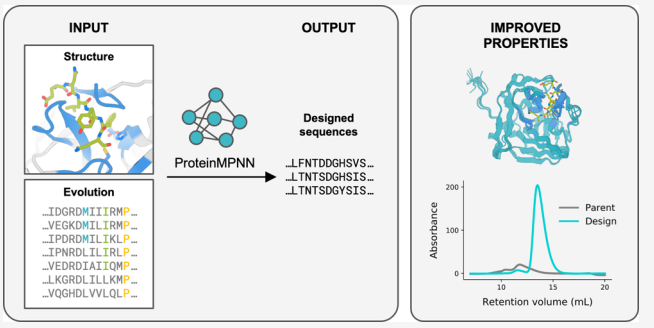
Kiera H. Sumida, Reyes Núñez-Franco, Indrek Kalvet, Samuel J. Pellock, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, Jue Wang, Yakov Kipnis, Noel Jameson, Alex Kang, Joshmyn De La Cruz, Banumathi Sankaran, Asim K. Bera, Gonzalo Jiménez-Osés, and David Baker*

Cite This: J. Am. Chem. Soc. 2024, 146, 2054–2061

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Natural proteins are highly optimized for function but are often difficult to produce at a scale suitable for biotechnological applications due to poor expression in heterologous systems, limited solubility, and sensitivity to temperature. Thus, a general method that improves the physical properties of native proteins while maintaining function could have wide utility for protein-based technologies. Here, we show that the deep neural network ProteinMPNN, together with evolutionary and structural information, provides a route to increasing protein expression, stability, and function. For both myoglobin and tobacco etch virus (TEV) protease, we generated designs with improved expression, elevated melting temperatures, and improved function. For TEV protease, we identified multiple designs with improved catalytic activity as compared to the parent sequence and previously reported TEV variants. Our approach should be broadly useful for improving the expression, stability, and function of biotechnologically important proteins.



INTRODUCTION

Evolution has optimized function over stability in most natural proteins;[1] as a result, they often exhibit poor solubility, thermostability, and expression in heterologous systems, all of which reduce the yield of functional protein.[2,3] Many protein-based therapeutics and catalysts are limited in their industrial application by low stability, making protein stabilization a research area of increasing interest.[4,5] Experimental methods such as directed evolution have been extensively used to optimize desirable features in proteins but are often prohibitively resource- and labor-intensive.[6,7] Computational tools have been developed to achieve the benefits of directed evolution while minimizing experimental screening.[8−11] PROSS (protein repair one-stop shop), for example, utilizes evolutionary information and Rosetta physics-based energy calculations to perform sequence redesign using a three-dimensional (3D) structure as input and has been shown to increase the soluble expression and thermostability of several natural proteins.[8] More recently, advances in deep learning-based modeling of proteins have been applied to generate new variants of natural proteins, including language models that generate sequences for a given enzyme family or function,[11] convolutional neural networks that leverage structural information for the prediction of gain-of-function mutations,[10] and shallow neural networks for guiding combinatorial directed evolution.[12]

Deep learning-based tools for protein sequence design have shown success in the generation of novel proteins with excellent expression, solubility, and sub-angstrom accuracy to design models.[11,13,14] ProteinMPNN generates highly stable sequences for designed backbones, and for native backbones, it generates sequences that are predicted to fold to the intended structures more confidently than their native sequences.[13] We reasoned that ProteinMPNN could be applied to protein stability optimization and set out to develop a strategy for applying ProteinMPNN to natural proteins to increase solubility and stability. We chose as model systems one of the first proteins whose structure was solved, the oxygen storage protein myoglobin, and the widely used protease from tobacco etch virus (TEV).

RESULTS

Protein Stabilization with ProteinMPNN. ProteinMPNN generates amino acid sequences that are predicted to fold into a given 3D structure. The method is purely

Received: October 4, 2023
Revised: December 3, 2023
Accepted: December 5, 2023
Published: January 9, 2024

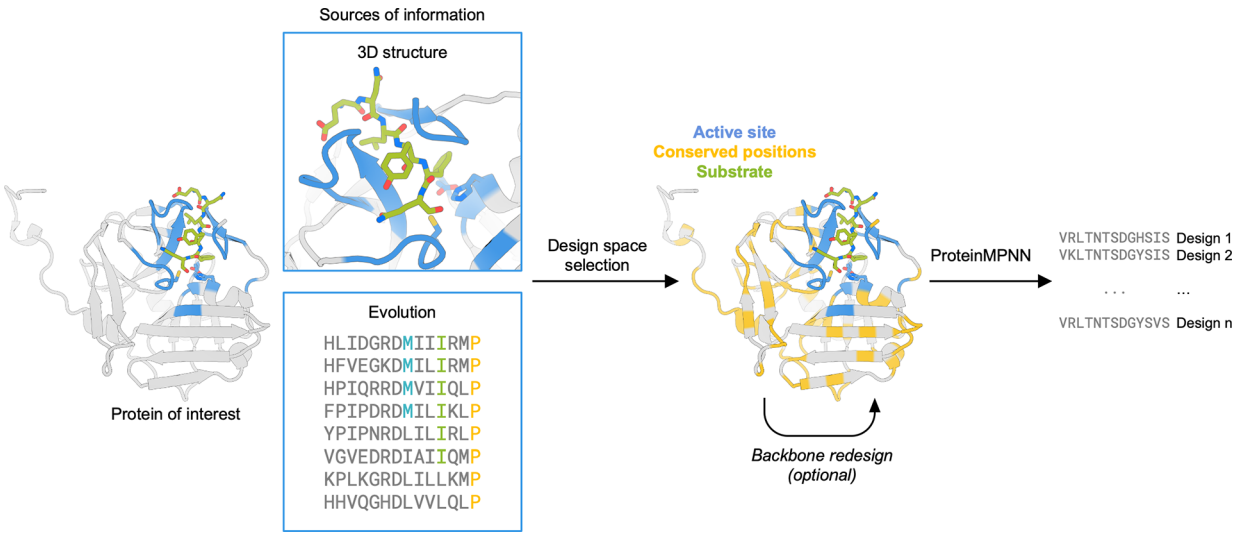


Figure 1. Design strategy for the optimization of protein expression and stability using ProteinMPNN. The design space is chosen to preserve the native protein function by fixing the amino acid identities of residues close to the ligand and those that are highly conserved in multiple sequence alignments. The protein backbone structure and fixed position information are input into ProteinMPNN, which generates new amino acid sequences likely to fold to the input structure. The backbone structure in loop regions can optionally be remodeled using RoseTTAfold joint inpainting to further idealize the input protein.

structure-based and does not have access to functional information. Therefore, to retain protein function during sequence design, additional information must be provided to the network. We experimented with a range of approaches to retain functionality during the design process. In all targets, to preserve the catalytic machinery and substrate-binding site, we fixed the amino acid identities of the first shell functional positions—defined as those within 7 Å of the substrate in a ligand-bound crystal structure complex. For TEV protease, we used evolutionary information to further identify residues critical to activity. In myoglobin, we performed a limited backbone redesign to further stabilize the structure. With the design space selected, we generated sequences with ProteinMPNN, predicted the structures with AlphaFold2,[15] and filtered by the predicted local distance difference test score (pLDDT) and Cα root-mean-square deviation (RMSD) to the input structure (Figure 1).

Design of Myoglobin Variants with Increased Stability. We first applied our design strategy to the model protein myoglobin. Myoglobin binds heme to carry oxygen in mammalian muscle tissue,[16] and has relevance in clinical applications as a biomarker,[17] as a versatile platform for biocatalytic applications,[18−20] and in food science as an ingredient in artificial meat products.[21−23] Current efforts to create more stable variants of myoglobin have focused on the stabilization of the globin fold through stapling with cysteine-reactive noncanonical amino acids.[24,25]

We applied the ProteinMPNN design protocol described above using a crystal structure of human myoglobin, nMb (PDB: 3RGK).[26] To preserve the oxygen storage function, we fixed the identities of 17 positions located around the heme ligand in the heme-bound structure (Figure 2A). Sixty sequences were generated with ProteinMPNN and evaluated for their likelihood to recapitulate the myoglobin backbone coordinates using AlphaFold2 single-sequence predictions (see Supporting Information). Eight of the designs did so with high confidence (pLDDT > 85.0 and Cα RMSD < 1.0 Å; analogous single-sequence prediction of the native sequence yielded pLDDT = 50.6 and Cα RMSD = 7.5 Å). Four designs with close structural agreement in the heme-binding region were selected for experimental testing.

We also explored the limited backbone redesign of poorly ordered regions to attempt to further stabilize the protein. The globin superfamily, of which myoglobin is a member, has a fold made up of eight alpha helical regions, with diversity in the termini and two loop regions flanking the heme-binding pocket[27−29] (Figure S1). We selected these less-conserved loop regions for backbone remodeling with RoseTTAFold joint inpainting (Figure 2A).[30] We generated two distinct sets of designs with structural remodeling: one with the region joining helices E and F redesigned and one additionally including the CD-loop region (Figure 2A). From these remodeled backbones, we again performed sequence design with ProteinMPNN, with the heme-binding site kept fixed as described above. Following filtering on structure prediction metrics (Figure S2), an additional 16 sequences were selected for experimental testing. All 20 tested myoglobin designs have 41−55% sequence identity with the most similar protein (a myoglobin in all cases) in the UniRef100 database[31] (Table S1).

Synthetic genes encoding the designs and the parent sequence, nMb, were expressed in *E. coli*. The heme-loaded *holo*-proteins were purified via immobilized metal affinity chromatography (IMAC) and size exclusion chromatography (SEC). All designs were expressed and were monomeric by SEC (Figure 2B). Thirteen of the twenty designs had higher levels (up to a 4.1-fold increase) of total soluble protein yield compared to that of native myoglobin (Figure 2C). All 20 designs had similar heme-binding spectra to native myoglobin, with agreement in the Soret maximum (407−413 nm vs 409 nm in native) and Q-band features (500, 537, 582, and 630 nm), suggesting the preservation of the native heme-binding mechanism (Figure S3).

The thermal stabilities of eight highly-expressing designs (six and two designed with and without backbone remodeling, respectively) were evaluated by circular dichroism (CD)

https://doi.org/10.1021/jacs.3c10941
J. Am. Chem. Soc. 2024, 146, 2054−2061

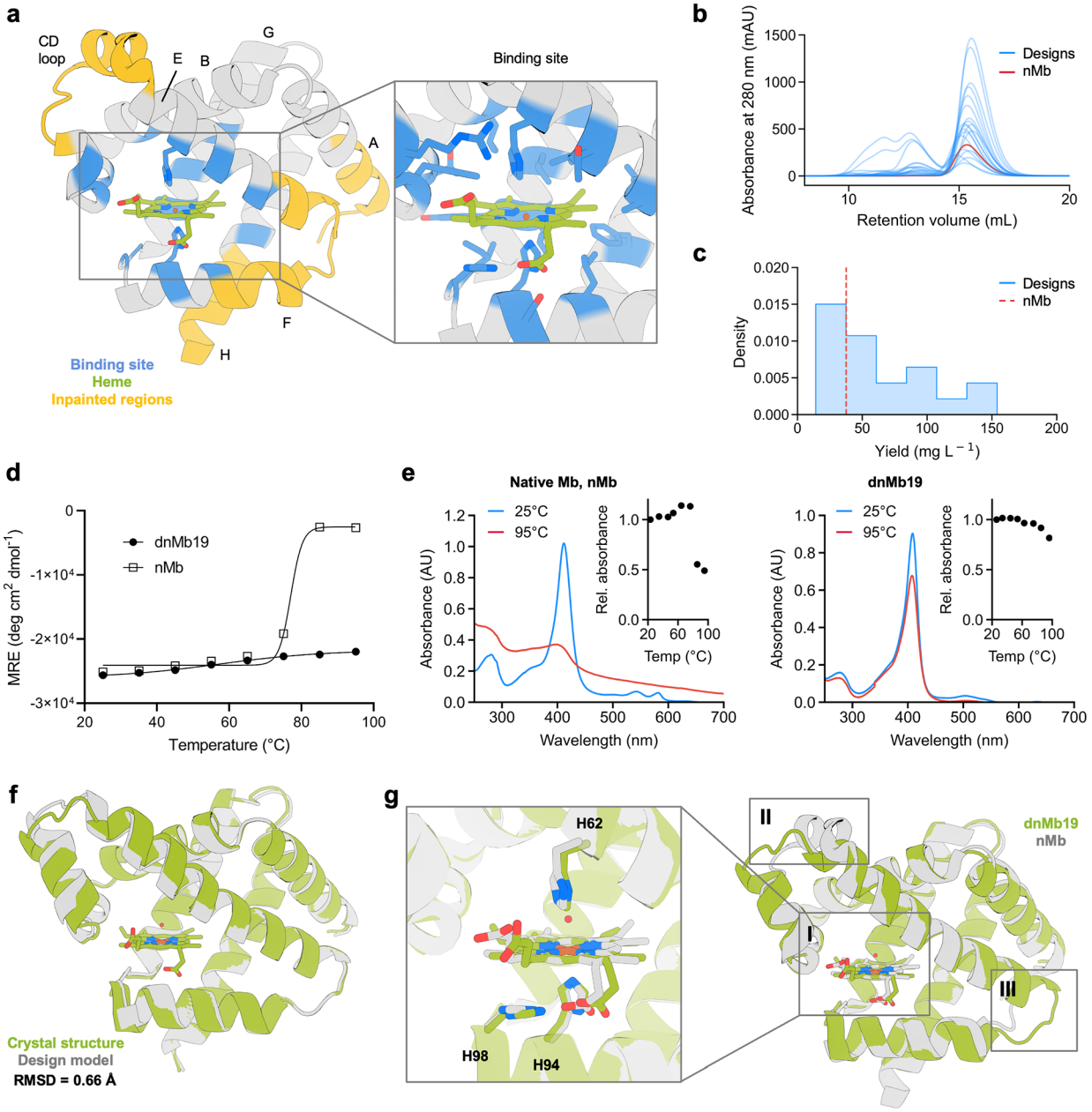


Figure 2. ProteinMPNN design improves myoglobin expression and thermostability. (a) Positions adjacent to the heme were kept fixed during the sequence design (shown in blue). Non-conserved regions (in yellow) were subjected to backbone remodeling. Inset shows the heme-binding site. (b) SEC traces of 20 designed myoglobin variants. (c) Soluble yield of myoglobin designs and native myoglobin (nMb, represented as a red dashed line). (d) CD melting temperature plots of dnMb19 compared to native myoglobin (signal reported in molar residue ellipticity (MRE)). (e) Absorbance plots of dnMb19 and native myoglobin (inset shows the temperature scan). (f) Structural alignment of the crystal structure (green) and AlphaFold2 (AF2) prediction (gray) of dnMb19. (g) Overlay of the crystal structure of native myoglobin (gray) and the crystal structure of dnMb19 (green, PDB: 8U5A). Non-conserved regions displayed in insets II and III were subjected to backbone redesign.

spectroscopy. All eight designs had higher melting temperatures than that of native myoglobin, with six remaining fully folded at 95 °C (native myoglobin melts at 80 °C; Figures 2D and S4). Heme binding was also evaluated over a temperature gradient to determine the functional thermal stability. All designs preserved heme binding at higher temperatures than native myoglobin (as monitored by changes to the Soret band wavelength and intensity in the UV/vis spectrum), with five designs maintaining significant heme-binding at 95 °C (Figure S5). One of the five designs, dnMb19, generated with the more aggressive backbone remodeling strategy, showed a much higher thermal stability of heme binding compared to native myoglobin (Figure 2E). Overall, remodeling regions of the myoglobin backbone with inpainting increased the success rate for retaining heme-binding at elevated temperatures.

To understand the structural basis of these improvements in stability, we determined the crystal structure of dnMb19 (2.0 Å resolution, PDB: 8U5A). We found that it closely agreed with the design model (0.66 Å Cα RMSD, Figure 2F), including the regions remodeled with inpainting. Native side chain contacts with the heme group are largely preserved in dnMb19 (Figure 2G, inset I). Outside of the heme-binding site, the crystal structure confirms the structural changes introduced by inpainting: the C and E helices were elongated as designed

https://doi.org/10.1021/jacs.3c10941
J. Am. Chem. Soc. 2024, 146, 2054−2061

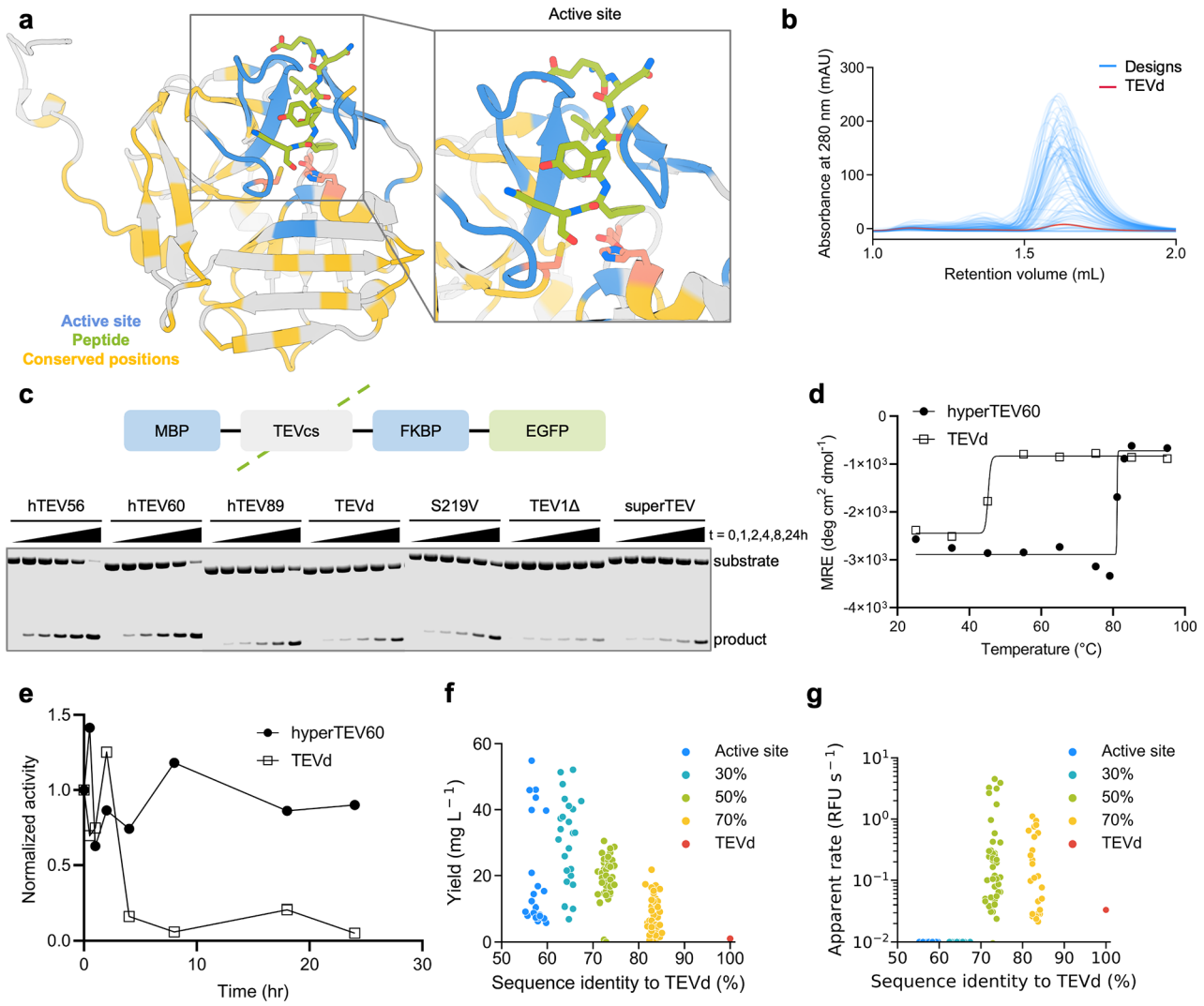


Figure 3. ProteinMPNN sequence design improves TEV protease expression, thermostability, and catalytic efficiency. (a) TEVd (PDB: 1LVM) input structure with positions fixed during redesign highlighted. Active site residues surrounding the substrate (blue), 50% most highly conserved residues (yellow), and catalytic residues (pink) are highlighted. Inset shows a zoomed-in view of the active site region. (b) SEC traces of the designed TEV variants. (c) Diagram of TEV substrate (top) and fluorescent gel image of TEV cleavage reactions at various time points (bottom). (d) CD melting temperature plots of the designed and native TEV (signal reported in molar residue ellipticity (MRE)). (e) Benchtop stability comparison of native TEVd and the designed variant assessed as activity measured over time incubated at 30 °C before inclusion in the assay. (f) Decreased evolutionary constraints correlate with higher soluble expression levels. Legend indicates regions fixed during the design (all designs have the active site fixed). (g) Designs made with the active site and 50% most conserved residues fixed during design exhibited the highest catalytic activity. Raw apparent rate is reported in relative fluorescence units (RFU) per second.

and connected by a new loop (Figure 2G, inset II); the loop connecting the E and F helices has a new conformation, and the F helix was straightened through the replacement of PRO88 with GLU89 (Figure 2G, inset III). The Cα RMSD over the inpainted regions between the crystal structure and the design model is 0.88 Å, with the largest deviation being in the CD-loop region (1.51 Å). These results illustrate the power of RoseTTAFold joint inpainting and ProteinMPNN to accurately remodel native protein backbones while increasing solubility, thermostability, and functional stability.

Design of TEV Protease Variants with Improved Stability and Catalytic Activity. To explore the utility of ProteinMPNN sequence design for stabilizing enzymes, we next applied our design strategy to the cysteine protease from tobacco etch virus (TEV). TEV protease is widely used in biotechnological applications to specifically cleave between glutamine and serine in its recognition sequence (ENLYFQ/S) to remove purification tags from recombinant proteins. However, TEV protease has suboptimal properties, including low soluble yield from heterologous expression, low thermostability, and poor catalytic activity. These properties often necessitate long incubation times and result in incomplete cleavage.[32]

We applied our sequence design strategy to TEV protease starting from an autolysis-resistant S219D variant, TEVd (PDB: 1LVM).[33] We defined the active site residues as described above to be fixed during redesign. We additionally fixed the amino acid identities of residues that are most conserved within the protein family (determined from a sequence alignment generated against UniRef30[31]), as residues distant from the active site can contribute significantly to function.[34] We ranked each amino acid identity at each position by the degree of conservation in the sequence alignment and varied the percentage of these most highly

https://doi.org/10.1021/jacs.3c10941
J. Am. Chem. Soc. 2024, 146, 2054−2061

Table 1. Kinetic Parameters for TEV Redesigns and the Parent TEV Variant

variant	k_{cat} (min^{-1})	K_m (μM)	k_{cat}/K_m (μM^{-1} min^{-1})	fold improvement in k_{cat}/K_m over parent
hyperTEV56	0.0106 ± 0.0005	1.4 ± 0.2	0.0077	20
hyperTEV60	0.014 ± 0.002	1.4 ± 0.4	0.01	26
hyperTEV89	0.0050 ± 0.0001	2 ± 1	0.0024	6.2
TEVd	0.0023 ± 0.0003	6 ± 3	0.00039	

conserved residues to fix during sequence redesign between 30 and 70%. We generated four distinct sets of designs that fixed the amino acid identities of just the active site residues or the active site residues and 30, 50, and 70% of the most conserved residues in the TEV family (Figure 3A, see Supporting Information). A total of 144 sequences were generated with ProteinMPNN, which were all predicted with high confidence to fold to the TEV structure by AlphaFold2 (pLDDT > 87.5; native TEV is predicted with pLDDT = 90) and possess 55 to 85% sequence identity to the parent sequence. All 144 designs were selected for experimental testing.

Synthetic genes encoding the designs, the parent sequence, TEVd, and several previously reported TEV variants were expressed in *E. coli*, and the resultant proteins were purified via IMAC and SEC. 134 of 144 designs solubly expressed and eluted as monomers by SEC (Figure 3B). 129 of 144 designs exhibited higher levels of soluble expression than TEVd (TEVd average yield = 1 mg/L culture, design average yield = 20.1 mg/L culture (Figure 3F)).

We evaluated catalytic activity using a previously described[7] coumarin derivative with 7-amino-4-trifluoromethylcoumarin conjugated to the C-terminus of the TEV substrate peptide Ac-ENLYFQ (Figure S7A). Purified protein was incubated with the peptide-coumarin substrate, and 64 designs displayed progress curves with fluorescence above the background, indicating substrate turnover (Figure S7B and S7C). Designs made with no evolutionary constraints had improved soluble expression over the parent but were not active on the peptide substrate, while designs with the highest activities were designed with the top 50% most conserved residues fixed (Figure 3F,G). We performed detailed kinetic analysis of three highly active designs from the 50% design method—hyperTEV56, hyperTEV60, and hyperTEV89—and the parent sequence TEVd.[8] The designs displayed improved catalytic efficiencies (k_{cat}/K_m) compared to TEVd, with up to 26-fold improvements (Table 1 and Figure S8).

Next, we tested the most active designs with a fusion protein substrate to assess their performance on the target application of tag removal. The designs and a set of previously engineered TEV proteases[32,33,35−37] were incubated at 30 °C with the fusion protein substrate MBP-TEVcs-FKBP-EGFP, where MBP is the maltose-binding protein, TEVcs is the TEV peptide cleavage site (ENLYFQS), FKBP is the FK506-binding protein, and EGFP is an enhanced green fluorescent protein. The extent of proteolysis was evaluated by monitoring the accumulation of the cleaved product via sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (Figure S9). Two designs, hyperTEV56 and hyperTEV60, exhibited significantly higher rates of cleavage of protein substrate compared to the parent TEVd, yielding 50% cleaved product at ∼4 h of incubation, while TEVd required 24 h to reach an equivalent yield. The designs also outperformed other published TEV variants, with 30% turnover for superTEV, 15% turnover for TEV1Δ, and 50% turnover for S219V at 24 h of incubation (Figures 3C and S10A). Straight-line fits of product accumulation and substrate depletion reveal catalytic efficiencies that corroborate those determined in the peptide assay (Figure S10B). In the peptide assay, the gains in catalytic efficiency are primarily due to increases in k_{cat}, which could reflect a higher fraction of enzyme in a catalytically competent state (see below).

Analysis by CD spectroscopy of TEVd and the most active design, hyperTEV60, indicated an approximate melting temperature of 84 °C for hyperTEV60, 40 °C higher than that of TEVd (Figures 3D and S11), and to the best of our knowledge, higher than that of any previously described TEV variant. To further probe the stability of the designed variant, TEVd and hyperTEV60 were incubated at 30 °C for various times and then used in the peptide−coumarin cleavage assay. After 4 h of incubation, hyperTEV60 retained 90% of its original cleavage activity, while TEVd was reduced to 15% of its original activity (Figure 3E), indicating a significant improvement in benchtop stability.

Given that the catalytic and substrate-binding residues were kept fixed during the design with ProteinMPNN, it is notable that significant improvements in k_{cat} were observed with both the peptide and protein substrates. Mutations distal to the active site can influence catalytic activity through the stabilization of catalytically productive conformational states[38,39] or global conformational changes.[40] To investigate if the stabilization of functional conformational states may be involved in activity enhancement, we performed microsecond molecular dynamics (MD) simulations on TEV−peptide complexes to probe the impact of the introduced mutations on the overall protein dynamics. A general rigidification of loop regions distributed across the structure was observed in designs compared to TEVd (Figure S12A). This backbone rigidification in distal regions not directly involved in substrate binding may be related to allosteric improvement of substrate binding, as reflected by the two- to threefold lower K_m values measured for the designed variants (Table 1). Rigidification in the region spanning residues 115 to 124 appeared to correlate with activity; the highest activity design, hyperTEV60, was most rigid, while TEVd and a design with no activity on the peptide substrate were the most flexible in this region (Figure S12B). These trends were also observed in the per-residue pLDDT analysis of AlphaFold2 ensemble predictions (Figure S12C). In all designs, we observed a decrease in the population of catalytically competent conformations of the Cys-His dyad ($d_{N\text{-}SH}$) compared to TEVd, but this shift was least significant in hyperTEV60, in agreement with its higher relative k_{cat} (Figure S13). These notable differences may begin to explain how ProteinMPNN enables substantial activity enhancements without explicit design elements to improve function. It is also possible that the major contribution to the increase in k_{cat} is from an increase in the fraction of the protein in the catalytically competent state more globally.

https://doi.org/10.1021/jacs.3c10941
J. Am. Chem. Soc. 2024, 146, 2054−2061

DISCUSSION

We show that the expression, stability, and function of native proteins can be improved using ProteinMPNN, guided by the available sequence and structural information. For both TEV protease and human myoglobin, multiple variants were identified that showed higher soluble yield and thermostability than the native protein. The best of the TEV protease designs have higher apparent catalytic efficiency on peptide and protein substrates than the parent enzyme and previously reported variants. While the optimal number of residues to maintain (and perhaps enhance) function may have to be determined empirically for each case, the simplicity of our procedure and the computational efficiency and ease of use of ProteinMPNN make this straightforward, and the number of variants that need to be tested is far smaller than that in typical experimental screens. We expect that our approach will be widely useful for improving the expression, stability, and function of biotechnologically important proteins.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.3c10941.

Computational and experimental methods, details of crystallographic data collection, sequences of designs, structural diversity of natural and designed myoglobin variants; *in silico* metrics of myoglobin designs grouped by a method; raw thermostability, spectrophotometry, and activity data; raw data from Michaelis Menten fitting of TEV activity assays; data from MD simulations; sequence similarity analysis of myoglobin designs; and mass spectrometry data for purified myoglobin designs (PDF)

Computational models of designed proteins (ZIP)

AUTHOR INFORMATION

Corresponding Author

David Baker − *Institute for Protein Design, Department of Biochemistry, and Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, United States*; Email: dabaker@uw.edu

Authors

Kiera H. Sumida − *Department of Chemistry and Institute for Protein Design, University of Washington, Seattle, Washington 98195, United States*; orcid.org/0000-0003-2773-9676

Reyes Núñez-Franco − *Center for Cooperative Research in Biosciences, Basque Research and Technology Alliance, Derio 48160, Spain*

Indrek Kalvet − *Institute for Protein Design, Department of Biochemistry, and Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, United States*; orcid.org/0000-0002-6610-2857

Samuel J. Pellock − *Institute for Protein Design and Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States*

Basile I. M. Wicky − *Institute for Protein Design and Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States*; orcid.org/0000-0002-2501-7875

Lukas F. Milles − *Institute for Protein Design and Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States*

Justas Dauparas − *Institute for Protein Design and Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States*

Jue Wang − *Institute for Protein Design and Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States*

Yakov Kipnis − *Institute for Protein Design, Department of Biochemistry, and Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, United States*

Noel Jameson − *Department of Chemistry, University of Washington, Seattle, Washington 98195, United States*; orcid.org/0000-0001-9231-3765

Alex Kang − *Institute for Protein Design, University of Washington, Seattle, Washington 98195, United States*

Joshmyn De La Cruz − *Institute for Protein Design, University of Washington, Seattle, Washington 98195, United States*

Banumathi Sankaran − *Berkeley Center for Structural Biology, Molecular Biophysics, and Integrated Bioimaging, Lawrence Berkeley Laboratory, Berkeley, California 94720, United States*

Asim K. Bera − *Institute for Protein Design and Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States*

Gonzalo Jiménez-Osés − *Center for Cooperative Research in Biosciences, Basque Research and Technology Alliance, Derio 48160, Spain; Ikerbasque, Basque Foundation for Science, Bilbao 48013, Spain*; orcid.org/0000-0003-0105-4337

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.3c10941

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank A. Lauko, C. Norn, A. Roy, and L. Stewart for helpful discussions. We thank L. Goldschmidt and K. VanWormer for computational and experimental support, respectively. We thank X. Li and M. Lamb for analytical services. We also thank Agencia Estatal Investigacion of Spain (PID2021-125946OB-I00, CEX2021-001136-S, predoctoral fellowship; G.J.O. and R.N.F.) for support of this work. Crystallographic data were collected at the Advanced Light Source (ALS), which is supported by the Director, Office of Science, Office of 20 Basic Energy Sciences, and US Department of Energy under contract number DE-AC02-05CH11231. Funding was also provided by a National Science Foundation (NSF) grant CHE-1629214 (A.K.B.), the Air Force Office of Scientific Research FA9550-18-1-0297 (S.J.P.), a Defense Threat Reduction Agency grant HDTRA1-19-1-0003 (S.J.P.), the Bill and Melinda Gates Foundation grant OPP1156262 (A.K., J.C.), the National Institute of Health's National Institute of Allergy and Infectious Disease (R0AI160052, A.B.K.), the DARPA program Harnessing Enzymatic Activity for Lifesaving Remedies (HEALR) (HR0011-21-2-0012, A.K.B.), the Audacious Project at the Institute for Protein Design (L.F.M., A.K., J.C., E.B., A.K.B., and D.B.), the Howard Hughes Medical Institute (I.K., Y.K., and D.B.), the Open Philanthropy Project Improving Protein

Design Fund (K.H.S., S.J.P., I.K., B.I.M.W., J.D., J.W., Y.K., A.K., J.C., E.B., A.K.B., and D.B.), Schmidt Futures (J.W.), a grant from the National Science Foundation (NSF) (DBI 1937533; D.B.), the Department of Energy ARPA-E Grant 2459-1671 (D.B.), an EMBO long-term fellowship (ALTF 139-2018; B.I.M.W.), a Washington Research Foundation Fellowship (S.J.P. and J.W.), an Alfred P. Sloan Foundation Matter-to-Life Program Grant (G-2021-16899; D.B.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019; L.F.M.), a Human Frontier Science Program Cross-Disciplinary Fellowship (LT000838/2018-C (I.K.), LT000395/2020-C (L.F.M.)), the National Institute of Health (R35 GM124773; N.J.), and a gift from Microsoft (J.D. and D.B.).

REFERENCES

(1) Beadle, B. M.; Shoichet, B. K. Structural Bases of Stability−function Tradeoffs in Enzymes. *J. Mol. Biol.* **2002**, *321* (2), 285−296.
(2) Magliery, T. J. Protein Stability: Computation, Sequence Statistics, and New Experimental Methods. *Curr. Opin. Struct. Biol.* **2015**, *33*, 161−168.
(3) Singh, A.; Upadhyay, V.; Upadhyay, A. K.; Singh, S. M.; Panda, A. K. Protein Recovery from Inclusion Bodies of Escherichia Coli Using Mild Solubilization Process. *Microb. Cell Fact.* **2015**, *14*, 41.
(4) Thomson, R. E. S.; Carrera-Pacheco, S. E.; Gillam, E. M. J. Engineering Functional Thermostable Proteins Using Ancestral Sequence Reconstruction. *J. Biol. Chem.* **2022**, *298* (10), No. 102435.
(5) Rathore, N.; Rajan, R. S. Current Perspectives on Stability of Protein Drug Products during Formulation Fill and Finish Operations. *Biotechnol. Prog.* **2008**, *24* (3), 504−514.
(6) Cobb, R. E.; Chao, R.; Zhao, H. Directed Evolution: Past, Present and Future. *AIChE J.* **2013**, *59* (5), 1432−1440.
(7) Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem., Int. Ed. Engl.* **2018**, *57* (16), 4143−4148.
(8) Goldenzweig, A.; Goldsmith, M.; Hill, S. E.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; Lieberman, R. L.; Aharoni, A.; Silman, I.; Sussman, J. L.; Tawfik, D. S.; Fleishman, S. J. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2018**, *70* (2), 380.
(9) Shroff, R.; Cole, A. W.; Diaz, D. J.; Morrow, B. R.; Donnell, I.; Annapareddy, A.; Gollihar, J.; Ellington, A. D.; Thyer, R. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth. Biol.* **2020**, *9* (11), 2927−2935.
(10) Lu, H.; Diaz, D. J.; Czarnecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; Lynd, N. A.; Ellington, A. D.; Alper, H. S. Machine Learning-Aided Engineering of Hydrolases for PET Depolymerization. *Nature* **2022**, *604* (7907), 662−667.
(11) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L., Jr; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; Naik, N Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nat. Biotechnol.* **2023**, *41*, 1099−1106.
(12) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine Learning-Assisted Directed Protein Evolution with Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (18), 8852−8858.
(13) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; de Haas, R. J.; Bethel, N.; Leung, P. J. Y.; Huddy, T. F.; Pellock, S.; Tischer, D.; Chan, F.; Koepnick, B.; Nguyen, H.; Kang, A.; Sankaran, B.; Bera, A. K.; King, N. P.; Baker, D. Robust Deep Learning−based Protein Sequence Design Using ProteinMPNN. *Science* **2022**, *378* (6615), 49−56.
(14) Wicky, B. I. M.; Milles, L. F.; Courbet, A.; Ragotte, R. J.; Dauparas, J.; Kinfu, E.; Tipps, S.; Kibler, R. D.; Baek, M.; DiMaio, F.; Li, X.; Carter, L.; Kang, A.; Nguyen, H.; Bera, A. K.; Baker, D. Hallucinating Symmetric Protein Assemblies. *Science* **2022**, *378* (6615), 56−61.
(15) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(16) Ordway, G. A.; Garry, D. J. Myoglobin: An Essential Hemoprotein in Striated Muscle. *J. Exp. Biol.* **2004**, *207* (Pt 20), 3441−3446.
(17) Hamm, C. W. Cardiac Biomarkers for Rapid Evaluation of Chest Pain. *Circulation* **2001**, *104* (13), 1454−1456.
(18) Bordeaux, M.; Tyagi, V.; Fasan, R. Highly Diastereoselective and Enantioselective Olefin Cyclopropanation Using Engineered Myoglobin-Based Catalysts. *Angew. Chem., Int. Ed. Engl.* **2015**, *54* (6), 1744−1748.
(19) Carminati, D. M.; Decaens, J.; Couve-Bonnaire, S.; Jubault, P.; Fasan, R. Biocatalytic Strategy for the Highly Stereoselective Synthesis of CHF2 -Containing Trisubstituted Cyclopropanes. *Angew. Chem., Int. Ed. Engl.* **2021**, *60* (13), 7072−7076.
(20) Brandenberg, O. F.; Fasan, R.; Arnold, F. H. Exploiting and Engineering Hemoproteins for Abiological Carbene and Nitrene Transfer Reactions. *Curr. Opin. Biotechnol.* **2017**, *47*, 102−111.
(21) Fraser, R.; Brown, P. O.; Karr, J.; Holz-Schietinger, C.; Cohn, E. Methods and Compositions for Affecting the Flavor and Aroma Profile of Consumables. US Patent, 9700067B2, 2017.
(22) Simsa, R.; Yuen, J.; Stout, A.; Rubio, N.; Fogelstrand, P.; Kaplan, D. L. Extracellular Heme Proteins Influence Bovine Myosatellite Cell Proliferation and the Color of Cell-Based Meat. *Foods* **2019**, *8* (10), 521.
(23) Devaere, J.; De Winne, A.; Dewulf, L.; Fraeye, I.; Šoljić, I.; Lauwers, E.; de Jong, A.; Sanctorum, H. Improving the Aromatic Profile of Plant-Based Meat Alternatives: Effect of Myoglobin Addition on Volatiles. *Foods* **2022**, *11* (13), 1985.
(24) Moore, E. J.; Zorine, D.; Hansen, W. A.; Khare, S. D.; Fasan, R. Enzyme Stabilization via Computationally Guided Protein Stapling. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (47), 12472−12477.
(25) Iannuzzelli, J. A.; Bacik, J.-P.; Moore, E. J.; Shen, Z.; Irving, E. M.; Vargas, D. A.; Khare, S. D.; Ando, N.; Fasan, R. Tuning Enzyme Thermostability via Computationally Guided Covalent Stapling and Structural Basis of Enhanced Stabilization. *Biochemistry* **2022**, *61* (11), 1041−1054.
(26) Hubbard, S. R.; Hendrickson, W. A.; Lambright, D. G.; Boxer, S. G. X-Ray Crystal Structure of a Recombinant Human Myoglobin Mutant at 2·8 Å Resolution. *J. Mol. Biol.* **1990**, *213* (2), 215−218.
(27) Keppner, A.; Maric, D.; Correia, M.; Koay, T. W.; Orlando, I. M. C.; Vinogradov, S. N.; Hoogewijs, D. Lessons from the Post-Genomic Era: Globin Diversity beyond Oxygen Binding and Transport. *Redox Biol* **2020**, *37*, No. 101687.
(28) Kapp, O. H.; Moens, L.; Vanfleteren, J.; Trotman, C. N.; Suzuki, T.; Vinogradov, S. N. Alignment of 700 Globin Sequences: Extent of Amino Acid Substitution and Its Correlation with Variation in Volume. *Protein Sci.* **1995**, *4* (10), 2179−2190.
(29) Gell, D. A. Structure and Function of Haemoglobins. *Blood Cells Mol. Dis.* **2018**, *70*, 13−42.
(30) Wang, J.; Lisanza, S.; Juergens, D.; Tischer, D.; Watson, J. L.; Castro, K. M.; Ragotte, R.; Saragovi, A.; Milles, L. F.; Baek, M.; Anishchenko, I.; Yang, W.; Hicks, D. R.; Expòsit, M.; Schlichthaerle, T.; Chun, J.-H.; Dauparas, J.; Bennett, N.; Wicky, B. I. M.; Muenks, A.; DiMaio, F.; Correia, B.; Ovchinnikov, S.; Baker, D. Scaffolding Protein Functional Sites Using Deep Learning. *Science* **2022**, *377* (6604), 387−394.
(31) Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters. *Bioinformatics* **2007**, *23* (10), 1282−1288.

https://doi.org/10.1021/jacs.3c10941
J. Am. Chem. Soc. 2024, 146, 2054−2061

(32) Blommel, P. G.; Fox, B. G. A Combined Approach to Improving Large-Scale Production of Tobacco Etch Virus Protease. *Protein Expr. Purif.* **2007**, *55* (1), 53−68.
(33) Phan, J.; Zdanov, A.; Evdokimov, A. G.; Tropea, J. E.; Peters, H. K., 3rd; Kapust, R. B.; Li, M.; Wlodawer, A.; Waugh, D. S Structural Basis for the Substrate Specificity of Tobacco Etch Virus Protease. *J. Biol. Chem.* **2002**, *277* (52), 50564−50572.
(34) Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **2009**, *138* (4), 774−786.
(35) Kapust, R. B.; Tözsér, J.; Fox, J. D.; Anderson, D. E.; Cherry, S.; Copeland, T. D.; Waugh, D. S. Tobacco Etch Virus Protease: Mechanism of Autolysis and Rational Design of Stable Mutants with Wild-Type Catalytic Proficiency. *Protein Eng.* **2001**, *14* (12), 993−1000.
(36) Sanchez, M. I.; Ting, A. Y. Directed Evolution Improves the Catalytic Efficiency of TEV Protease. *Nat. Methods* **2020**, *17* (2), 167−174.
(37) Correnti, C. E.; Gewe, M. M.; Mehlin, C.; Bandaranayake, A. D.; Johnsen, W. A.; Rupert, P. B.; Brusniak, M.-Y.; Clarke, M.; Burke, S. E.; De Van Der Schueren, W.; Pilat, K.; Turnbaugh, S. M.; May, D.; Watson, A.; Chan, M. K.; Bahl, C. D.; Olson, J. M.; Strong, R. K. Screening,Large-Scale Production and Structure-Based Classification of Cystine-Dense Peptides. *Nat. Struct. Mol. Biol.* **2018**, *25* (3), 270−278.
(38) Otten, R.; Pádua, R. A. P.; Bunzel, H. A.; Nguyen, V.; Pitsawong, W.; Patterson, M.; Sui, S.; Perry, S. L.; Cohen, A. E.; Hilvert, D.; Kern, D. How Directed Evolution Reshapes the Energy Landscape in an Enzyme to Boost Catalysis. *Science* **2020**, *370* (6523), 1442−1446.
(39) Jiménez-Osés, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N. The Role of Distant Mutations and Allosteric Regulation on LovD Active Site Dynamics. *Nat. Chem. Biol.* **2014**, *10* (6), 431−436.
(40) Ishida, T. Effects of Point Mutation on Enzymatic Activity: Correlation between Protein Electronic Structure and Motion in Chorismate Mutase Reaction. *J. Am. Chem. Soc.* **2010**, *132* (20), 7104−7118.

https://doi.org/10.1021/jacs.3c10941
J. Am. Chem. Soc. 2024, 146, 2054−2061