Computational Stabilization of a Non-heme Iron Enzyme Enables Efficient Evolution of New Function

Brianne King

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee

Jesse Zalatan, Chair

Jorge Marchand

Lauren Rajakovich

Program Authorized to Offer Degree

Chemistry

University of Washington

Abstract

Computational Stabilization of a Non-heme Iron Enzyme Enables Efficient Evolution of New Function

Brianne King

Chair of the Supervisory Committee:

Jesse Zalatan

Department of Chemistry

Biocatalysis is a promising and sustainable alternative to conventional chemical manufacturing, which is responsible for a significant portion of global energy consumption. A major gap, however, is engineering enzymes for industrially relevant novel reactivity outside of their biological contexts. Combining the power of C-H functionalization with the superfamily of non-heme iron enzymes can address this gap and open the door to a suite of new-to-nature biocatalysts that can be readily incorporated into synthetic chemistry workflows. This thesis provides an overview of the conserved non-heme iron enzyme mechanism, which gives context to both substrate promiscuity and novel reactivity. Because wild-type enzymes often exhibit low reactivity for promiscuous substrates or reactions, protein engineering is necessary to optimize activity. Recent engineering and directed evolution efforts with non-heme iron enzymes are described, as well as general engineering challenges within this superfamily. One challenge that we explored in depth

is how to address activity-stability tradeoffs during directed evolution. Emerging deep learning methods for enzyme stabilization offer a potential solution to these tradeoffs, though their effectiveness for industrially relevant enzymes remains to be fully evaluated. Here, we evaluate the use of the deep-learning tool ProteinMPNN for redesigning a non-heme amino acid iron hydroxylase with promiscuous C-H hydroxylation activity towards a carboxylic acid substrate. We describe the process by which critical residues in our candidate enzyme were fixed prior to design to maintain catalytic function, and compare directed evolution trajectories of both wild-type and a stabilized redesign. In alignment with previous findings on how stability can promote evolvability, we found that a stabilized starting point leads to more efficient directed evolution. Together, our results suggest that user-friendly deep-learning tools like ProteinMPNN could be readily incorporated into enzyme engineering workflows to generate novel and robust biocatalysts.

Acknowledgements

Pursuing a PhD takes an equal mix of determination and curiosity, a bit of delusion, a lot of patience (which I don't have a lot of), and acceptance of how much we do not know. It is both a personal struggle of will and a community effort in survival. For me, it has been an arduous task of pushing myself to my limits while staving off burn-out. A lot of people say that "it is a marathon, not a sprint." I guess that is true if you run a marathon as short sprints, with long breaks laying flat out in front of aid stations, gulping down nebulous electrolyte drinks, and shoving fistfuls of off-brand puffy Cheetos in your mouth while staring at the sky wondering "what the fuck am I doing this for?" Like most of the masochistic outdoor activities I enjoy, I suppose pursuing a PhD is type II fun. Importantly, type II fun is best enjoyed with people around you. In getting to this point, I have so many people to thank. Mentors, friends, and family have all played a massive role in my success.

I have been incredibly fortunate to have amazing mentors who have provided support in so many forms. My graduate advisor, Jesse Zalatan, created an environment where I had freedom to explore big ideas and take scientific risks that were values-aligned for me. That environment is a gift, and I would not be where I am today without his support. We spent long stretches of our mentor-mentee relationship on what I like to call a "growth edge." Jesse pushed me to be a better scientist, and I am endlessly thankful for the high bar he set. I have pushed Jesse to be a better mentor, and while I am not sure he is thankful for that, he nevertheless rose to the challenge. Thank you, Jesse. I would also like to thank my committee members – Dusty Maly, Jorge Marchand, and Lauren Rajakovich – for helping me to shape my work to where it is today. I feel lucky to not only have mentors from this stage in my career, but mentors who have been a source of support and encouragement over the last sixteen years. To Ardi Kveven and Josh Searle at ORCA, thank you

for planting the seed of what it means to be a scientist and for the opportunities over these many years to engage with students and share what I am passionate about. To Greg O'Neil, Clint Spiegel, and Spencer Anthony-Cahill at Western – thank you for your continued advice, perspectives, and encouragement. I feel a profound sense of gratitude that I can call you both mentors and dear friends.

A key piece of the puzzle in getting to the end of my PhD has been the support of my lab mates. A big thanks goes to Noel Jameson, whose discussions over the years have greatly helped me think through puzzling data and plan next steps in my experiments. A lot can also be said for struggling in unison – without Noel's comradery, this journey would have been a lot harder. I will miss our walks to the DM to spend our grad school funny money on chicken tendies. To all the Zally Gallies, thank you for the many hours of yapping and fun adventures. Thank you to a select few of you for keeping me on my toes, for several reasons that shall remain between all of us. To all my mentees over the years – Jon Zhang, Caleb Kono, Daniel Ong, Jolene Nguyen, and Jess Caruso – thank you for letting me be a part of your scientific journeys and I cannot wait to see what you all accomplish. It truly takes a village to get through graduate school and I could not have asked for a better group of people to be on the struggle bus with. In addition to my lab mates, I have to thank my fellow graduate student organizers for fighting to make the department a better place.

Thank you to all my friends for being there for me throughout this process, especially Julie and Mollie. Thank you for celebrating the ups and listening to me rant about the downs. Thank you for your thoughtful perspectives, and for the sheer joy that comes from strong female friendships. Thank you to all of my barn friends, both two- and four-legged. My sanity was kept in check through the many hours of simply being around horses. To my family, I would not be

where I am today without you. A career in science is also an endeavor in creativity, and my mom is one of the most creative people I know. Even if I have made a poor financial decision in going into academia, thank you to my dad for being endlessly proud of my accomplishments. And thanks to my brother, too, for having absolutely no clue what it is that I do but being excited for me anyway. Thank you to Kevin, Dani, Moss, David and Cynthia – the many adventures in Seattle, the San Juans, and in Utah have all provided needed respites from the environment in academia and I am so grateful to have you in my life. Most importantly, thank you to my husband, Donald. You have been endlessly supportive of my scientific journey, and there have been long stretches where it certainly has not been easy. Thank you for always showing up, and for keeping me grounded in living a full and wonderful life outside of the lab. And saving the best for last, thank you to Chimi. You are the best cat in the world, and when this thesis is good and old and covered in dust on a shelf, I will remember you.

Table of Contents

Chapter 1: Introduction

Biocatalysis offers a valuable and sustainable alternative to current practices in chemical manufacturing, an industry which accounts for 10% of the global energy demand.[1] However, the scope of enzymatic reactions outside of a biological context is limited. To meet the demands of a sustainable chemical industry, an outstanding challenge is to rapidly engineer enzymes for new-to-nature reactions. One reaction class where new approaches to catalysis are needed is C-H functionalization, which allows for expedient diversification of simple or complex small molecules to a range of polyfunctional compounds.[2] Members of the non-heme iron enzyme superfamily (Fe(II)/αKGs) are a potential rich source of new C-H functionalization biocatalysts due to their ability to selectively activate C-H bonds which leads to diverse reaction outcomes.[3] This chapter covers the general mechanism of Fe(II)/αKGs, which provides context for both reaction and substrate promiscuity. Recent engineering efforts showcase how the mechanism can be co-opted for the successful generation of novel non-heme iron biocatalysts and highlight engineering challenges. Finally, general instability as a barrier during protein engineering of Fe(II)/αKGs is discussed in addition to the potential of protein stabilization methods in addressing this challenge.

1.1 General mechanism of non-heme iron α-KG dependent enzymes

Understanding the catalytic mechanism of enzymes that show promise in development of novel biocatalysts is a critical first step in proposing both new substrates and new reaction pathways. The subset of non heme Fe(II) enzymes that rely on the co-substrate αKG utilize a radical mechanism that proceeds through a common Fe(IV)-oxo intermediate (Figure 1E).[4] This intermediate gives rise to a wide diversity of reactions such as C-H hydroxylation, halogenation, epoxidation, desaturation, ring-expansion and contraction, and cyclization (Figure 1E).[3] The consensus mechanism begins with iron binding to a 2-His-1-carboxylate facial triad followed by binding of αKG (Figure 1B). The introduction of a substrate molecule leads to subsequent binding

of molecular oxygen and the formation of a putative Fe(III)-superoxide radical species, which then attacks the iron-bound αKG to form a Fe(II)-peroxy-succinate species (Figure 1C and 1D).[5] Subsequent O-O cleavage and release of CO_2 leads to the Fe(IV)-oxo intermediate (Figure 1E). The Fe(IV)-oxo intermediate then participates in hydrogen atom transfer (HAT) with a substrate C-H bond leading to a substrate carbon-centered radical (Figure 1F). In hydroxylases, product formation occurs after hydroxyl rebound with the carbon-centered radical from the Fe(III)-OH intermediate (Figure 1G). The Fe(IV)-oxo intermediate described above can also form in the absence of substrate or with weakly binding substrate, which can lead to a side reaction where substrate-uncoupled hydroxylation of enzyme active site residues leads to enzyme inactivation.[6,7] It may be important to consider the substrate-uncoupled pathway when exploring Fe(II)/αKG promiscuity with new substrates. However, many engineering campaigns with Fe(II)/αKGs do not report the uncoupled reaction rate,[8–11] and focus engineering efforts on improving overall substrate turnover to levels practical for chemical synthesis. Overall, the general Fe(II)/αKG mechanism described above can be adapted to novel reactivity by taking advantage of the powerful Fe(IV)-oxo intermediate and its participation in diverse radical reaction pathways with a wide range of substrates.

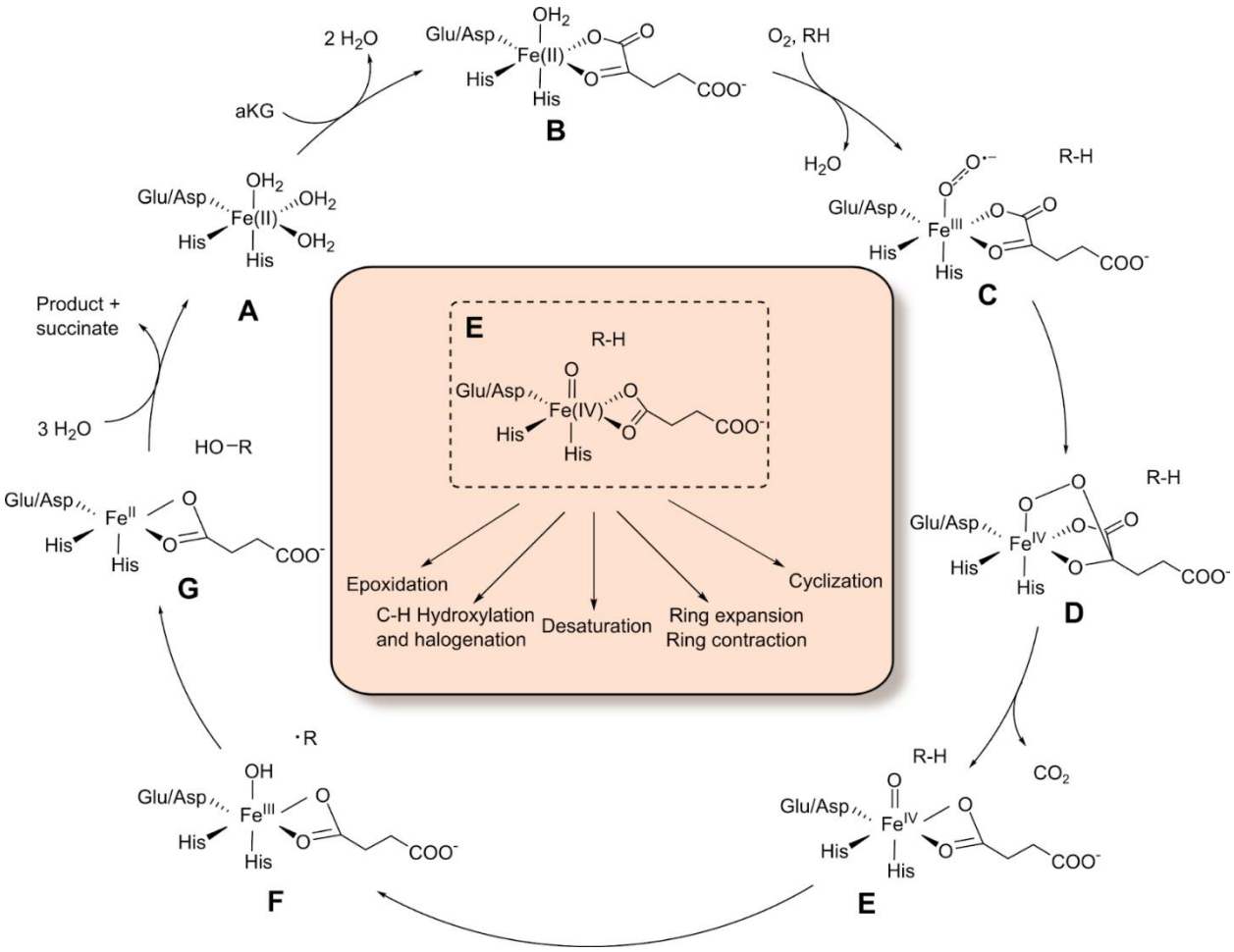


Figure 1. General mechanism of Fe(II)/αKG dependent enzymes.

1.2 Recent applications of non-heme iron enzymes in chemical synthesis

In the last decade, Fe(II)/αKGs have been increasingly utilized in both chemoenzymatic synthesis and as standalone biocatalysts for new-to-nature reactions (Figure 2A-H).[8–16] Generally, new reactions are discovered when wild-type enzymes are challenged with new substrates under native or alternative reaction conditions. Novel function can arise from both the native reaction pathway or an alternative reaction pathway with a non-native substrate. For example, co-opting the native C-H hydroxylation reaction pathway in the Fe(II)/αKG enzyme GriE towards a substrate analog containing an azide functional group leads to a hydroxylated product used in the synthesis of manzacidin C (Figure 2A).[12] Likely, substrate positioning allowed for the maintenance of site-

selectivity for hydroxyl rebound. Conversely, alternative reaction outcomes can arise when reactive functional groups like azides are positioned close to the iron center. This alternative reactivity is observed with the leucine hydroxylase PolL.[13] When an azide containing substrate analog is introduced to PolL, the reaction pathway shifts from one of hydroxy radical rebound to either nitrene formation or oxime formation with either pathway leading to a nitrile product (Figure 2B).[13] Another example of how positioning reactive functional groups close to the Fe(II)/αKG iron center can lead to diverse reaction outcomes is with the enzyme LdoA, where introduction of an olefin substrate analog results in epoxidation instead of hydroxyl rebound (Figure 2C).[14] Diverse reactivity can additionally be exploited by taking advantage of the non-heme iron center and alternative co-factors in the absence of molecular oxygen. For example, nitrene transfer, azidation, and fluorine transfer have all been observed with Fe(II)/αKGs under anaerobic conditions (Figure 2E-F).[8–11] Mutagenesis can also unveil previously undetectable promiscuity in Fe(II)/αKGs. For example, the lysine hydroxylase Hydrox, which natively has no detectable halogenation activity, was converted to a radical halogenase through mutagenesis and subsequent DNA shuffling (Figure 2H).[16] Halogenases utilize a 2-His iron binding site with no carboxylate ligand, leaving an open coordination site for a halogen ion that can react with the substrate carbon-centered radical to form a new carbon-halogen bond.[17] Halogenation over hydroxyl rebound is thought to be controlled through precise positioning of substrate in the active site, and this positioning can be perturbed with mutagenesis.[18] The above examples illustrate how utilizing a Fe(II)/αKG Fe(IV)-oxo intermediate, or the conserved 2-His-1-carboxylate itself, can lead to multiple reaction outcomes. Overall, the diversity of reactions possible with Fe(II)/αKGs continues to expand beyond their biological scope and this important enzyme superfamily will continue to serve as a rich new source of biocatalysts for chemical synthesis.

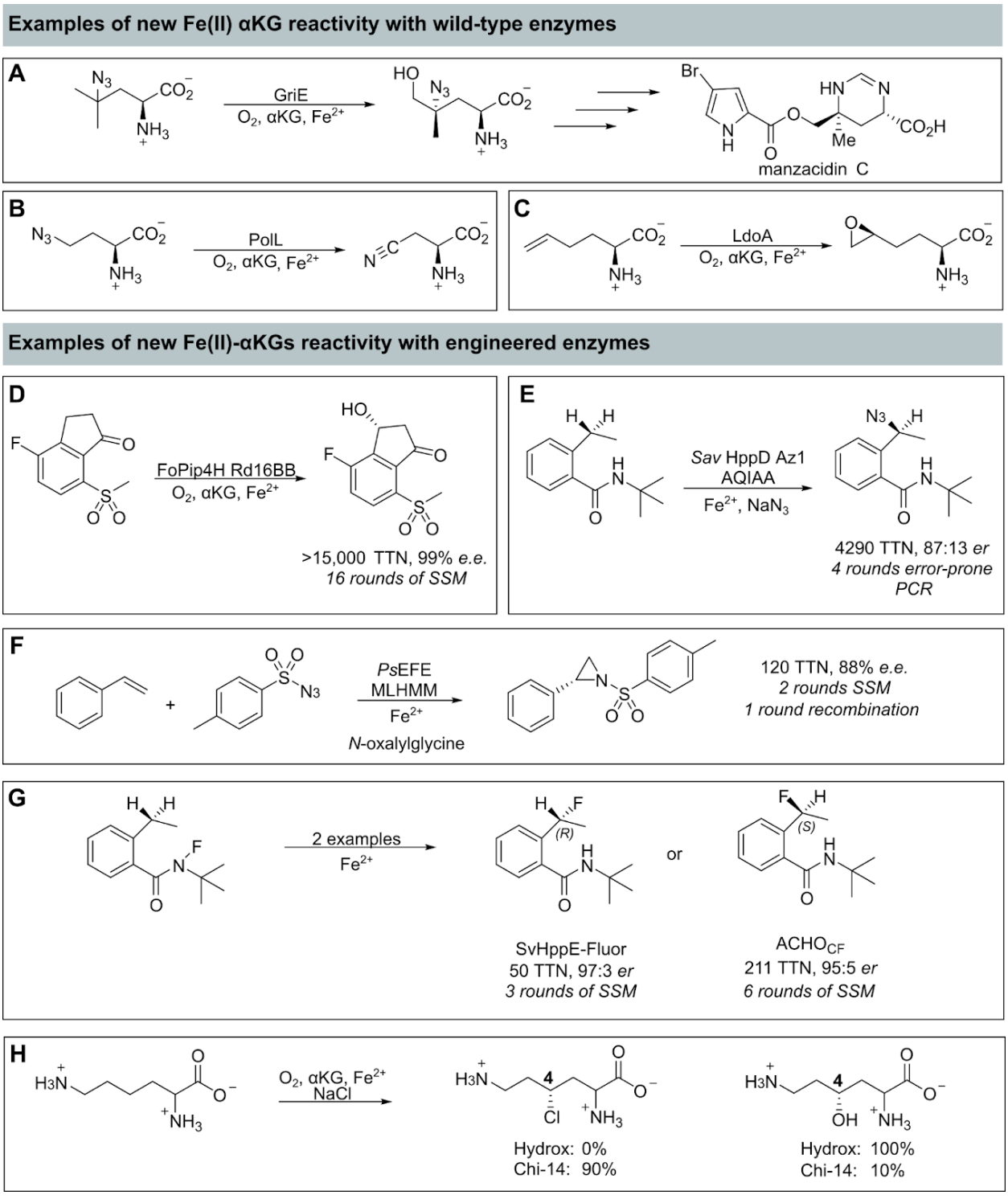


Figure 2. Select examples of novel reactions with wild-type and mutant Fe(II)/αKGs. References are in corresponding order to A-H.[8–16]

1.3 Fe(II)/αKG engineering and activity-stability tradeoffs.

While new reactions are being continuously discovered with Fe(II)/αKGs, initial reactivity from wild-type enzymes is often unsuitable for industrial processes. Diverse low level reactivity of Fe(II)/αKGs can effectively be improved through protein engineering, which is most often accomplished using directed enzyme evolution.[19] However, enzyme engineering campaigns for industrial processes are often incredibly time and resource intensive. For example, Merck recently incorporated an engineered Fe(II)/αKG hydroxylase into the synthesis of the kidney cancer drug belzutifan. To reach the target goal of a total turnover number (TTN) of >15,000, an intensive 16 rounds of site-saturation mutagenesis (SSM) across the entire enzyme structure were carried out.[15] One of key challenges in Fe(II)/αKG engineering that may require such extreme effort is our lack of knowledge of which residues in an enzyme may be important in modulating different reaction parameters. For example, we know that altering residues distal to the active site in many enzymes has a direct effect on catalysis.[20] Similarly, we know that substrate specificity is likely encoded globally in an enzyme structure, but we cannot predict unique positions to target towards altering specificity.[21] While we can partially overcome this gap in knowledge through brute-force resource-intensive screening, most often enzyme mutagenesis efforts are limited to the enzyme active site.[8,22] Because active site mutations can be destabilizing,[23,24] this approach carries the risk of activity-stability tradeoffs in Fe(II)/αKG engineering which can limit their effective use in industrial synthesis. There are several approaches for identifying or generating stable enzymes. Two common approaches are homology screening and ancestral sequence reconstruction (ASR).[25,26] Additional computational design and deep-learning based methods continue to emerge.[27,28] There is a critical need to evaluate the effectiveness of these methods as well as their generalizability and ease of use towards enzyme engineering.

1.4 Conclusions

Fe(II)/αKGs hold promise in their ability to carry out diverse new-to-nature reactions. While that promise is being realized, there are still several engineering challenges within this enzyme family that may limit their practical application in synthesis. Coupling our lack of knowledge on Fe(II)/αKG global structure-function effects, the frequent focus on active-site engineering and the prevalence of active-site stability-activity tradeoffs suggests that utilizing methods for enzyme stabilization may open the door to broad generalizability of Fe(II)/αKGs as robust biocatalysts.

References

[1] P. G. Levi, J. M. Cullen, *Environ. Sci. Technol.* **2018**, *52*, 1725–1734.
[2] J. F. Hartwig, M. A. Larsen, *ACS Central Science* **2016**, *2*, 281–292.
[3] C. Q. Herr, R. P. Hausinger, *Trends in Biochemical Sciences* **2018**, *43*, 517–532.
[4] M. S. Islam, T. M. Leissing, R. Chowdhury, R. J. Hopkinson, C. J. Schofield, *Annual Review of Biochemistry* **2018**, *87*, 585–620.
[5] S. G. Pati, C. E. Bopp, H.-P. E. Kohler, T. B. Hofstetter, *ACS Catal.* **2022**, *12*, 6444–6456.
[6] E. I. Solomon, T. C. Brunold, M. I. Davis, J. N. Kemsley, S.-K. Lee, N. Lehnert, F. Neese, A. J. Skulan, Y.-S. Yang, J. Zhou, *Chem. Rev.* **2000**, *100*, 235–350.
[7] R. Myllylä, K. Majamaa, V. Günzler, H. M. Hanauske-Abel, K. I. Kivirikko, *Journal of Biological Chemistry* **1984**, *259*, 5403–5405.
[8] N. W. Goldberg, A. M. Knight, R. K. Zhang, F. H. Arnold, *Journal of the American Chemical Society* **2019**, *141*, 19585–19588.
[9] J. Rui, Q. Zhao, A. J. Huls, J. Soler, J. C. Paris, Z. Chen, V. Reshetnikov, Y. Yang, Y. Guo, M. Garcia-Borràs, X. Huang, *Science* **2022**, *376*, 869–874.
[10] Y. Yang, L.-P. Zhao, B. K. Mai, L. Cheng, F. Gao, Y. Zhao, R. Guo, H. Wu, Y. Zhang, P. Liu, **2024**, DOI 10.26434/chemrxiv-2024-pt58m.
[11] Q. Zhao, Z. Chen, J. Soler, X. Chen, J. Rui, N. T. Ji, Q. E. Yu, Y. Yang, M. Garcia-Borràs, X. Huang, *Nat. Synth* **2024**, *3*, 958–966.
[12] C. R. Zwick, H. Renata, *Journal of the American Chemical Society* **2018**, *140*, 1165–1169.
[13] M. Davidson, M. McNamee, R. Fan, Y. Guo, W. C. Chang, *Journal of the American Chemical Society* **2019**, *141*, 3419–3423.
[14] L. Cha, S. Milikisiyants, M. Davidson, S. Xue, T. Smirnova, A. Smirnov, Y. Guo, W. Chang, *Biochemistry* **2020**, *59*, 1961–1965.
[15] W. L. Cheung-Lee, J. N. Kolev, J. A. McIntosh, A. A. Gil, W. Pan, L. Xiao, J. E. Velásquez, R. Gangam, M. S. Winston, S. Li, K. Abe, E. Alwedi, Z. E. X. Dance, H. Fan, K. Hiraga, J. Kim, B. Kosjek, D. N. Le, N. S. Marzijarani, K. Mattern, J. P. McMullen, K. Narsimhan, A. Vikram, W. Wang, J.-X. Yan, R.-S. Yang, V. Zhang, W. Zhong, D. A. DiRocco, W. J. Morris, G. S. Murphy, K. M. Maloney, *Angewandte Chemie International Edition* **2024**, *63*, e202316133.
[16] M. E. Neugebauer, E. N. Kissman, J. A. Marchand, J. G. Pelton, N. A. Sambold, D. C. Millar, M. C. Y. Chang, *Nat Chem Biol* **2022**, *18*, 171–179.
[17] M. E. Neugebauer, K. H. Sumida, J. G. Pelton, J. L. McMurry, J. A. Marchand, M. C. Y. Chang, *Nat Chem Biol* **2019**, *15*, 1009–1016.
[18] M. L. Matthews, C. M. Krest, E. W. Barr, F. H. Vaillancourt, C. T. Walsh, M. T. Green, C. Krebs, J. M. Jr. Bollinger, *Biochemistry* **2009**, *48*, 4331–4343.
[19] C. Zeymer, D. Hilvert, *Annual Review of Biochemistry* **2018**, *87*, 131–157.
[20] J. Gu, Y. Xu, Y. Nie, *Biotechnology Advances* **2023**, *63*, 108094.
[21] E. E. Wrenbeck, L. R. Azouz, T. A. Whitehead, *Nature Communications* **2017**, *8*, 15695.
[22] M. Goldsmith, D. S. Tawfik, *Methods in Enzymology* **2013**, *523*, 257–283.
[23] N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, *PLoS Comput Biol* **2008**, *4*, e1000002.
[24] E. M. Meiering, L. Serrano, A. R. Fersht, *Journal of Molecular Biology* **1992**, *225*, 585–589.
[25] H. J. Atkinson, J. H. Morris, T. E. Ferrin, P. C. Babbitt, *PLOS ONE* **2009**, *4*, e4345.

[26] L. O. Chisholm, K. N. Orlandi, S. R. Phillips, M. J. Shavlik, M. J. Harms, *Annual Review of Biophysics* **2024**, *53*, 127–146.
[27] A. Goldenzweig, S. J. Fleishman, *Annual Review of Biochemistry* **2018**, *87*, 105–129.
[28] K. H. Sumida, R. Núñez-Franco, I. Kalvet, S. J. Pellock, B. I. M. Wicky, L. F. Milles, J. Dauparas, J. Wang, Y. Kipnis, N. Jameson, A. Kang, J. D. L. Cruz, B. Sankaran, A. K. Bera, G. Jiménez-Osés, D. Baker, **2023**, 2023.10.03.560713.

Chapter 2: Computational Stabilization of a Non-heme Iron Enzyme Enables Efficient Evolution of New Function

Brianne R. King[1], Kiera H. Sumida[1,2], Jessica L. Caruso[1], David Baker[2,3], & Jesse G. Zalatan[1]

[1]Department of Chemistry

[2]Institute for Protein Design

[3]Howard Hughes Medical Institute

University of Washington, Seattle, WA 98195, USA

Published as a research article in *Angewandte Chemie, International Edition* on October 12th, 2024.

DOI: 10.1002/anie.202414705

2.1 Abstract

Deep learning tools for enzyme design are rapidly emerging, and there is a critical need to evaluate their effectiveness in engineering workflows. Here we show that the deep learning-based tool ProteinMPNN can be used to redesign Fe(II)/αKG superfamily enzymes for greater stability, solubility, and expression while retaining both native activity and industrially-relevant non-native functions. This superfamily has diverse catalytic functions and could provide a rich new source of biocatalysts for synthesis and industrial processes. Through systematic comparisons of directed evolution trajectories for a non-native, remote C(sp^3)-H hydroxylation reaction, we demonstrate that the stabilized redesign can be evolved more efficiently than the wild-type enzyme. After three rounds of directed evolution, we obtained a 6-fold activity increase from the wild-type parent and an 80-fold increase from the stabilized variant. To generate the initial stabilized variant, we identified multiple structural and sequence constraints to preserve catalytic function. We applied these criteria to produce stabilized, catalytically active variants of a second Fe(II)/αKG enzyme, suggesting that the approach is generalizable to additional members of the Fe(II)/αKG superfamily. ProteinMPNN is user-friendly and widely-accessible, and our results provide a framework for the routine implementation of deep learning-based protein stabilization tools in directed evolution workflows for novel biocatalysts.

2.2 Introduction

Directed evolution is a powerful method to generate enzymes for new chemical transformations.[1,2] However, catalytic functional groups often have destabilizing effects on protein structure, and altering active site groups for new reactions can lead to unstable, non-functional proteins.[3–11] Initiating a directed evolution campaign from a stabilized variant can be an effective way to overcome this problem.[12,13] Typically, a starting point for directed evolution is obtained by screening a library of candidate enzymes for a desired promiscuous activity. If a

thermostable homolog with similar catalytic properties is identified, it can then be used as a starting point for evolution of the desired function.[14,15] Alternatively, there are a variety of strategies to produce stable variants using directed evolution,[16,17] ancestral reconstruction,[14] protein recombination,[18,19] or computational engineering.[20–22] These methods are often time- and resource-intensive, highlighting the need for simple and accessible alternatives.

An important class of enzymes where stabilization would be useful is the non-heme iron(II) α-ketoglutarate-dependent oxygenase (Fe(II)/αKG) superfamily.[23–26] These enzymes have emerged as a rich source of potential new biocatalysts. Fe(II)/αKG enzymes can perform remote, asymmetric C(sp^3)-H oxyfunctionalization reactions on small molecule substrates using a conserved, radical-mediated mechanism. These transformations are synthetically challenging.[27,28] and a biocatalytic alternative could allow expedient and sustainable diversification of simple building blocks to a range of complex polyfunctional compounds. The advantages offered by this enzyme family include a high degree of chemical flexibility in the iron-containing active site due to multiple open coordination sites, the utilization of benign molecular oxygen as an oxidant, and use of the inexpensive and readily available co-factor αKG. However, Fe(II)/αKGs can be relatively unstable,[29,30] which may limit their practical applications in organic synthesis and in industrial process applications.

The recent use of an Fe(II)/αKG in an industrial-scale drug biosynthesis pathway highlights both the potential advantages and drawbacks of this family for biocatalysis. An engineered Fe(II)/αKG was used to catalyze an enantioselective C(sp^3)-H hydroxylation to produce a key intermediate for the anti-cancer drug belzutifan.[30] The reaction could be performed on kilogram-scale and bypassed five steps of the pre-existing chemical synthesis route. Notably, this effort required an extensive, large-scale directed evolution campaign. Furthermore, early rounds of

screening yielded stabilizing mutations before significant improvements in turnover could be obtained in later rounds. These findings highlight the potential utility of Fe(II)/αKG for practical, industrial-scale green chemistry but also the importance of enzyme stability in the evolution of new function.

Recent applications of deep learning to protein design have provided new and relatively straightforward methods to stabilize protein scaffolds,[22,31,32] and there is broad interest in applying these approaches to directed evolution.[33] Here we demonstrate that the deep learning-based tool ProteinMPNN[31,32] enables more efficient optimization of a synthetically-relevant, non-native C(sp^3)-H hydroxylation reaction in an Fe(II)/αKG family member. A critical step was identifying appropriate design criteria to prevent modification of residues important for catalytic function, which includes both active site and remote positions. With a stabilized starting point for site-saturation mutagenesis, we observed substantially larger increases in non-native activity compared to the same mutations in the wild-type parent enzyme. This systematic comparison of the wild-type parent and the stabilized redesign provides a critical benchmark for the field to evaluate the effectiveness of these tools. We suggest that this designed stabilization approach should be routinely used in future directed evolution campaigns with the Fe(II)/αKG superfamily and will likely be effective in a broad range of other enzyme families. There have been many recent reports applying machine learning tools to design variant libraries or pick residues for functional optimization,[22,33–38] and stabilized scaffolds can readily be coupled to these approaches for tailored engineering.

2.3 Results and Discussion

2.3.1 Fe(II)/αKGs with promiscuous activity for C-H hydroxylation of free carboxylate substrates

Within the Fe(II)/αKG enzyme superfamily, free amino acid hydroxylases are attractive candidates for engineering new reactions.[23–26] Because Fe(II)/αKG amino acid hydroxylases

already have catalytic machinery to interact with amine and carboxylate functional groups in amino acids, we hypothesized that they might have promiscuous activity for substrates containing only an amine or only a carboxylate. These molecules are important feedstocks for early-stage oxyfunctionalization reactions in multi-step syntheses. Selectivity for remote C(sp^3)-H hydroxylation reactions has been historically difficult to achieve with traditional transition metal catalysis, and a biocatalytic process could offer improved regio- and stereoselectivity.[27,28]

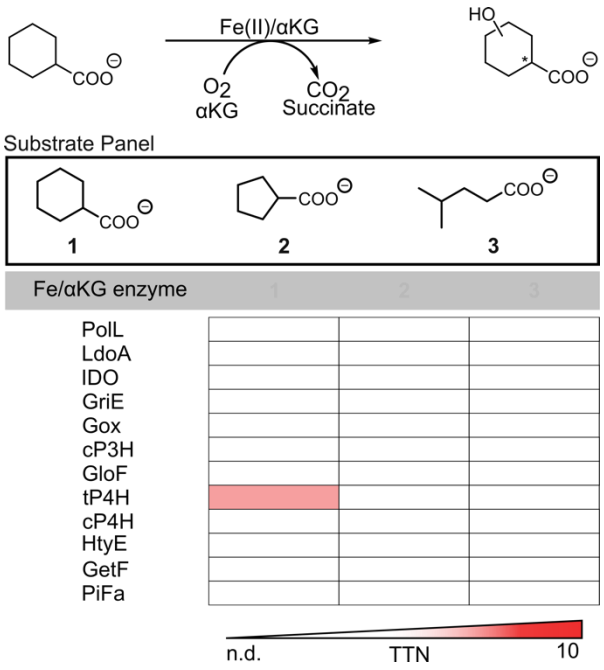


Figure 1. Initial whole-cell reaction screen data with a panel of Fe(II)/αKG amino acid hydroxylases and free acid substrate analogues. Each enzyme was screened against all three substrates. A white panel indicates that product was not detectable (n.d.), which is the case for all reactions except tP4H with substrate **1**. Reactions were performed in whole cell from 50 mL expression cultures where whole cell volume was 1/20th the expression volume. Reactions were carried out in MOPS (pH 7.0, 50 mM) with 20 mM substrate, 60 mM αKG (as disodium salt), 1 mM ferrous ammonium sulfate, and 1 mM L-ascorbic acid.

We initially screened a panel of 12 Fe(II)/αKG amino acid hydroxylases for the ability to hydroxylate free carboxylates (Figure 1). The enzymes in this panel were chosen for their ability to hydroxylate free amino acids and their ease of expression, handling, and purification (Table S6). We chose a set of candidate carboxylate substrates (**1-3**) that are structurally analogous to the

native amino acid substrates L-pipecolic acid (L-Pip), L-proline (L-Pro), and L-leucine (L-Leu). We used whole-cell biocatalysis and liquid chromatography-mass spectrometry (LC-MS) to detect products. We confirmed that the native amino acid reaction products are detectable with all 12 members of the enzyme panel (Table S6). We then screened for promiscuous activity with carboxylates, and observed that one enzyme, tP4H,[39] has detectable activity with substrate **1** (Figure 1). The total turnover number (TTN) with this substrate was ~5 after 24 hr incubation with 10 μM enzyme, ~130-fold lower than the TTN for the corresponding native amino acid substrate. The reaction of tP4H with substrate **1** gives the *trans* product with a *d.r.* of 4:1 (Figure S4). tP4H produces exclusively *trans* product with its native substrate L-Pip,[39] suggesting that the free carboxylate substrate **1** and the native substrate are positioned similarly in the enzyme active site with respect to the iron center. To confirm that tP4H and not a contaminating enzyme was responsible for the observed non-native activity, we mutated active site residues that are involved in Fe(II), substrate, or αKG binding. Because tP4H does not have an experimental structure, we identified these active site residues using an Alphafold2[40] model (Figure S9) and comparisons to structures of the highly homologous Fe(II)/αKG enzyme GriE.[41,42] In all cases, active site mutations produced activity decreases for the non-native substrate **1** (Figure 2). We also observed increased product yield with increasing wild-type tP4H concentration (Figure 2). The overall yields remain relatively low, which is typical for promiscuous, non-native reactions.[1] Together these results confirm that tP4H is responsible for the non-native reaction.

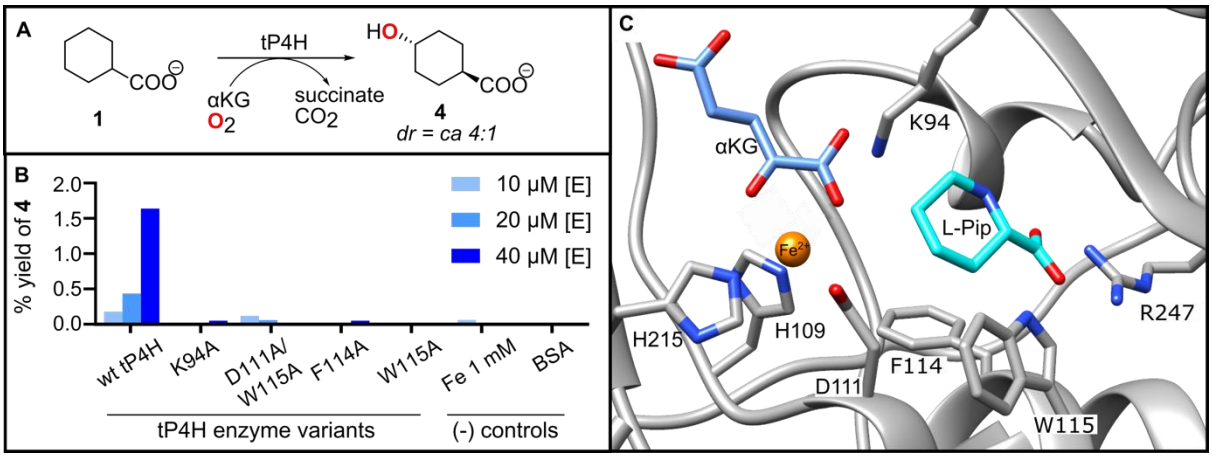


Figure 2. Validation of tP4H activity with free acid **1.** (A) Reaction of tP4H with substrate **1** to form *trans*-4-hydroxycyclohexane carboxylic acid **4**. (B) Yield of **4** after reaction of **1** with tP4H variants, as well as negative control reaction with Fe(II) and bovine serum albumin (BSA). The Fe 1 mM control was run in the absence of added enzyme. Purified enzyme concentration was varied between 10-40 μM with 20 mM **1**, 40 mM αKG, 1 mM ferrous ammonium sulfate, and 1 mM L-ascorbic acid in MES buffer (50 mM, pH 6.8). Reactions were carried out for 24 hours at 25 °C and quantified with analytical LC-MS. (C) Structural model of the tP4H showing key active site residues. Fe(II), αKG, and L-Pip were modeled in Chimera.[43]

2.3.2 Stabilization of tP4H with ProteinMPNN

To improve tP4H activity towards substrate **1**, we began a directed evolution campaign but quickly encountered limitations due to poor enzyme stability. First, we found that tP4H variants were difficult to express and purify due to enzyme insolubility. Additionally, we found that the parent wild-type tP4H enzyme loses activity with time (Figure 3). These observations are consistent with prior reports on tP4H behavior.[39] It is possible for enzyme stability to improve during directed evolution, whether by selecting more stable variants during each round or by chance. For example, the evolved Fe(II)/αKG PsEFE had slight stability improvements compared to wild-type, despite the researchers not selecting for improved stability.[44] In another case, the Fe(II)/αKG UbP4H was successfully screened for improved stability after initial screening rounds destabilized the enzyme.[29] However, the benefits of starting with a highly stable enzyme for

directed evolution are well-established.[12–14] A stabilized protein scaffold could potentially increase the population of active, properly-folded protein or provide access to other mutants that otherwise do not fold.

We used the deep learning-based tool ProteinMPNN[31,32] to generate stabilized variants of tP4H through sequence design. We first used ProteinMPNN to redesign the entire tP4H sequence. Unsurprisingly, we found that the predicted sequences eliminated key active site residues, which is likely to disrupt enzymatic activity (Figure S10). This behavior is consistent with the well-established propensity for catalytic active site residues to be destabilizing.[3–10] To preserve catalytic function, we fixed the active site residues in all subsequent design efforts. We defined the active site as any residues that contact the amino acid substrate, Fe(II), or the αKG cofactor, based on our Alphafold2 model (Figure S9) and comparisons to the structure of the closely-related enzyme GriE bound to L-Leu.[42] Because other residues throughout the protein could also be important, we also tested four additional strategies using either sequence conservation or distance metrics. To identify important conserved residues, we constructed a multiple sequence alignment (MSA) and selected tP4H residues conserved in at least 35%, 70%, or 95% of sequences (Methods and Table S4). Alternatively, we fixed any residues with side chains within a 10 Å sphere from the substrate binding pocket. Using these five starting points (fix active site only, active site + 35%/70%/95% conservation, active site + 10 Å sphere), we generated 48 ProteinMPNN sequences per method and selected 4 each (20 total) for activity screens. Selection was based on calculated top-ranked Cα-RMSD values matched to the input tP4H structure. We obtained only one variant that had any detectable activity, with catalytic efficiency ~35-fold lower than wild-type tP4H (Figure S11). This variant was designed from the sequence where >35% conserved residues are fixed, which constrains more residues than the >70% or >95% cutoffs. This

result suggests that even weakly conserved residues may need to be fixed to maintain activity. Further analysis of the variant with detectable activity revealed a ~2-fold increase in the K_M for the αKG cofactor and a ~3-fold decrease in k_{cat} (Figure S11). We identified two residues in proximity to αKG, L228 and V230, that were mutated in the redesigned sequence. These sequence changes may have contributed to improved stability at the expense of cofactor binding and positioning, leading to the decrease in activity. Notably, L228 and V230 were fixed in the designs generated from fixed active site + 10 Å sphere, but none of these designs had detectable activity. Taken together, these findings suggest that additional criteria will be needed to identify critical functional residues that should be fixed prior to sequence redesign.

To generate stabilized variants that maintain catalytic activity, we performed another set of ProteinMPNN sequence redesigns with three new strategies to fix important residues. In each case, we fixed the active site as defined above plus residues L228 and V230. For the first approach, we fixed all residues at tP4H positions conserved in 35% of the MSA. We chose this cutoff because it was the only one from our initial set that produced a stabilized variant with any detectable activity, and we expected that fixing L228 or V230 could further improve these designs. For the second and third approaches, we identified highly conserved positions regardless of whether the wild-type tP4H residue is the most highly conserved amino acid. These strategies were based on previous work suggesting that more stringent constraints are necessary to maintain activity in ProteinMPNN redesign.[32] Every tP4H amino acid position was ranked based on the % conservation of the most frequent amino acid present in the MSA, and the top 50% or 70% were fixed. These positions were fixed as the wild-type tP4H residue, even if they were different from the consensus most frequent amino acid in the MSA. Together with fixed active site residues, these criteria resulted in 148/272 (54%) or 198/272 (73%) fixed residues across the entire 272 amino

acid protein. Using these three strategies, we selected 32 designs each of 48 generated for a total of 96 sequences (Supplementary spreadsheet – ProteinMPNN sequences_metrics). Of these designs, 69 expressed detectable quantities of protein by SDS-PAGE and 11 had detectable activity above background for the native substrate. For the active enzymes we proceeded to measure thermostability and kinetic parameters for the native L-Pip substrate and the promiscuous carboxylate substrate **1**. The variant with the highest k_{cat} for L-Pip was R2_11 (Table S7), with a k_{cat} of 0.10 s^{-1} compared to 0.14 s^{-1} for wild-type. R2_11 was designed from the method where the top 70% ranked conserved residues were fixed, and had 44 designed mutations compared to the wild-type sequence (Figure 3). R2_11 has modestly slower (~3-fold) non-native carboxylate hydroxylase activity compared to wild-type tP4H and exhibits an 11 °C increase in thermal melting temperature (T_m) as measured by temperature-dependent circular dichroism (CD) spectroscopy (Figure 3). When activity is measured as a function of time, R2_11 maintains activity over a timescale of days, which is a substantial improvement compared to wild-type tP4H (Figure 3). The modest decrease in promiscuous activity is unsurprising because ProteinMPNN does not consider catalytic activity, and there is no expectation that global protein stabilization would either maintain or optimize a non-native reaction.

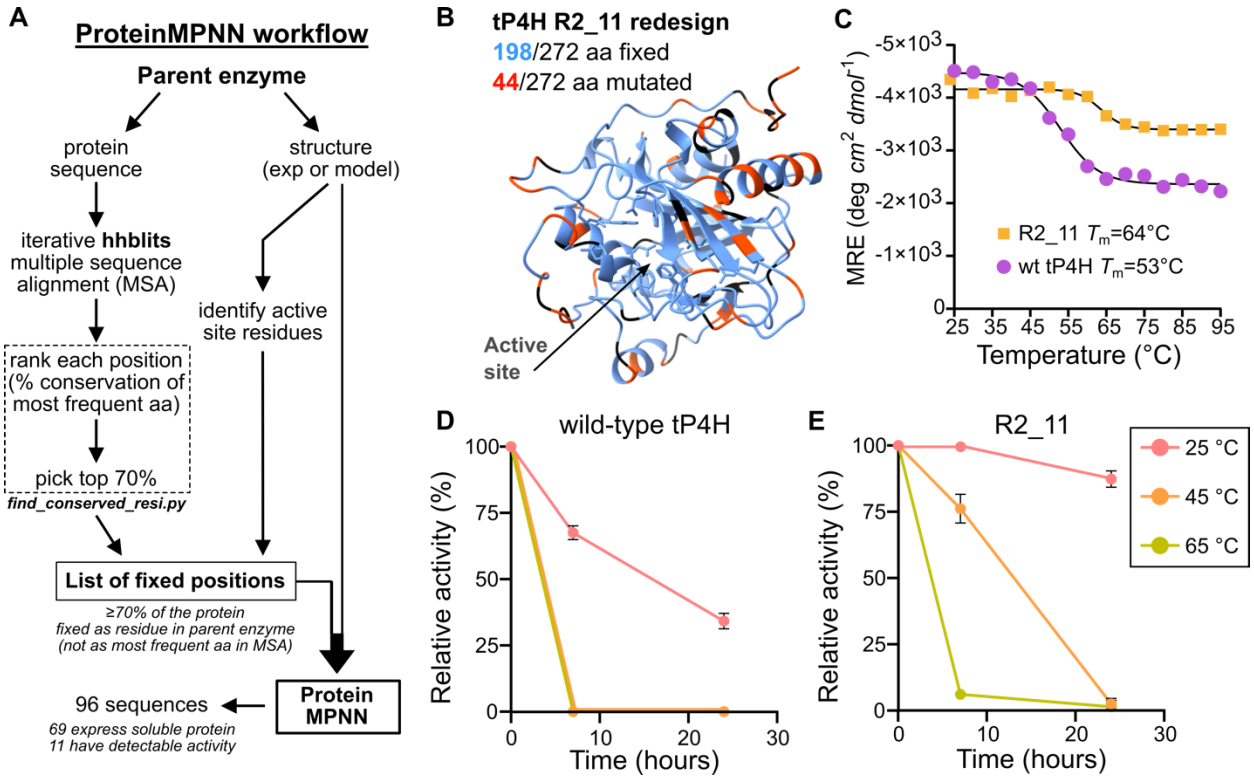


Figure 3. Stability and activity of wild-type tP4H and ProteinMPNN design R2_11. (A) Flowchart of the ProteinMPNN workflow that successfully produced stabilized variants that retained catalytic activity. See Supporting Information for detailed guidelines for each computational step. (B) tP4H structure (Alphafold2 model, Figure S9) color-coded to show sites fixed in the design process (blue, Supplementary spreadsheet – ProteinMPNN sequences_metrics) and sites mutated in the ProteinMPNN R2_11 redesign (orange-red). Sites colored black were neither fixed nor redesigned in the R2_11 variant. Side chains for first shell active site residues (Table S4) are shown in blue. (C) Temperature-dependent CD spectroscopy of wild-type tP4H and R2_11. T_m values were calculated using the Boltzmann sigmoid function in GraphPad Prism. (D) Activity-stability analysis of wild-type tP4H. (E) Activity-stability analysis of R2_11. For D and E, relative activity was determined using PBP assay described in Supplementary Information. Values are mean ± SD for three replicates.

2.3.3 Stabilization of GriE with ProteinMPNN

After successfully identifying sequence constraints for ProteinMPNN-mediated stabilization of tP4H while maintaining catalytic function, we evaluated whether the same approach would be effective with a second Fe(II)/αKG amino acid hydroxylase, GriE. This enzyme could benefit from stabilization because, although it expresses well and is soluble, it loses

activity at room temperature over 24 hours (Figure S12). As with tP4H, we fixed the top 70% ranked conserved residues along with catalytic residues identified from the GriE crystal structure (Supplementary spreadsheet – ProteinMPNN sequences_metrics).[42] We generated 32 redesigned sequences and found that 29 were expressed as soluble enzyme, and 27 showed activity with the GriE native substrate L-Leu. The top design based on stability and kinetic parameters, GM_A9, showed a similar catalytic efficiency (k_{cat}/K_M) and a ~4-fold decrease in k_{cat} compared to wild-type GriE (Figure S12 & S13, Table S8). One design, GM_A11, had a 2-fold faster initial rate with L-Leu compared to GM_A9 but this design was unstable by temperature-dependent CD and thus was not chosen for further analysis. A decrease in k_{cat} is not surprising given that increased stability could reduce conformational flexibility and negatively impact catalytic function.[4,6,7]

We next screened the stabilized GriE redesign GM_A9 for substrate promiscuity. Previously, wild-type GriE has been shown to accept substrates with increased substrate chain lengths but has weaker activity towards substrates with substitution at C3.[45] We chose two previously identified non-native substrates to test: L-norleucine (L-Nle) and L-allo-isoleucine (L-allo-Ile). L-Nle was chosen as a representative substrate with increased chain length compared to L-Leu. L-allo-Ile was chosen because it has a methyl group substitution at C3. L-isoleucine also has a C3 methyl group, but it is not detectably hydroxylated by wild-type GriE[45] and was therefore not included in this analysis. We observed detectable activity with L-Nle but not for L-allo-Ile (Table S9). Similar to wild-type GriE, the GM_A9 variant maintained a preference for the extended chain L-Nle substrate over the C3-substituted L-allo-Ile substrate. The GM-A9 reactions with L-Leu and L-Nle were 11- and 4-fold slower than wild-type GriE reactions, respectively (Table S9). These results suggest that our ProteinMPNN protocol can be readily applied to other

Fe(II)/αKG enzymes to stabilize proteins while maintaining synthetically-relevant catalytic function that can be a foothold for further optimization by directed evolution.

2.3.4 Directed evolution of wild-type tP4H for carboxylate C-H hydroxylation activity

We next sought to improve the non-native carboxylate C(sp^3)-H hydroxylase activity through directed evolution. We prioritized tP4H because carboxylate hydroxylase activity was detectable in both the wild-type and ProteinMPNN-stabilized variant, which allows for direct comparisons. We conducted three rounds of directed evolution for both enzymes by varying first- and second-shell substrate binding residues identified in the active site from our Alphafold2 structural model. We defined the first shell as any residues that contact the amino acid substrate, based on comparisons to the structure of ligand-bound GriE.[42] We defined the second shell as any residues that make contacts with first shell residues. We used the 22c-trick method for single site-saturation mutagenesis at each target position, and we screened 70 colonies for each position to ensure >95% library coverage (Table S1).[46]

We first performed directed evolution with wild-type tP4H. For the first screening round, we chose three tP4H active site residues based on their potential role in substrate specificity: H58, F114, and L174. Based on our tP4H structural model (Figure S9B), H58 likely contacts the amine of native amino acid substrates and is presumably not needed or detrimental for carboxylate substrates that lack an amine. F114 likely provides a substrate hydrophobic contact, and L174 is part of a loop that could affect substrate binding. We screened whole cell biocatalysis reactions in 96 well plates for improved TTN and 80% *trans* selectivity in reactions with substrate **1**. Based on production of hydroxylated product **4**, the top 5% of mutants were chosen for validation with purified enzymes. We obtained several variants with modest activity improvements, and the best performer was mutant H58L with a TTN of 7 (Figure 4A). In a second round starting from H58L,

we rescreened mutants at F114 and L174 and screened an additional 14 first and second shell residues (Table S1). We identified the improved variant H58L/W170Q with a TTN of 15. In a third round starting from H58L/W170Q, we screened 9 residues that showed activity increases in previous rounds and identified the improved variant H58L/W170Q/E118K with a TTN of 31 (Figure 4A & Table S1). Overall, after three rounds of directed evolution for improved carboxylate hydroxylase activity with wild-type tP4H we obtained a 6-fold improvement in TTN and maintained >80% selectivity for the *trans* reaction product (Figure S4).

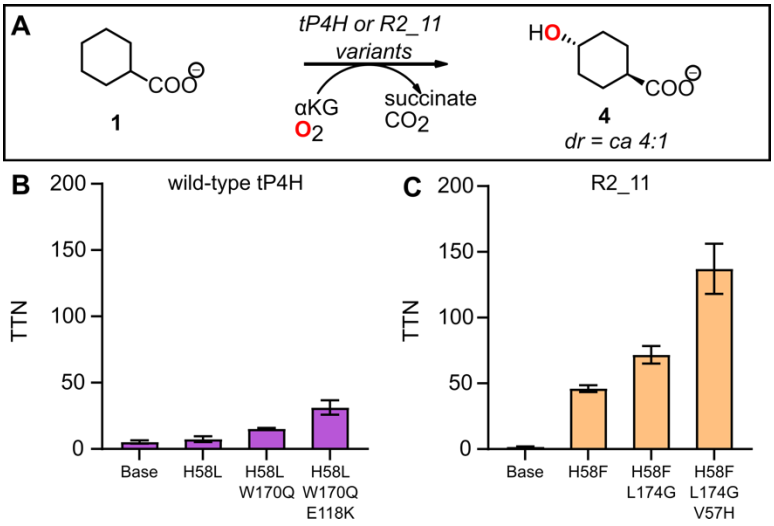


Figure 4. (A) Reaction scheme for C-H hydroxylation of substrate **1** with tP4H, R2_11, and associated variants. (B) Directed evolution of wild-type tP4H. (C) Directed evolution of stabilized variant R2_11. Reactions were carried out for 24 hr at 25 °C using purified enzyme (10-20 μM) in MES buffer (50 mM, pH 6.8), with 20 mM cyclohexane carboxylic acid **1**, 40 mM αKG, 1 mM ferrous ammonium sulfate, and 1 mM ascorbic acid. Concentration of **4** in quenched reaction samples was quantified by analytical LC-MS analysis. For B and C, values are mean ± SD for three replicates.

2.3.4 Directed evolution of ProteinMPNN-stabilized tP4H for carboxylate C-H hydroxylation activity.

To optimize carboxylate hydroxylase activity in ProteinMPNN-stabilized tP4H, we conducted directed evolution using a similar strategy to our approach with wild-type tP4H, with minor modifications. In the first round of site saturation mutagenesis, we started with a larger pool of 19 first- and second-shell residues including the three sites from prior round one (H58, F114, and L174) and the 16 additional sites from prior round 2 (Table S1). As before, we screened whole cell biocatalysis reactions for improved TTN and >80% *trans* selectivity with carboxylic acid substrate **1**. The top hit was H58F, which is the same position but a different mutant than the previous round one winner, H58L. R2_11_H58F displayed a 27-fold increase in TTN relative to the parent R2_11 (Figure 4B). This effect is substantially bigger than the <2-fold improvement obtained with H58L relative to wild-type tP4H. Notably, the 27-fold increase in a single round from stabilized R2_11 was already larger than the total 6-fold improvement from three rounds of directed evolution from wild-type tP4H. Given the strong performance of the H58F mutant, we also evaluated its effect in the wild-type tP4H background and observed a small, <2-fold increase in TTN, similar to the effect of H58L on wild-type tP4H (Figure S14). Thus, the strong, 27-fold improvement with the H58F mutant depends on the context of the stabilized R2_11 backbone. Context-dependent activity increases have previously been observed in stabilized variants, supporting the general idea that stability can promote evolvability.[12]

In a second round of screening from R2_11_H58F, we selected the 18 residues that were screened in prior round two from wild-type tP4H (Table S1). We retained this large pool of residues to ensure a direct comparison to the tP4H directed evolution workflow. This round identified improved variant L174G. R2_11_H58F/L174G has a TTN of 72. This TTN is a 1.6-fold improvement from parent R2_11_H58F and outstrips any variant obtained from the wild-type tP4H backbone (Figure 4B). In a third round of screening from R2_11_H58F/L174G, we selected

9 residues that were screened in prior round 3 from wild-type tP4H (Table S1). This round identified the improved variant V57H with a TTN of 138, a 1.7-fold improvement from the previous round.

Overall, the ProteinMPNN-stabilized tP4H directed evolution campaign produced an 80-fold improvement in TTN from the base R2_11 redesign, compared to a modest 6-fold improvement in the wild-type tP4H evolutionary trajectory. Although the R2_11 parent starts ~3-fold slower than wild-type tP4H, the much larger improvement over three rounds of directed evolution produced an R2_11 triple mutant with a 4.5-fold higher TTN than the triple mutant obtained from wild-type tP4H (Figure 4).

In addition to a more efficient directed evolution trajectory, the R2_11 triple mutant maintains high stability relative to the triple mutant derived from wild-type tP4H (Figure 5A, Figure S15), with only a modest decrease in thermal stability compared to the R2_11 parent. Higher stability allows reactions to be run more efficiently, both at higher temperatures and for less time. For example, after 6 hours at 35 °C, the R2_11 triple mutant reaches a mean TTN of 142 for the non-native reaction with carboxylate **1** to form product **4** with 4:1 selectivity for the *trans* reaction product (Figure 5B). The TTN after 6 hours at 35 °C is comparable to the TTN after 24 hours at 25 °C. In contrast, the tP4H triple mutant shows a slight decrease in TTN at 35 °C, likely due to enzyme instability at higher temperatures (Figure 5B). The stability profile of the R2_11 triple mutant suggests that this enzyme will be more robust towards further engineering compared to the tP4H triple mutant. Future engineering efforts with the R2_11 mutant could include improvements to key reaction metrics like turnover, selectivity, and increased substrate scope.

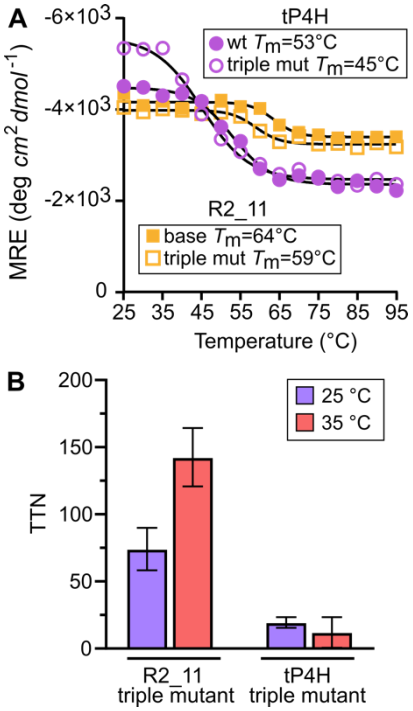


Figure 5. (A) Temperature dependent CD of the R2_11 and tP4H parent enzymes and triple mutants. T_m values were calculated using the Boltzmann sigmoid function in GraphPad Prism. (B) TTN for formation of **4** (*d.r.* 4:1, Figure S4) with the R2_11 and tP4H triple mutants at two different temperatures. Reactions were carried out for 6 hrs. at 25 °C and 35 °C using purified enzyme (15 μM) in MES buffer (50 mM, pH 6.8), with 20 mM cyclohexane carboxylic acid **1**, 40 mM αKG, 1 mM ferrous ammonium sulfate, and 1 mM ascorbic acid. Values are mean ± SD for three replicates.

2.4 Conclusions

Directed evolution is a powerful tool to engineer enzymes for new-to-nature reactions. However, many enzyme starting points for evolution may lack the stability required to reach user-defined optimum fitness after multiple rounds of mutagenesis. Here we show that the deep learning-based tool ProteinMPNN can be used to stabilize the Fe(II)/αKG enzyme superfamily members tP4H and GriE with straightforward sequence constraints to maintain catalytic activity. Consistent with previous results using ProteinMPNN,[32] the top tP4H design was identified by using the most conservative of our chosen methods for fixing residues during sequence redesign.

Applying the same method to the related enzyme GriE readily produced stabilized variants with catalytic activity.

Wild-type and redesigned tP4H both exhibit novel reactivity towards remote C(sp^3)-H hydroxylation of a free carboxylic acid substrate. We directly compared evolutionary trajectories of wild-type tP4H with the stabilized variant R2_11 and demonstrated superior performance of the stabilized redesign variant. Future work will determine if this design method is generalizable to optimize directed evolution for other enzymes and enzyme families. Further improvements to deep learning models, or an improved understanding of the underlying mechanisms for enzyme stabilization, could be necessary for broad generalizability. Additional systematic comparisons will also be necessary to evaluate ProteinMPNN relative to other methods for enzyme stabilization. For example, PROSS[20,47] uses physics-based energy calculations to generate stabilized sequences, and MutCompute[22,48] uses a deep learning approach to identify individual point mutations. Both approaches are distinct from the complete sequence redesign produced from ProteinMPNN. Directly comparing the ability and efficiency for each approach to generate stable variants for directed evolution could identify tradeoffs and potential advantages for each method. User-friendly computational tools are rapidly emerging, and our work suggests that these tools should be routinely incorporated into enzyme engineering workflows to efficiently optimize catalytic fitness for new biocatalysts.[33]

Supporting Information

Supporting Information includes materials, experimental and analytical methods, compound characterization data (Figure S4–S7), and enzyme characterization data (Figure S3, Figure S8, Figure S11–S15 and Tables S6–S9) (PDF). A supplementary spreadsheet includes accession numbers for all Fe(II)/αKG enzymes used in this work, full nucleotide and amino acid sequences for all reported wild-type and enzyme variants, oligonucleotide sequences for cloning and for all ProteinMPNN designs, ProteinMPNN design screen criteria results (XLSX). The supplementary file find_conserved_resi.txt contains the python script to parse multiple sequence alignments. The authors have cited additional references within the Supporting Information.[49-57]

Acknowledgments

We thank Dr. Wolfgang Hüttel at the University of Freiburg and Hans Renata at Rice University for the donation of the wild-type tP4H and GriE expression vectors, respectively, and for their advice on tP4H and GriE reactions in various formats. We also thank Jonathan Zhang and Susanna Vazquez Torres for their early contributions to Fe(II)/αKG reaction screening, Jue Wang for assistance with the python script to parse sequence alignments, and Dr. Martin Sadilek in University of Washington Mass Spectrometry Facility for his continued support and helpful advice in analytical method development. This work was supported by U.S. National Institutes of Health grants T32GM008268 (B.R.K., J.L.C.) and R35GM124773 (J.G.Z.), and by the Open Philanthropy Project Improving Protein Design Fund (K.H.S., D.B.).

Conflict of Interests

The authors declare no conflict of interest.

2.5 Supporting Information

Computational Stabilization of a Non-heme Iron Enzyme Enables Efficient Evolution of New Function

Authors

Brianne R. King, Kiera H. Sumida, Jessica L. Caruso, David Baker, and Jesse G. Zalatan*

Materials and Methods

Starting materials and product markers

All starting materials and product markers were used as received from commercial suppliers, unless otherwise noted.

Analytical instrumentation

HPLC-UV-MS analysis for native enzyme reactions after Fmoc-Cl derivatization was performed with an Agilent 1100 LC equipped with a Zorbax C18 column (4.6 x 100 mm, 2.1 μm ID) with DAD detection connected to a Bruker IonTrap MS. HPLC-MS analysis for site-saturation mutagenesis library screening was performed on a Waters Xevo-TQ equipped with a Waters BEH-Amide column (2.1 x 50mm 1.7 μm, plus guard column). Continuous fluorescence measurements were collected on a PerkinElmer plate-reader. Circular dichroism measurements were collected with a JASCO J-1500 instrument using a 1 mm pathlength cuvette.

General procedure for FMOC-Cl derivatization of Fe/αKG amino acid substrates and products

To 100 μL of quenched biocatalytic reactions was added 100 μL of 200 mM sodium borate (pH 8.0) and 20 μL of 10 mM 9-fluorenylmethyl chloroformate (Fmoc-Cl) in acetone. After vortexing for 60 seconds, 20 μL of 100 mM 1-adamantanamine solution in acetone was added to stop the reaction followed by vortexing for 60 seconds. 50 μL of the resultant derivatized mixture was added to a 50:50 mixture of water and acetonitrile and submitted for LC/MS analysis. LC/MS analysis for derivatized amino acid products was carried out using an Agilent 1100 LC connected to a DAD detector and a Bruker IonTrap mass spec in positive mode. Separation was carried out on a Zorbax C18 column (4.6 x 100 mm, 2.1 μm ID) with an acid modified ACN/water gradient.

DNA and protein sequences for wild-type and base MPNN designs

Materials

DNA oligonucleotides were purchased from IDT. PCR was carried out using the Phusion® High-Fidelity PCR kit (New England Biolabs). Cloning of DNA fragments into linearized vectors was carried out using In-Fusion Snap Assembly Master Mix (Takara). All DNA and amino acid sequences used in this work are supplied in a supplementary Excel document.

Plasmids and Proteins

tP4H sequences

The gene encoding *Dactylosporangium sp.* L-proline-4-hydroxylase *tP4H* (Uniprot ID O06499) was contained within a pET-28a(+) vector with an N-term 6xHis-tag. This plasmid, hereby referred to as pET28a-tP4H, was generously donated by Prof. Wolfgang Hüttel's lab at the University of Freiburg, Germany.[1] ProteinMPNN designs were purchased as IDT eBlocks that were cloned directly into a pET-28a(+) vector backbone between NcoI and XhoI for initial screening. All other DNA sequences are supplied in a supplementary Excel spreadsheet. For MPNN design directed evolution, the gene encoding R2_11 was cloned into a pET-22b(-) vector with carbenicillin resistance between XbaI and XhoI.

tP4H DNA sequence and In-Fusion overhangs for pET-28a(+).15 bp In-Fusion overhangs are bolded, and restriction sites are highlighted in red:

5′aggagatataccatgggcagcagccatcatcatcatcacagcagcggcctggtgccgcg
cggcagccatatggctagcatgactggtggacagcaaatgggtcgggatccgaattcgagctcc
gtcgacggtatcgataagcttgatatcgaattcctgcagcccgacATGCTGACCCCGACCGAAC
TGAAACAGTACCGTGAAGCTGGTTACCTGCTGATCGAAGACGGTCTGGGTCCGCGTGAAGTTGA
CTGCCTGCGTCGTGCTGCTGCTGCTCTGTACGCTCAGGACTCTCCGGACCGTACCCTGGAAAAA
GACGGTCGTACCGTTCGTGCTGTTCACGGTTGCCACCGTCGTGACCCGGTTTGCCGTGACCTGG
TTCGTCACCCGCGTCTGCTGGGTCCGGCTATGCAGATCCTGTCTGGTGACGTTTACGTTCACCA
GTTCAAAATCAACGCTAAAGCTCCGATGACCGGTGACGTTTGGCCGTGGCACCAGGACTACATC
TTCTGGGCTCGTGAAGACGGTATGGACCGTCCGCACGTTGTTAACGTTGCTGTTCTGCTGGACG
AAGCTACCCACCTGAACGGTCCGCTGCTGTTCGTTCCGGGTACCCACGAACTGGGTCTGATCGA
CGTTGAACGTCGTGCTCCGGCTGGTGACGGTGACGCTCAGTGGCTGCCGCAGCTGTCTGCTGAC
CTGGACTACGCTATCGACGCTGACCTGCTGGCTCGTCTGACCGCTGGTCGTGGTATCGAATCTG
CTACCGGTCCGGCTGGTTCTATCCTGCTGTTCGACTCTCGTATCGTTCACGGTTCTGGTACCAA
CATGTCTCCGCACCCGCGTGGTGTTGTTCTGGTTACCTACAACCGTACCGACAACGCTCTGCCG
GCTCAGGCTGCTCCGCGTCCGGAATTCCTGGCTGCTCGTGACGCTACCCCGCTGGTTCCGCTGC
CGGCTGGTTTCGCTCTGGCTCAGCCGGTTTAAgatgggggatccactagttctagagcggccgc
a**cctcgagcaccaccacc 3'**

tP4H protein sequence:

MLTPTELKQYREAGYLLIEDGLGPREVDCLRRAAAALYAQDSPDRTLEKDGRTVRAVHG
CHRRDPVCRDLVRHPRLLGPAMQILSGDVYVHQFKINAKAPMTGDVWPWHQDYIFWA
REDGMDRPHVVNVAVLLDEATHLNGPLLFVPGTHELGLIDVERRAPAGDGDAQWLPQL
SADLDYAIDADLLARLTAGRGIESATGPAGSILLFDSRIVHGSGTNMSPHPRGVVLVTYNR
TDNALPAQAAPRPEFLAARDATPLVPLPAGFALAQPV*

272 amino acids

R2_11 base DNA sequence and In-Fusion overhangs for pET-28a(+). 15 bp In-Fusion overhangs are bolded, and restriction sites are highlighted in red:

5′aggagatataccatggcagcagccatcatcatcatcacagcagcggcctggttggcag
cATGCTGACCGATGAAGAACTGAAACGCTATAACGAACTGGGCTATCTGCTGATTGAAGATGGC
CTGGGCCCGGAAGAAGTGGCAGTTTTACGTCGCGCGGCGGATGAACTGTTTGCGGAAGATAGCC
CGGATCGCACCCTGGAGAAAGATGGCGTTACCGTGCGCAGCGTGCATGGTTGCCATCGTCGTAA
CCCAGTGTGCGCAGATTTAGTTCGTCATCCGCGTCTGCTGGGTCCGGCGCAGCAGATTTTAAGC
GGCGAAGTGTATGTGCATCAGTTTAAAATTAACGCCAAAGCGCCGATGCGTGGCGATGTGTGGC
CGTGGCATCAGGATTATATCTTTTGGAACCGCGAAGATGGCATGGATAAACCGCATGTGGTGAA
CGTGGCGGTGCTGCTGGATGAAGCGACCCATCTGAACGGCCCGTTACTGTTTGTGCCGGGCACC
CATGAACTGGGTGAAATTGATGTTGCGCGCCGTAATCCACCGGGCGATGGTCCGGATCAATGGT
TACCGCAGCTGAGCGCGGATCTGGATTATGCGATTGATGATGATCTGCTGGCGACCCTGACCGA
TGGTCGCGGCATTGATTCTGCAACCGGCAAAGCGGGCAGCATTCTGCTGTTTGATAGCCGCATT
GTGCATGGCAGTGGTCGCAACATGAGCCCATTTCCGCGTCGTGTGGCGCTGGTGACCTATAACC
GCACCGATAACGCCTTACCGGAACAGGATAATCCGCGCCCGGAATTTCTGGCAGCGCGTGATGC
AACCCCATTAACCCCGCTGCCGGAAGGCTTTCGTCTGGCGGATCCACCGTAA**ctcgagcaccac
cacc 3′**

R2_11 base protein sequence:

MLTDEELKRYNELGYLLIEDGLGPEEVAVLRRAADELFAEDSPDRTLEKDGVTVRSVHG
CHRRNPVCADLVRHPRLLGPAQQILSGEVYVHQFKINAKAPMRGDVWPWHQDYIFWN
REDGMDKPHVVNVAVLLDEATHLNGPLLFVPGTHELGEIDVARRNPPGDGPDQWLPQLS
ADLDYAIDDDLLATLTDGRGIDSATGKAGSILLFDSRIVHGSGRNMSPFPRRVALVTYNRT
DNALPEQDNPRPEFLAARDATPLTPLPEGFRLADPP*

272 amino acids

GriE sequences

The gene encoding *Streptomyces sp.* L-leucine hydroxylase *GriE* (Uniprot ID A0A0E3URV8) was contained within a pET-22b(-) vector with a C-term 6xHis-tag. This plasmid, hereby referred to as pET22b-GriE, was generously provided by Prof. Hans Renata (Rice University, USA).[2] ProteinMPNN designs were purchased as IDT eBlocks that were cloned directly into a pET-22b(+) vector backbone between XbaI and XhoI. All other DNA sequences are supplied in a supplementary Excel spreadsheet.

GriE DNA sequence and In-Fusion overhangs for pET-22b(-).15 bp In-Fusion overhangs are bolded, and restriction sites are highlighted in red:

5′acaattcccctctagaaataatttgtttaactttaagaaggagatatacatATGCAGCTGA
CCGCCGATCAGGTGGAAAAATATAAATCCGATGGATACGTTCTGCTTGAAGGCGCTTTCAGCCC
AGAAGAGGTTCACGTCATGCGTCAGGCACTGAAAAAGGATCAAGAAGTCCAAGGACCTCATCGG
ATTTTAGAAGAAGATGGTCGCACGGTCCGCGCACTGTATGCTAGTCACACAAGACAAAGCGTAT
TCGATCAGTTGTCTCGCTCTGACAGATTACTGGGACCCGCCACACAGCTGTTAGAGTGCGACTT
ATACATTCACCAATTCAAGATTAATACTAAGCGCGCTTTTGGTGGAGATAGCTGGGCATGGCAC
CAGGACTTTATCGTGTGGCGCGATACTGACGGTTTACCTGCCCCGCGTGCCGTCAATGTCGGCG
TCTTTTTATCGGACGTGACAGAGTTTAATGGGCCAGTGGTCTTTTTATCTGGCTCCCACCAGCG
TGGTACAGTGGAGAGAAAGGCACGGGAGACGTCACGCTCAGACCAGCATGTGGACCCGGATGAT
TATTCTATGACGCCAGCTGAGCTGAGCCAAATGGTGGAAAAACATCCAATGGTCTCCCCAAAGG
CCGCAAGCGGTTCTGTAATGTTGTTCCACCCAGAGATAATACACGGATCGGCACCAAACATCTC
GCCGTTTGCTCGTGACCTGTTGATCATTACATACAACGACGTGGCGAACGCCCCGAAACCGGCA
GGAGAACCACGCCCGGAATACGTTATCGGTCGTGATACTACGCCACTGGTGTCTCGCTCGGGAC
CATTACACGAAGCAGCGGAGAGCCGGCTTGCC**ctcgag**caccaccacc 3′

GriE protein sequence:

MQLTADQVEKYKSDGYVLLEGAFSPEEVHVMRQALKKDQEVQGPHRILEEDGRTVRAL
YASHTRQSVFDQLSRSDRLLGPATQLLECDLYIHQFKINTKRAFGGDSWAWHQDFIVWR
DTDGLPAPRAVNVGVFLSDVTEFNGPVVFLSGSHQRGTVERKARETSRSDQHVDPDDYS
MTPAELSQMVEKHPMVSPKAASGSVMLFHPEIIHGSAPNISPFARDLLIITYNDVANAPKP
AGEPRPEYVIGRDTTPLVSRSGPLHEAAESRLA

270 amino acids

GM_A9 sequence and In-Fusion overhangs for pET-22b(-).15 bp In-Fusion overhangs are bolded, and restriction sites are highlighted in red:

5′acaattcccctctagaaataatttgtttaactttaagaaggagatatacatATGCAGCTGA
CCGATGCGCAGGTGGAACAGTATAAAAGCGATGGCTATGTGCTGCTGGAAGGCGCGTTTAGCCC
GGAAGAAGTGGATATTATGCGTAAAGCGCTGGCGAAAGATGCGGAAGTTGAAGGCCCGCATCGC
ATTATGGAAGAAGATGGCAAAGCGGTGCGCGCGCTGTATGCGAGCCATAAACGCCAGAGCGTGT
TTGATCAGCTGAGCCGCAGCGATCGCCTGCTGGGCCCGGCGACCCAGCTGCTGGAATGCGATCT
GTATATTCATCAGTTTAAAATTAACACCAAACGCGCGTTTGGCGGCAGCGCGTGGGCGTGGCAT
CAGGATTTTATTGTGTGGCGCGATACCGATGGCCTGCCGGCGCCGCGCGCGGTGAACGTTGGCG
TGTTTCTGAGCGATGTGACCGAATTTAACGGCCCGGTGGTGTTCCTGAGCGGTAGCCATCAGAA
AGGCACCCTGGAACGCAAACGTCGCGCGACCAGCGTGAGCGATGAACATGTGGATCCGCGCGAT
TATAGCATGACCCCGGCGGAACTGGAGAAAATGGTGAAAGAACATCCGATGGTGAGCCCGAAAG
CCGCGAGCGGCAGCGTGCTGCTGTTTCATCCGGAAGTGATTCATGGCAGCTTTCCGAACATTAG
CCCGTTTGCGCGTGATCTGCTGATTATTACCTATAATGATGTGAACAACGCGCCGAAACCGGCG
GGCACCCCGCGCCCGGAATATGTGATTGGCCGTGATACCACCCCGCTGGTGAGCGAAAGCGGCC
CGCTGCAT**ctcgag**caccaccacc 3′

GM_A9 protein sequence:

MQLTDAQVEQYKSDGYVLLEGAFSPEEVDIMRKALAKDAEVEGPHRIMEEDGKAVRAL
YASHKRQSVFDQLSRSDRLLGPATQLLECDLYIHQFKINTKRAFGGSAWAWHQDFIVWR

DTDGLPAPRAVNVGVFLSDVTEFNGPVVFLSGSHQKGTLERKRRATSVSDEHVDPRDYS
MTPAELEKMVKEHPMVSPKAASGSVLLFHPEVIHGSFPNISPFARDLLIITYNDVNNAPKP
AGTPRPEYVIGRDTTPLVSESGPLH

262 amino acids – Note the C-term sequence "EAAESRLA" was excluded from ProteinMPNN design as it was not resolved in the crystal structure for GriE (PDB ID: 5NCI).[3]

Protein overexpression

Chemically competent *E. coli* BL21(DE3) cells were transformed with plasmids containing *tP4H, GriE*, associated variants, and ProteinMPNN designs using a standard heat-shock protocol. Starter cultures of LB with the appropriate antibiotic were inoculated from a single *E. coli* colony on an agar plate encoding the protein of interest and grown overnight to stationary phase at 37 °C. Expression cultures of Terrific Broth supplemented with the appropriate antibiotic were inoculated with the starter cultures (2% v/v) and shaken at 37 °C at 200 rpm in a ThermoFisher Scientific MaxQ800 shaker. When expression cultures reached OD_{600} of ~0.4-0.5 (typically 2-3 hours), they were cooled to 18 °C and protein expression was induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG, 0.5 mM). Cultures were incubated at 18 °C and 200 rpm overnight (16-24 hours). Cells were pelleted by centrifugation (4 °C, 4000 rpm, 10 minutes). Cell pellets were resuspended in MES buffer (50 mM, pH 7.0) for whole-cell reactions. For reactions with lysate, whole-cell suspensions in MES buffer (50 mM, pH 7.0) were lysed by sonication on ice (70% power, 30 seconds, 1 second on/1 second off, repeated 3 times). Clarified lysate was prepared by centrifugation of lysed cells (4 °C 10,000 rpm, 15 min) and used directly in biocatalytic reactions.

Protein purification

Large-scale purification: Harvested cell pellets from 1 L overexpression were resuspended on ice in 20 mL of Lysis Buffer (25 mM Tris pH 8.0, 150 mM NaCl, 5 mM Imidazole, 5% Glycerol) containing protease inhibitor (Pierce Protease Inhibitor Mini Tablets EDTA-free, ThermoFisher A32955). Resuspended cells were lysed by sonication (70% power, 30 seconds, 1 second on/1 second off, repeated 3 times), and clarified lysate was prepared by centrifugation (16,000 rpm, 20 min, 4 °C). Clarified lysate was incubated with Ni-NTA resin (3 mL bed volume in Econo-Pac® Chromatography Columns, Bio-Rad) for 30 min at 4 °C on a rocker. The resin was washed successively with 3 column volumes of each of the following buffers: Lysis buffer (25 mM Tris pH 8.0, 150 mM NaCl, 5 mM Imidazole, 5% Glycerol) NaCl wash buffer (25 mM Tris, pH 8.0, 1M NaCl, 5 mM imidazole, 5% glycerol), and Ni-wash buffer (25 mM Tris, pH 8.0, 150 mM NaCl, 15 mM imidazole, 5% glycerol). Enriched His-tagged protein was eluted from the resin by

incubation for 10 min with Ni-elution buffer (5-10 mL, 25 mM Tris, pH 8.0, 150 mM NaCl, 250 mM, 10% glycerol). The eluted proteins were dialyzed into buffer suitable for enzymatic reactions (MES or MOPS 50 mM, pH 7.0, 10% glycerol). Protein concentration was determined by Bradford assay. Dialyzed protein used directly or was aliquoted into PCR or Eppendorf tubes, flash frozen in liquid nitrogen, and stored at -80 °C for further use. Figure S3 shows representative SDS gel samples for wild-type tP4H and R2_11 and associated variants from directed evolution rounds.

Small-scale purification: Harvested cell pellets from 25-50 mL expressions were resuspended on ice in 1-2 mL of lysis buffer (25 mM Tris pH 8.0, 150 mM NaCl, 5 mM Imidazole, 5% Glycerol) containing protease inhibitor (Pierce Protease Inhibitor Mini Tablets EDTA-free, ThermoFisher A32955). Resuspended pellets were lysed by sonication on ice in 1.5 mL Eppendorf tubes (70% power, 15 seconds with 1 second on/1 second off pulse, repeated 2 times). Clarified lysate was prepared by centrifugation (12,000 x g, 15 min, 4 °C). Clarified lysate was incubated with Ni-NTA resin in a fresh Eppendorf tube (300 µL of a 1:1 suspension of Ni-NTA resin to lysis buffer) at 4 °C for 30 min. Ni-NTA resin and His-tagged protein was spun down for 30 seconds at 12,000 x g, and supernatant was removed via pipetting. The resin bed was then washed in a similar manner with 3x1 mL washes of the following buffers: Lysis buffer (25 mM Tris pH 8.0, 150 mM NaCl, 5 mM Imidazole, 5% Glycerol) NaCl wash buffer (25 mM Tris, pH 8.0, 1M NaCl, 5 mM imidazole, 5% glycerol), and Ni-wash buffer (25 mM Tris, pH 8.0, 150 mM NaCl, 15 mM imidazole, 5% glycerol). Enriched His-tagged protein was eluted from the resin by 10-minute incubation with Ni-elution buffer (0.5-1 mL, 25 mM Tris, pH 8.0, 150 mM NaCl, 250 mM, 10% glycerol). Supernatant enriched with purified enzyme was removed after centrifugation to pellet Ni-NTA resin. Purified proteins were buffer exchanged into storage and reaction buffer (MES or MOPS 50 mM, pH 7, 10% glycerol) using an appropriately sized Zeba Spin Desalting Column with 7K MWCO. Protein concentration was determined by Bradford assay. Dialyzed protein was used directly or aliquoted into PCR or Eppendorf tubes, flash frozen in liquid nitrogen, and stored at -80 °C for further use.

Site-saturation mutagenesis and library screening

Cloning

Site-saturation mutagenesis (SSM) was carried out using the 22c-trick method.[4] All oligos used can be found in a supplementary Excel spreadsheet. 22c-trick oligos were used in conjunction with carbenicillin forward and reverse oligos to amplify two plasmid fragments for In-Fusion assembly. The carbenicillin overlap was used as a positive control for correct plasmid assembly under

antibiotic selection. All complete PCR reactions were treated with DpnI to eliminate template plasmid DNA. PCR products were isolated by gel electrophoresis using 1% agarose gels with visualization using Invitrogen SYBR Safe dye. Gel fragments were isolated after gel extraction (ThermoFisher GeneJet Gel Extraction kit). Each SSM library was constructed with two purified linear DNA fragments using In-Fusion assembly (Takara). After In-Fusion assembly, constructed plasmids were transformed into chemically competent NEB® Turbo Competent *E. coli* cells. Transformants were grown in 5 mL of LB-carbenicillin at 37 °C for 10-12 hrs to perform a quick quality control (QQC) check. After incubation, plasmid libraries were isolated after mini-prep (ThermoFisher GeneJet Plasmid Mini-prep kit) and sent for Sanger sequencing (Azenta). Confirmed libraries were then transformed into electrocompetent E. cloni EXPRESS BL21(DE3) cells (Lucigen, catalog #: 60300-2) and plated on LB-carb agar plates after a 1 hr outgrowth at 37 °C. The plated cells were incubated at 37 °C for 14-16 hr and stored at 4 °C until colony picking.

Sites chosen in each round of directed evolution for both wild-type tP4H and the R2_11 redesign are shown in Table S1.

Table S1. Summary of directed evolution screening for cyclohexane carboxylic acid (substrate 1) with tP4H and stabilized variant R2_11.

Enzyme	Directed Evolution Round	# of sites targeted	# of clones screened	Residues targeted	Winning Variant
tP4H	1	3	210	H58, F114, L174	H58L
	2	16	1120	V57, F93, I95, W106, P107, Y112, I113, F114, W115, E118, D119, W170, L174, T232, V231, F250	H58L/W170Q
	3	8	560	V57, I95, P107, I113, E118, L174, V231, F250	H58L/W170Q/E118K
R2_11	1	17	1190	V57, H58, F93, I95, W106, P107, Y112, I113, F114, W115, E118, D119, W170, L174, T232, V231, F250	H58F
	2	16	1120	V57, F93, I95, W106, P107, Y112, I113, F114, W115, E118, D119, W170, L174, T232, V231, F250	H58F/L174G
	3	8	560	V57, I95, P107, I113, E118, L174, V231, F250	H58F/L174G/V57H

Library Expression

Single colonies from the LB-carb agar plates for each SSM library were picked with sterile toothpicks and used to inoculate starter cultures using 0.5 mL LB-carb into 96-well deep well plates (Costar 3961). 70 colonies were picked for every library to ensure >95% library coverage.[4] For a positive control, 5 colonies harboring plasmid encoding for the parent of the round were picked. For a negative control, 5 colonies harboring plasmid encoding for maltose-binding protein (MBP) were picked. After colony picking, toothpicks were removed, and plates were covered in foil and incubated at 37 °C for 14-16 hr in a ThermoFisher MaxQ 8000 shaker set to 250 rpm. Expression cultures (1 mL, TB-carb) were inoculated with 50 µL of starter cultures. In parallel, glycerol stocks were prepared in 350 µL 96-well plates (USA Scientific catalog #: 1830-9610) by

mixing 50 µL of starter cultures with 50 µL sterile 1:1 glycerol:water. Glycerol stocks were sealed with cold-storage aluminum foil seals (VWR catalog #: 89049-034) and stored at -80 °C until hit identification. Inoculated expression cultures were incubated at 37 °C at 250 rpm for 3 hours. After 3 hours, plates were chilled on ice for 20 minutes before induction with IPTG (0.5 mM final concentration), and then incubated at 18 °C at 250 rpm for 18-20 hours. Cells were pelleted by centrifugation (4000 rpm, 10 minutes, 4 °C). After supernatant was discarded, plates containing pelleted cells were sealed with silicone mats and stored at -20 °C for at least 24 hours before use.

Library reactions in whole-cell

Frozen cell pellets in 96-well deep well plates were thawed on at room temperature for 10 minutes and then placed on ice. Cells were resuspended with 300 µL MES buffer (50 mM, pH 6.8). Substrate (125 mM in MES buffer, 96 µL, 20 mM final concentration), α-ketoglutarate disodium salt (250 mM in MES buffer, 96 µL, 40 mM final concentration), L-ascorbate (25 mM in water, 24 µL, 1 mM final concentration), and ferrous ammonium sulfate (12.5 mM, 24 µL, 0.5 mM final concentration) were added successively to resuspended cells. Plates were then covered loosely with foil and shaken for 24 hours at 25 °C. Reactions were quenched by addition of 300 µL acetonitrile. Plates were spun down to pellet cell debris and precipitated proteins. For individual well reaction analysis, 25 µL of the resultant supernatant was added to 225 µL of an 80:20 mixture of ACN:water with 10 mM ammonium acetate and 0.4% v/v ammonium hydroxide. Samples in later rounds of directed evolution were analyzed in a pooled manner. Pooled samples were prepared by combining 10 µL of six consecutive samples (A1-A6, A7-A12, etc) to a total volume of 60 µL. Each plate accounted for 16 pooled samples. The 60 µL sample mix was added too 190 µL of 80:20 mixture of ACN:water with 10 mM ammonium acetate and 0.4% v/v ammonium hydroxide. All LC/MS samples were filtered through a 96-well Pall AcroPrep™ Advance 96-well filter plate (350 µL, 0.2 µm Supor membrane), which was mounted onto a 96-well polypropylene sample plate (USA Scientific catalog #: 1830-9610). The stacked plates were spun down for 1 min at 4000 rpm. The plate containing filtrate was then sealed with heat-sealing aluminum foil. Samples were either stored at -20 °C or submitted directly to LC/MS analysis.

Library analysis

LC/MS Analysis was carried out using a Waters Acquity 2D UHPLC coupled to a Waters Xevo-TQ mass spec. Column: Water BEH-Amide, 2.1x50mm 1.7 µm particle size equipped with a guard column. Mobile phase A: 100% H_2O with 10 mM $NH_4CH_3COO^-$ and 0.04% NH_4OH. Mobile phase B: 95:5 ACN:water with 10 mM $NH_4CH_3COO^-$ and 0.04% NH_4OH.

Gradient:

Time	A	B	Flow rate
Initial	2.5	97.5	0.5 ml/min
3	20	80	--
4	2.5	97.5	--
5.5	2.5	97.5	--

Mass spec parameters – ESI mode: negative. Capillary voltage: 1 kV. Cone voltage: 30 V. Carboxylic acid product markers were directly injected into the source and the mass spec parameters were tuned for optimal ionization. The LC gradient conditions were optimized for separation of products from each other and from compounds in the reaction mixture. Multiple-reaction monitoring (MRM) was employed to detect specific fragments of product marker parent ions – specifically loss of water and formate. Ion counts were measured using Waters MassLynx.

Library hit selection

For directed evolution rounds where every sample was analyzed by LC/MS, fold changes in ion counts for increased hydroxylation product were measured against the parent control samples. Mean turnover was calculated, and hits one standard deviation from the mean were selected for validation. For directed evolution rounds where samples were pooled, fold changes in ion counts for increased hydroxylation product were measured against a set of pooled parent samples. The mean fold change was calculated from parent, and pooled sets 1 SD from the mean were deconvoluted by analyzing individual samples. From the deconvoluted sample set, the same analysis relative to parent was carried out and hits were chosen if they met the 1 SD cutoff. In both cases, where samples were all analyzed individually or from pooled sets, typically ~5% of variants screened were selected for validation.

Library hit validation

After selecting hits, cells from the corresponding glycerol stocks were streaked out on LB-carb agar plates. Streaked plates were incubated at 37 °C for 14-16 hours. Single colonies were picked for each hit to inoculate starter cultures (5 mL LB-carb). Starter cultures were grown at 37 °C for

10-12 hours. Plasmids were isolated from starter cultures and sent for sequencing. Unique variants were then re-expressed in 25-50 mL expressions. Variants were assayed using either clarified lysate or as pure enzymes isolated using the small-scale purification procedure described above. For validation with clarified lysates, enzyme concentration was estimated by SDS-PAGE using a calibration curve of protein at a known concentration followed by gel imaging on a LiCor Odyssey IR gel scanner and analysis of bands with ImageStudio Lite.

Small-scale validation reactions were carried out in 96-well deep well plates (Costar 3961). To enzyme (10-20 μM final concentration from purification or clarified lysate) in reaction buffer (MES 50 mM, pH 6.8) was added substrate (125 mM in MES buffer, 48 μL, 20 mM final concentration), α-ketoglutarate disodium salt (250 mM in MES buffer, 48 μL, 40 mM final concentration), L-ascorbate (25 mM in water, 12 μL, 1 mM final concentration), and ferrous ammonium sulfate (12.5 mM, 12 μL, 0.5 mM final concentration). Reaction plates were loosely covered with foil and incubated in a shaker at 25 °C for 24 hours. All reactions were run in triplicate. Reactions were quenched by addition of 150 μL acetonitrile. Total turnover (TTN) was calculated using a product marker calibration curve. In the case of product 4, calibration curves were made with both cis and trans product markers. Total turnover was calculated by dividing the product concentration by the enzyme concentration in the reaction. Stereoselectivity was calculated using the ratio of cis:trans TTNs. Samples for LC/MS analysis were prepared as described above. After LC/MS analysis, the variant with the highest mean fold change from parent that also retained 80% stereoselectivity was chosen as the winner of the round.

Scale-up biocatalytic reaction with R2_11 H58F/L174G/V57H and product isolation

To validate product identification, we performed a scale-up reaction, isolated the product, and characterized by NMR. To purified R2_11 H58F/L174G/V57H triple mutant (40 μM) in reaction buffer (MES 50 mM, pH 6.8) in a 50 mL Erlenmeyer flask was added cyclohexane carboxylic acid (Substrate **1**, 20 mM final concentration), α-ketoglutarate disodium salt (40 mM final concentration), L-ascorbate (1 mM final concentration), and ferrous ammonium sulfate (0.5 mM final concentration). The total reaction volume was 21 mL. The flask was shaken at 180 rpm at 25 °C for 24 hours. The reaction was quenched by acidification to pH ~1 with 1M HCl. The aqueous mixture was extracted with 2x15 mL of EtOAc. The combined organic layer was washed 3x15 mL with saturated brine, dried over sodium sulfate, and concentrated *in vacuo*. Purification by flash

column chromatography afforded a 4-hydroxylcyclohexane carboxylic acid (**4**, 32 mg, 30% yield) as a mixture of *cis* and *trans* isomers, consistent with previous observations (Figure S4). ^{1}H NMR (300 MHz, D_2O): δ3.79 (m, 1H), 3.52 (m, 1H), 2.41 (m, 1H), 2.29 (m, 1H), 1.89 (m, 4H), 1.74 (m, 2H), 1.57 (m, 6H), 1.35 (d, J = 13 Hz, 2H), 1.18 (q, J = 11 Hz, 2H) (Figure S6).

Continuous fluorescence-based assay for Fe(II)/αKGs ("PBP assay") for initial rates measurements and Michaelis-Menten analyses

General set-up

For Michaelis-Menten kinetic analysis of Fe(II)/αKGs used in this study, we developed a continuous coupled assay that takes advantage of a common mechanism in Fe(II)/αKGs.[5] Enzyme-catalyzed substrate oxidation is coupled to decomposition of α-ketoglutarate into succinate and CO_2. We used succinyl-CoA synthetase, which utilizes succinate, CoA, and ATP to generate succinyl-CoA along with ADP and inorganic phosphate (P_i). We detected P_i by fluorescence emitted from an engineered phosphate binding protein (PBP, ThermoFisher Scientific PV4407).[6] Production of P_i is equivalent to Fe(II)/αKG product generation (Figure S1). This assay can also be used to measure substrate uncoupled enzyme turnover. For comparison, a previously-reported Fe(II)/αKGs coupled assay used succinyl-CoA synthetase followed by two more coupled enzyme steps resulting in NADH consumption, which can be monitored with a continuous absorbance-based readout.[7]

Continuous fluorescence assays were carried out in 384 well-plates (Corning 3572) using a PerkinElmer EnVision plate reader. The top mirror module used was barcode 401. The excitation filter used was 405 nm (barcode 302) and the emission filter used was 450 nm (barcode 303). Typical assay components and amounts are shown in Table S2. Assays were set up in the following manner: a master mix was made fresh with substrate, buffer, $MgCl_2$, αKG, Fe(II), ascorbate, ATP, and Co-enzyme A. Master mix was distributed to individual wells, and PBP was added following addition of succinyl-CoA synthetase. This assay mixture was allowed to incubate at room-temperature for 5-10 minutes. Commercial sources of αKG are contaminated with low levels of succinate, and this incubation period allows for consumption of succinate prior to addition of Fe(II)/αKG enzyme. Time-course assays were initiated by addition of Fe(II)/αKGs. For background negative controls, either no Fe(II)/αKG was added or no substrate was added. For the negative control where no substrate is added, any signal generated above the no Fe(II)/αKG control reactions is likely the result of substrate uncoupled Fe(II)/αKG turnover.[7] A calibration curve for

the phosphate sensor was used to calculate [P_i] which is directly proportional to [product] from Fe(II)/αKG turnover.

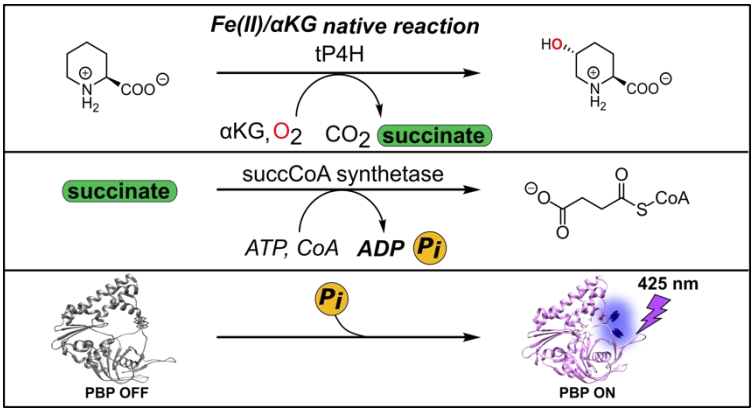


Figure S1. PBP assay general overview for Fe(II)/αKG oxidation reaction monitoring.

Table S2. General set-up for PBP assay on a 15 µL scale in a 384-well plate.

Reagent	Stock concentration (mM)	Volume (µL)	Assay concentration	Unit
Substrate	125	2.40	20	mM
Asc	25	0.30	0.5	mM
Fe	25	0.30	0.5	mM
αKG	10	0.4500	300.0	µM
Fe(II)/αKG	0.003	1.00	0.2	µM
ATP (10 mM)	10	0.45	0.3	mM
CoA (10 mM)	10	0.45	0.3	mM
SucCD	0.005	1.50	0.5000	µM
$MgCl_2$ (100 mM)	100	1.50	10.0	mM
PBP (100 µM)	0.1	3.00	20.000	µM
Assay Buffer	50 mM MOPS pH 7 50 mM NaCl	3.65		
Total volume		15.00		

Reagents and sources

Reagent	Source	Storage and composition
Substrate: Cyclohexane carboxylic acid	Sigma Aldrich Cat# 101834	125 mM in Assay Buffer, stored at r.t.
Substrate: L-pipecolic acid	TCI America Cat# P1404	125 mM in Assay Buffer, stored at r.t.
αKG	ChemImpex Cat# 00070	10 mM in assay buffer
L-ascorbate	Sigma Aldrich Cat# 255564-5G	Solid aliquots stored at r.t. – MilliQ water added at time of assay for 25 µM stock
ATP	ThermoFisher Scientific Cat# R1441	100 µL 10 mM stock diluted in MilliQ water, stored at -20 °C
Coenzyme A	Sigma Aldrich Cat# C3144	Solid aliquots stored at -20 °C, cold Assay Buffer added at time of assay for 10 mM stock
Assay Buffer: 50 mM MOPS 150 mM NaCl	MOPS: MilliPore Sigma Cat# 475922100GM NaCl: Fisher Scientific Cat# BP358-212	Prepared using in-house MilliQ water, stored at r.t.
$MgCl_2$	Fisher Scientific Cat# BP214	100 mM in MilliQ water, stored at r.t.
Fe(II) as ferrous ammonium sulfate	Sigma Aldrich Cat# 215406	Solid aliquots stored at r.t., MilliQ water added at time of assay for 25 µM stock
Phosphate binding protein (PBP)	ThermoFisher Scientific Cat# PV4407	Aliquoted and stored at -80 °C – diluted to 100 µM at time of assay in cold Assay Buffer
Succinyl CoA synthetase	Plasmid: Addgene cat# 83324; enzyme prepped in house	Aliquoted and stored in in 100 mM Tris, 150 mM NaCl, 20% glycerol pH 7.5 at -20 °C

Purification of succinyl-CoA synthetase

We found the lowest background fluorescence rate with succinyl-CoA synthetase prepared in-house after Ni-affinity purification followed by size exclusion chromatography (SEC). The SucCD gene was obtained from Addgene Cat# 83324. This plasmid did not encode for a 6xHis-tag fusion needed for Ni-affinity purification, so the SucCD gene was amplified and cloned into a standard

pET22b vector between NdeI and XhoI restriction sites to add a C-terminal 6xHis-tag onto Chain D (α subunit) of SucCD.[8] SucCD was first Ni-affinity purified and dialyzed into 100 mM Tris, 150 mM NaCl, 20% glycerol pH 7.5. Ni-affinity purified SucCD was spin concentrated to to ~50-70 μM and loaded onto an AKTA FPLC equipped with an SEC column. In all preps, a higher MW species elutes from the column that are most likely SucCD oligomers based on SDS-PAGE analysis. Fractions containing purified heterotetrameric SucCD were identified by SDS-PAGE. Purified SucCD fractions were tested in the coupled assay to determine background noise. Any fractions containing higher MW species gave significant background fluorescence in the absence of Fe(II)/αKG enzyme. Therefore, only purified heterotetrameric SucCD was aliquoted and stored at -80 °C for assay use.

Assay validation

We validated the fluorescence based coupled assay by comparing assay derived Michaelis-Menten parameters with literature reported values for GloF and HtyE. These results are shown in Table S3.

Table S3. Observed kinetic parameters for Fe(II)/αKGs HtyE and GloF compared to literature values.[9]

Enzyme	Experimental k_{cat} (s^{-1})	Literature k_{cat} (s^{-1})	Experimental K_M (mM L-Pro)	Literature K_M (mM L-Pro)
GloF	0.064	0.13	7.1	8.7
HtyE	0.18	0.65	8.3	4.2

ProteinMPNN computational sequence redesign

ProteinMPNN multiple sequence alignment (MSA) input

To generate the MSA for tP4H redesign, four iterative HHblits[10] searches were performed against the UniRef30 database (accessed June 30, 2022 at E-value cutoffs of 1e-50, 1e-30, 1e-10, and 1e-4, and the final result was filtered for 90% identity redundancy, 50% coverage, and 30% minimum query identity.

Fixed Residue Selection

For tP4H, eight methods for fixed residue selection were employed, shown in Table S4. After testing five methods in the first set of experiments, three new methods (methods 6-8) were employed. Methods 1 and 2 fix residues near the active site. In methods 3-6, residues conserved in at least X% of sequences were determined by calculating the frequency of the native amino acid identity at each position in each multiple sequence alignment (MSA) sequence. If the amino acid identity was conserved in more than X% of sequences in the alignment, that position was fixed during design. Method 6 is a repeat of method 3 with two additional positions (228 and 230) near the binding pocket fixed. In methods 7 and 8, additional residues were fixed based on alternative conservation criteria. Highly conserved positions were determined by calculating the frequency of each amino acid at each position, identifying the most highly conserved amino acid, and ranking all positions by the frequency of the most highly conserved amino acid. This step was performed using the find_conserved_resi.py script (included with Supporting Information as find_conserved_resi.txt).[11] The top 50% (method 7) or 70% (method 8) of positions were fixed during sequence design. For GriE, only method 8 was employed for design sequence generation after fixing active site residues. Complete lists of residues fixed in each of these methods are provided in the Supplementary spreadsheet (ProteinMPNN sequences_metrics tab).

Table S4. ProteinMPNN sequence generation methods for fixed residue selection.

Method	Description
1	Fix active site residues: tP4H: H58, Q92, K94, N96, K98, W106, H109, Q110, D111, F114, W115, Q173, H215, S217, R226, R247 GriE: Y59, Q93, K95, N97, K99, W107, H110, Q111, D112, V115, S116, V144, H169, V170, D171, P172, H210, A212, R221, L223, I225, R242, V246
2	Fix active site (residues from Method 1) and 10 Å sphere around active site. The 10 Å sphere was defined as any residues containing sidechain atoms within 10 Å of any of the active site residues.
3	Fix active site (residues from Method 1) and residues conserved in at least 35% of MSA sequences
4	Fix active site (residues from Method 1) and residues conserved in at least 70% of MSA sequences
5	Fix active site (residues from Method 1) and residues conserved in at least 95% of MSA sequences
6	Method 3 + L228 and V230
7	Fix active site (Residues shown in Method 1 + L228 and V230) and 50% most highly conserved residues from MSA
8	Fix active site (Residues shown in Method 1 + L228 and V230 for tP4H) and 70% most highly conserved residues from MSA. For GriE redesign only this method was used.

ProteinMPNN Design of tP4H and GriE

The structure of tP4H was predicted with AlphaFold2[12] and used as structural input to ProteinMPNN.[13] The code used in this work is available on the ProteinMPNN github repository (commit 0a72127, June 9, 2022). Active site and conserved residues for tP4H and GriE were excluded from design as described in Table S4. Cysteine was excluded from the amino acid identities that could be installed during design. Three temperature sampling parameters (0.1, 0.2, and 0.3) were used during design.[14] A model of ProteinMPNN trained with 0.2 Å noise applied to training set protein backbones was used to perform sequence generation.

Sequences generated with ProteinMPNN were predicted with AlphaFold2, using model 3 with 6 recycling steps. Both designs and native tP4H and GriE predicted with low confidence if given only the single sequence and minimal recycling steps; we found that structural templating with MSAs was necessary for accurate prediction. To generate MSAs of each design for structure prediction, the MSA of the parent sequence was used, and the parent sequence was swapped for the design sequence. All sequences generated were predicted with Cα RMSD < 2.0 Å and pLDDT

> 85.0 and were predicted to maintain critical structural features in the active site. For methods 6-8, we ordered the top ranked 32 out of 48 sequences based on top Cα RMSD values.

The following command was used to perform sequence design with ProteinMPNN on tP4H.

```
python $MPNN_PATH/protein_mpnn_run.py \
        --jsonl_path ../parsed_pdbs_bb.jsonl \
        --chain_id_jsonl ../assigned_chains.jsonl \
        --fixed_positions_jsonl ../masked_pos.jsonl \
        --out_folder $MPNN_OUTDIR \
        --num_seq_per_target 16 \
        --sampling_temp “0.1 0.2 0.3” \
        --batch_size 8 \
        --omit_AAs=’XC’
```

Where ../assigned_chains.jsonl contains the parsed PDB chain information: {“tP4H”: [[“A”]] }

This script generates 16 sequences for each sampling temp for a total of 48 sequences. The -omit_AAs line excludes cysteine residues from being installed during the design.

Sets of designs were distinguished by selection of fixed residues (Table S4).

Design cloning and screening

All ProteinMPNN design sequences can be accessed in a supplementary Excel spreadsheet. Designs were purchased as IDT eBlocks and were used directly in In-Fusion reactions to construct design encoded plasmids. The eBlock 5’- and 3’-overhangs are shown below. eBlocks for tP4H were used directly in In-Fusion reactions with a pET-28a(+) vector that was digested with NcoI and XhoI. eBlocks for GriE were used directly in In-Fusion reactions with a pET-22b(-) vector that was digested with XbaI and XhoI. General eBlock designs are shown below:

General ProteinMPNN design eBlock DNA sequence and In-Fusion overhangs for pET-28a(+). 15 bp In-Fusion overhangs are bolded, and restriction sites are highlighted in red:

5’aggagatataccatgggcagcagccatcatcatcatcacagcagcggcctggttggcag
c**DESIGNSEQUENCEHERE**c**tcgag**caccaccacc 3’

General ProteinMPNN design eBlock DNA sequence and In-Fusion overhangs for pET-22b(-). 15 bp In-Fusion overhangs are bolded, and restriction sites are highlighted in red:

5′
acaattcccctctagaaataattttgtttaactttaagaaggagatatacatDESIGNSEQUENCEHEREctcgagcaccaccacc 3′

After In-Fusion reactions, constructed plasmids were transformed directly into BL21(DE3) cells. After a 1 hour outgrowth, starter cultures were prepared with 1 mL of LB in a 96-well deep well plate inoculated with transformed cells. Starter cultures were incubated in a shaker at 37 °C for 12-14 hours. Expression cultures (1 mL, TB-carb) were inoculated with 50 µL of starter cultures. In parallel, glycerol stocks were prepared in 350 µL 96-well plates (USA Scientific catalog #: 1830-9610) by mixing 50 µL of starter cultures with 50 µL sterile 1:1 glycerol:water. Glycerol stocks were sealed with cold-storage aluminum foil seals (VWR catalog #: 89049-034) and stored at -80 °C until design hit identification. Glycerol stocks were sealed with cold-storage aluminum foil seals (VWR catalog #: 89049-034) and stored at -80 °C until hit identification. Inoculated expression cultures were incubated at 37 °C at 250 rpm for 3 hours. After 3 hours, plates were chilled on ice for 20 minutes before induction with IPTG (0.5 mM final concentration), and then incubated at 18 °C at 250 rpm for 18-20 hours. Cells were pelleted by centrifugation (4000 rpm, 10 minutes, 4 °C). After supernatant was discarded, plates containing pelleted cells were sealed with silicone mats and stored at -20 °C for at least 24 hours before use. Expression of designs was tested with SDS-PAGE. The general workflow for assessing hits for tP4H ProteinMPNN (Methods 6-8, Table S4) is shown in Figure S2. A similar workflow was used for GriE. Designs passing each criteria can be found in Supplementary Spreadsheet – ProteinMPNN sequences_metrics.

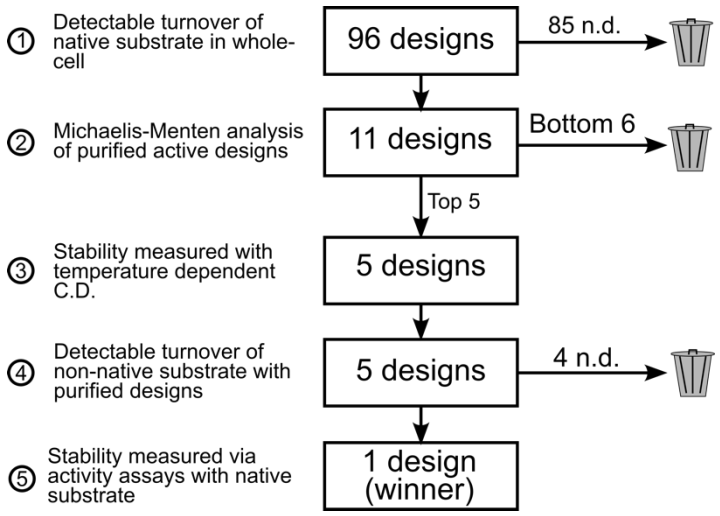


Figure S2. Selection criteria and workflow for the second set of tP4H ProteinMPNN design screening.

Initial activity screen in whole-cell: Initial screening was carried out to identify designs that turned over the native tP4H and GriE substrates L-pipecolic acid and L-leucine, respectively. Frozen cell pellets in 96-well deep well plates were thawed on at room temperature for 10 minutes and then placed on ice. Cells were resuspended with 300 µL MES buffer (50 mM, pH 6.8). Amino acid substrate (125 mM in MES buffer, 96 µL, 20 mM final concentration), α-ketoglutarate disodium salt (250 mM in MES buffer, 96 µL, 40 mM final concentration), L-ascorbate (25 mM in water, 24 µL, 1 mM final concentration), and ferrous ammonium sulfate (12.5 mM, 24 µL, 0.5 mM final concentration) were added successively to resuspended cells. Plates were then covered loosely with foil and shaken for 24 hours at 25 °C. Reactions were quenched by addition of 300 µL acetonitrile. Plates were spun down to pellet cell debris and precipitated proteins. For individual well reaction analysis, 25 µL of the resultant supernatant was added to 225 µL of an 80:20 mixture of ACN:water with 10 mM ammonium acetate and 0.4% v/v ammonium hydroxide. LC/MS was run using the method described above in the Site Saturation Mutagenesis Library Analysis workflow.

Michaelis-Menten kinetics screen with purified tP4H designs: Any designs that turned over native substrates were expressed and purified using the small-scale Ni-affinity purification protocol described above. Initial rates were measured using the PBP assay described above. For tP4H, Michaelis-Menten analysis was carried out where the K_M for each design for αKG was determined first by fixing L-pipecolic acid at 20 mM and measuring initial rates with the following

concentrations of αKG: 0, 6, 19, 56, 167 and 500 μM. For Michaelis-Menten analysis with tP4H using L-pipecolic acid, a saturating concentration of αKG (300 μM) was used along with 200 nM enzyme. Initial rates were then measured at the following concentrations of L-pipecolic acid: 0, 0.25, 0.74, 2.2, 6.7, 20, and 40 mM. The top 5 designs with the highest k_{cat} were selected for analysis by temperature dependent circular dichroism (CD) spectroscopy (Figure S8).

Michaelis-Menten kinetics with purified GriE designs: Any GriE designs that turned over L-leucine with yield at least 2-fold below the wild-type GriE control were selected for purification and initial rate analysis. This criteria included 27 enzymes out of 32 designs. At this stage only observed rates were measured at 300 μM aKG and 20 mM L-leucine, without varying concentrations, to determine Michaelis-Menten parameters. After the initial rates were measured, the top 5 were selected for analysis by temperature dependent circular dichroism (CD) spectroscopy. After temperature dependent CD, the most stable and active design GM_A9 was characterized with a full Michaelis-Menten analysis. The K_M of αKG for GM_A9 was determined first by fixing L-leucine at 20 mM and the following αKG concentrations were used with 200 nM enzyme to measure initial rates: 800, 400, 200, 100, 50, 25, and 12.5 μM. Then, a saturating concentration of 300 μM αKG was used along with 200 nM enzyme. Initial rates were measured at the following concentrations of L-leucine: 10, 3.3, 1,1, 0.37, 0.12, 0.04, and 0.01 mM.

Temperature dependent CD spectroscopy: To determine secondary structure and thermostability of design candidates from the second-pass screen described above, CD measurements were carried out on a JASCO J-1500 instrument using a 1 mm pathlength cuvette. Samples of purified protein were prepared at 0.4 mg/mL in 50 mM potassium phosphate buffer, pH 7.0. The sample temperature was ramped from 25 °C to 95 °C with full spectrum scans from 190 nm to 260 nm performed after each 10 °C interval. The molar residue ellipticity (MRE) at 220 nm was plotted over the temperature gradient to visualize temperature of unfolding. Representative CD data for tP4H and GriE can be found in Figure S8 and Figure S12B, respectively.

Non-native activity screen with purified enzyme: To determine if designs reacted with a non-native free acid substrate, purified enzymes were tested in small-scale biocatalytic reactions with cyclohexane carboxylic acid (substrate **1**). Reactions were run in the same manner described in SSM Library Hit Validation. LC/MS was used to determine TTN values.

Stability-activity assay with purified enzyme: Design stability was assessed by measuring initial rates at 25, 45 and 65 °C at t=0, 7 hours, and 24 hours unless otherwise stated. Initial rates were measured using the PBP assay described above.

2.6 Experimental Data

SDS-PAGE of tP4H variants and ProteinMPNN redesigns

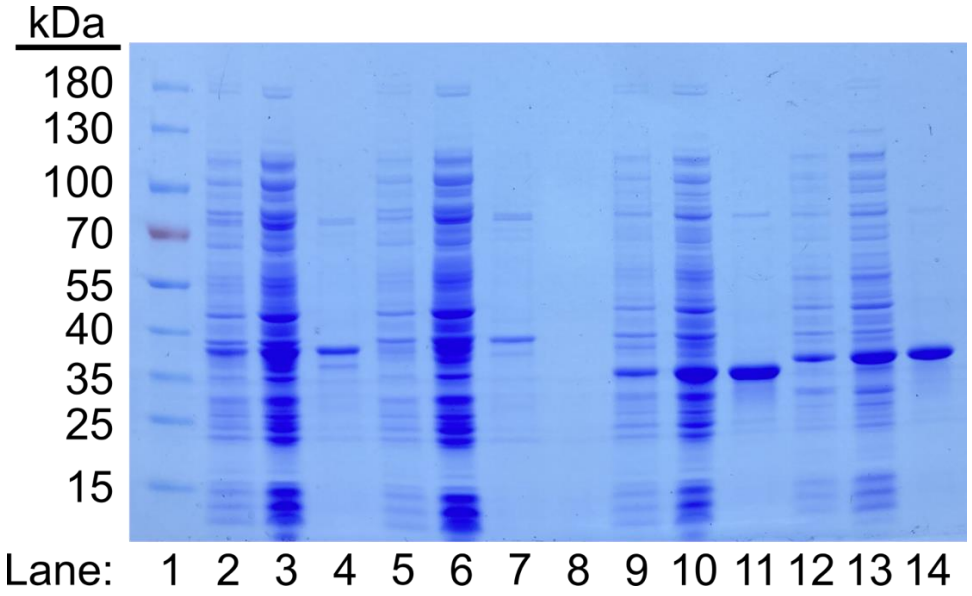


Figure S3. SDS-PAGE gel of wild-type tP4H, tP4H H58L/W170Q/E118K, R2_11, and R2_11 H58F/L174G/V57H. For each construct, gel samples are shown for post-expression from cell culture, lysate, and after Ni-affinity purification. All samples collected were diluted in the same way, so differences in staining intensity reflect differences in sample protein concentration.

Table S5. Gel lanes and descriptors for Figure S3.

Lane	Description	Lane	Description
1	Ladder	8	Blank
2	wt tP4H post expression	9	R2_11 base design post expression
3	wt tP4H lysate	10	R2_11 base design lysate
4	wt tP4H purified	11	R2_11 base design purified
5	tP4H H58L/W170Q/E118K post expression	12	R2_11 H58F/L174G/V57H post expression
6	tP4H H58L/W170Q/E118K lysate	13	R2_11 H58F/L174G/V57H lysate
7	tP4H H58L/W170Q/E118K purified	14	R2_11 H58F/L174G/V57H purified

LC-MS characterization of substrate 1 hydroxylation products and calibration curves

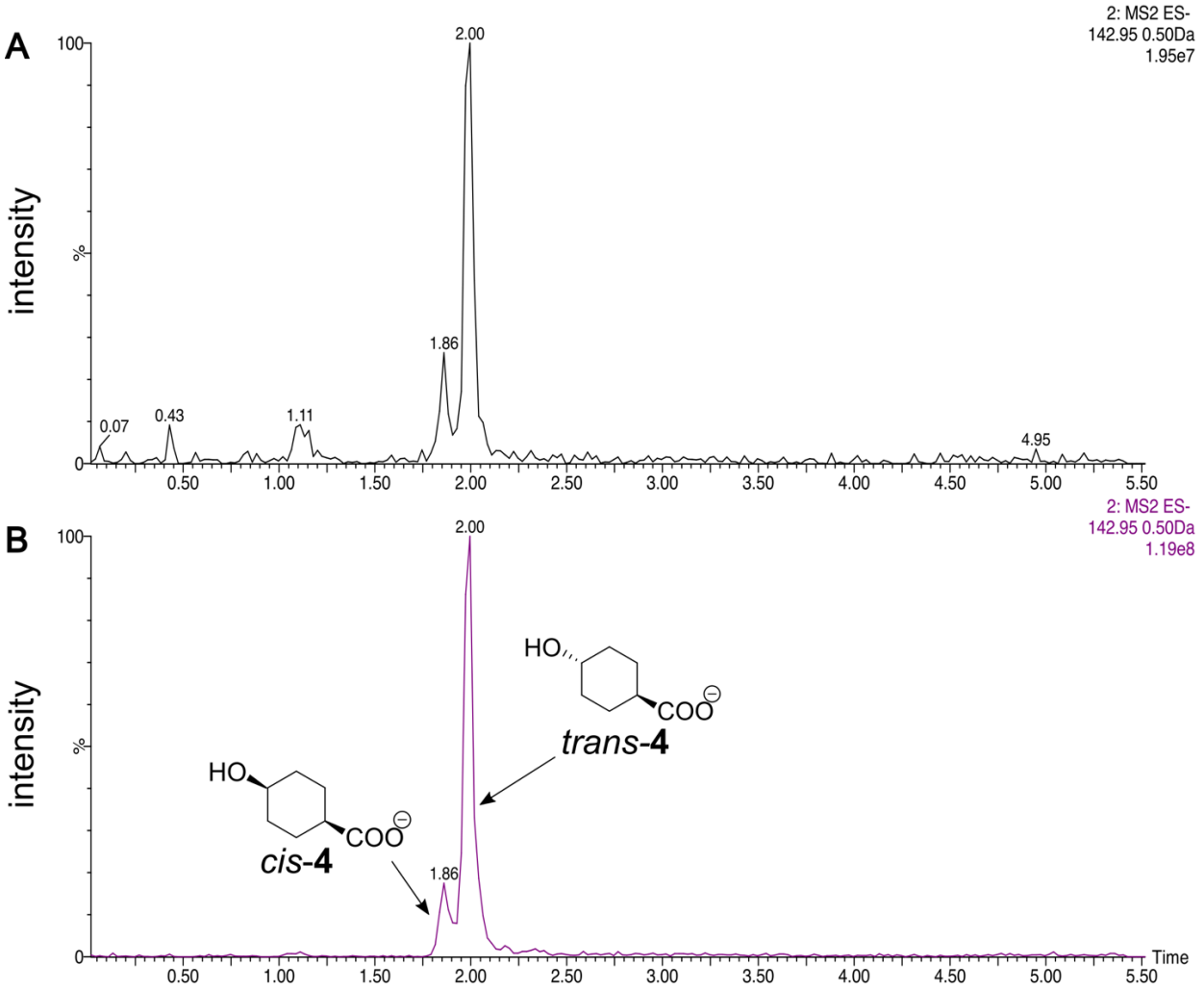


Figure S4. A) LC-MS extracted ion chromatogram (EIC) of a representative reaction sample with the R2_11 H58F/L174G/V57H triple mutant using the method described above in "Library Analysis." The 4:1 ratio of *trans* to *cis* was determined from calibration curves constructed from co-injected markers at varying ratios (Figure S5). The 4:1 selectivity was consistent across all mutants from both wild-type tP4H and R2_11. B) LC-MS EIC of a co-injection of authentic *cis*-**4** and *trans*-**4** product markers using the same method

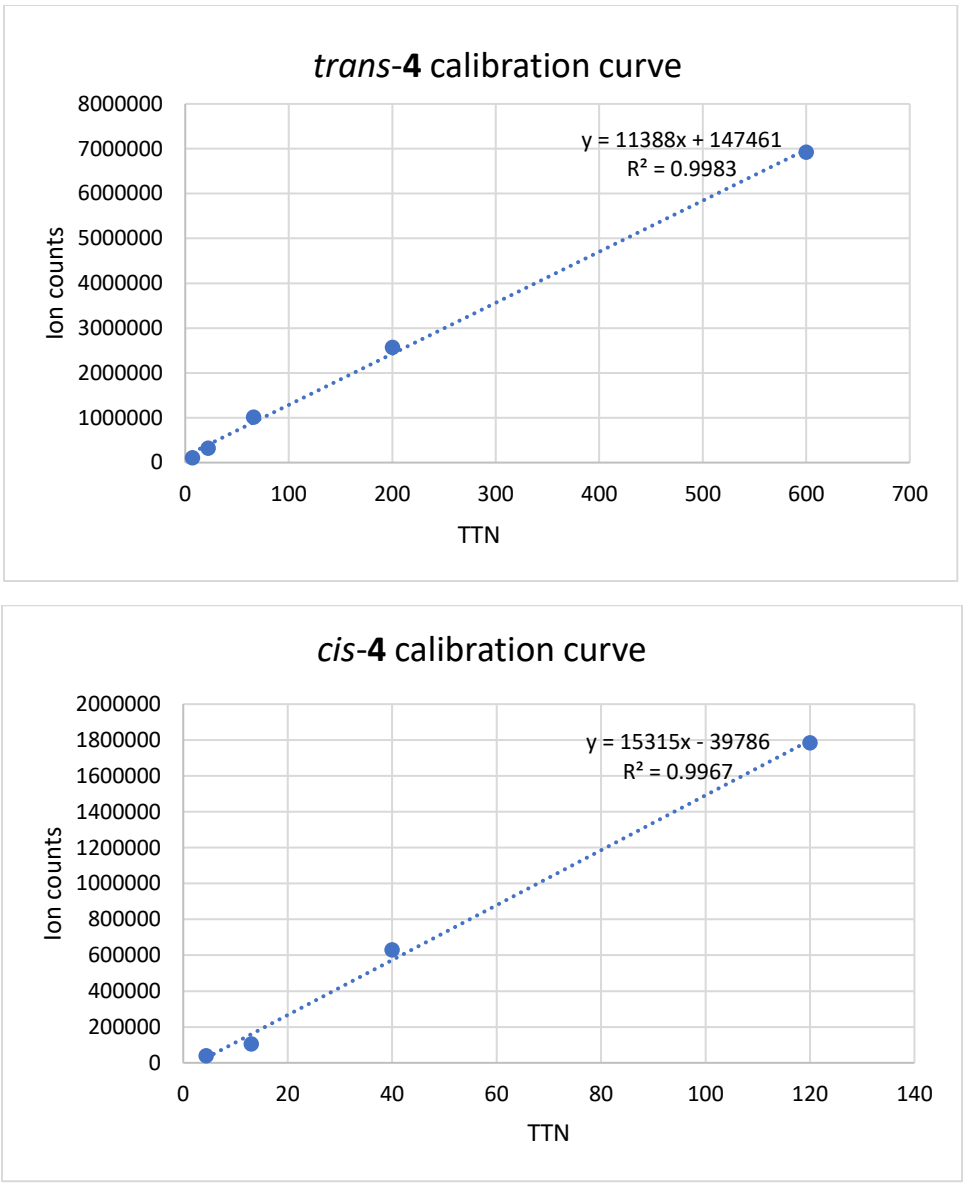


Figure S5. Calibration curves for authentic trans-4-hydroxycyclopentane carboxylic acid product markers. Ion counts were obtained from the extracted ion chromatograms for m/s 142.95 ([M]-1) using the LC-MS method described in the "Library Analysis" section above. A) Calibration curve using product marker for *trans*-4-hydroxycyclohexane carboxylic acid (Combi-blocks Cat. #: OR-5210). B) Calibration curve using product marker for *cis*-4-hydroxycyclohexane carboxylic acid (Combi-blocks Cat. #: QG-7784).

NMR spectra of isolated product 4 and product markers

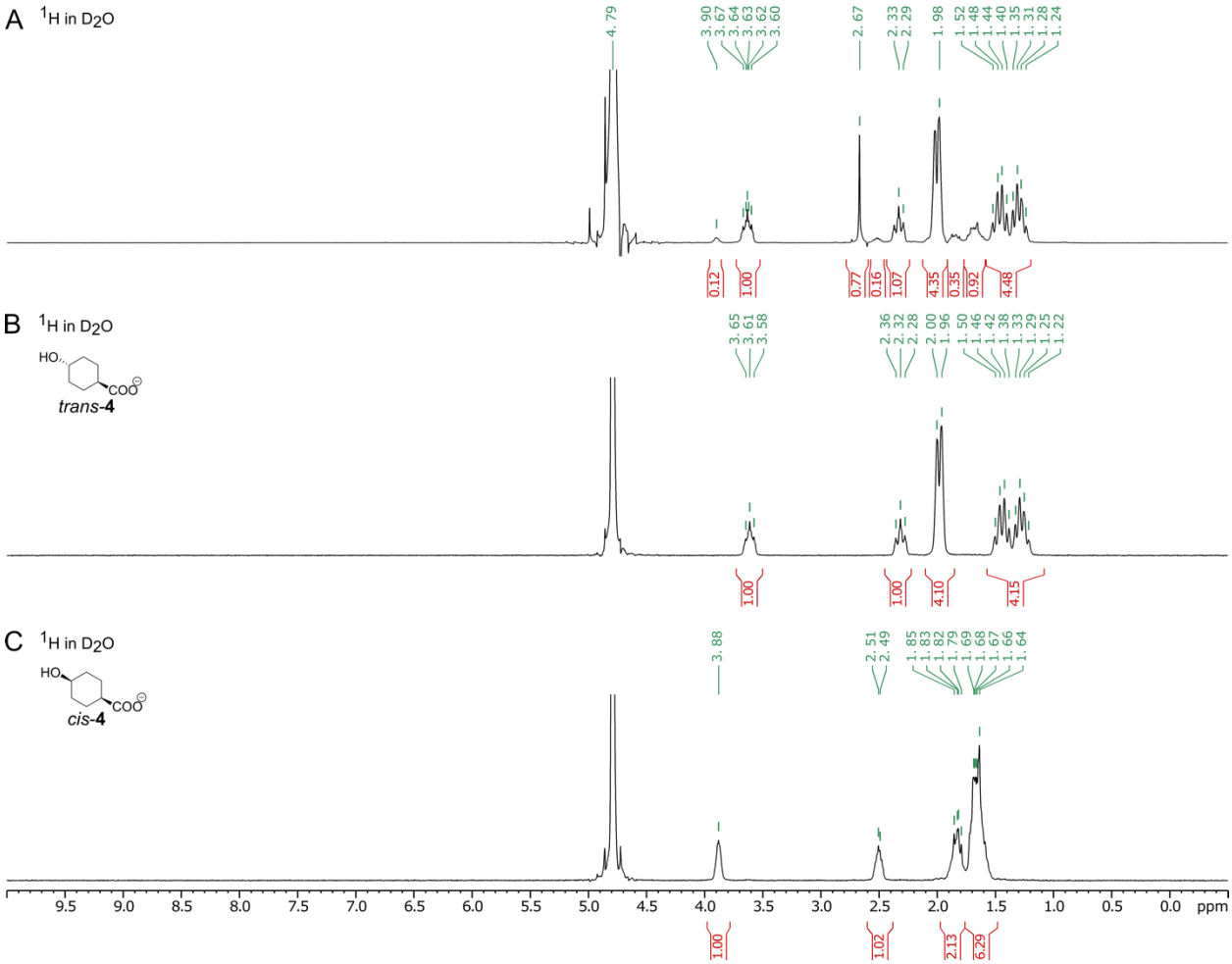


Figure S6. A) ^{1}H NMR spectrum (300 MHz, D_2O) of isolated **4** as a mixture of *cis* and *trans* isomers after flash silica gel chromatography from a scale-up biocatalytic reaction with R2_11 H58F/L174G/V57H triple mutant. B) ^{1}H NMR spectrum (300 MHz, D_2O) of *trans*-**4** authentic product marker (Combi-blocks Cat. #: OR-5210). C) ^{1}H NMR spectrum (300 MHz, D_2O) of *cis*-**4** authentic product marker (Combi-blocks Cat. #: QG-7784. For biocatalytic reaction conditions and an isolation procedure, see section titled "Scale-up biocatalytic reaction with R2_11 H58F/L174G/V57H and product isolation." on Page 14. In panels A-C, green labels are chemical shifts and red values are peak integrations.

LCMS traces for reactions of GriE and GM_A9

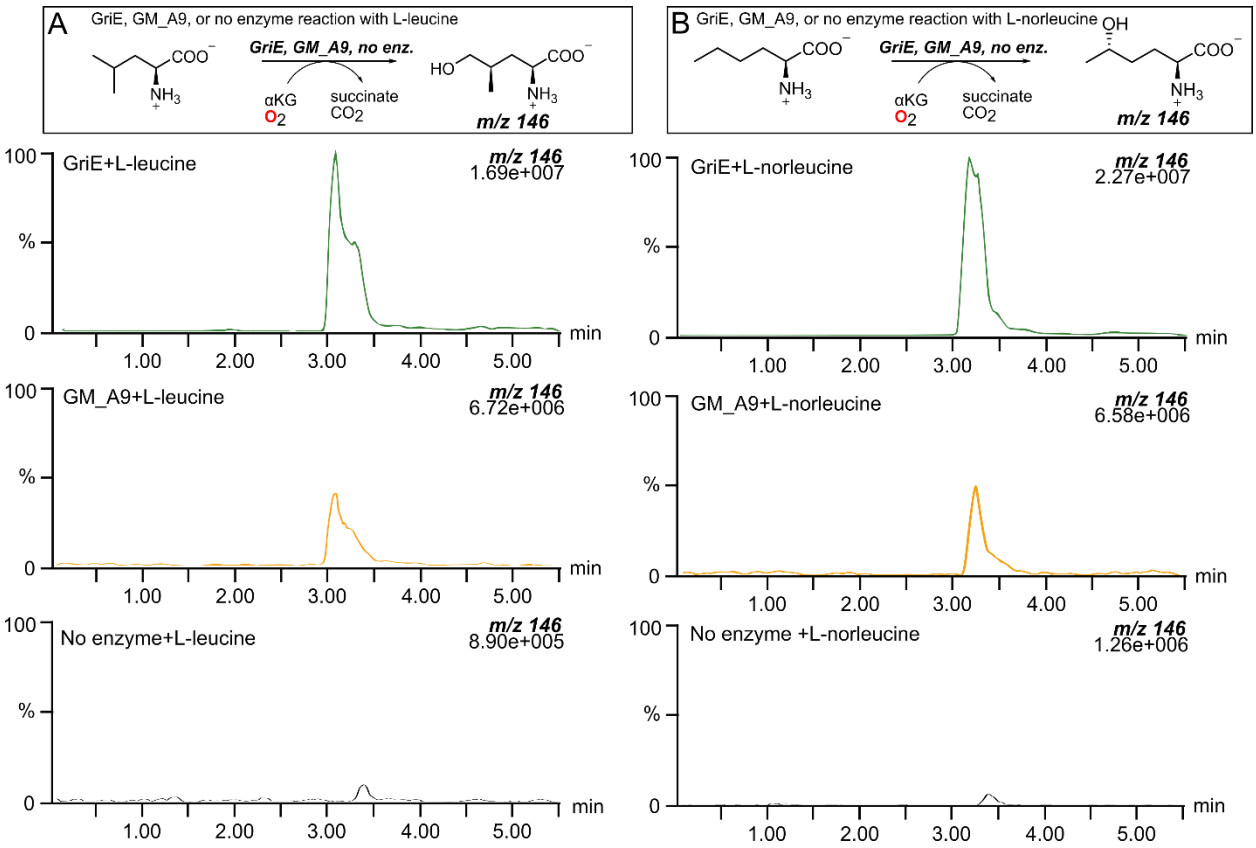


Figure S7. LCMS data were collected on a Waters Xevo TQ using the method described under "Library Analysis." Reactions were carried out with purified enzyme (15 μM) in reaction buffer (MES 50 mM, pH 6.8) with substrate (20 mM), αKG (40 mM), ferrous ammonium sulfate (0.5 mM) and ascorbic acid (1 mM). The extracted m/z as well as ion intensity are listed in the upper right-hand corner of each trace. A) Extracted ion chromatograms for reaction samples of GriE with L-leucine, GM_A9 with L-leucine, and a no enzyme with L-leucine control. B) Extracted ion chromatograms for reaction samples of GriE with L-norleucine, GM_A9 with L-norleucine, and a no enzyme with L-norleucine control.

Circular dichroism data for wild-type tP4H and redesigned variants

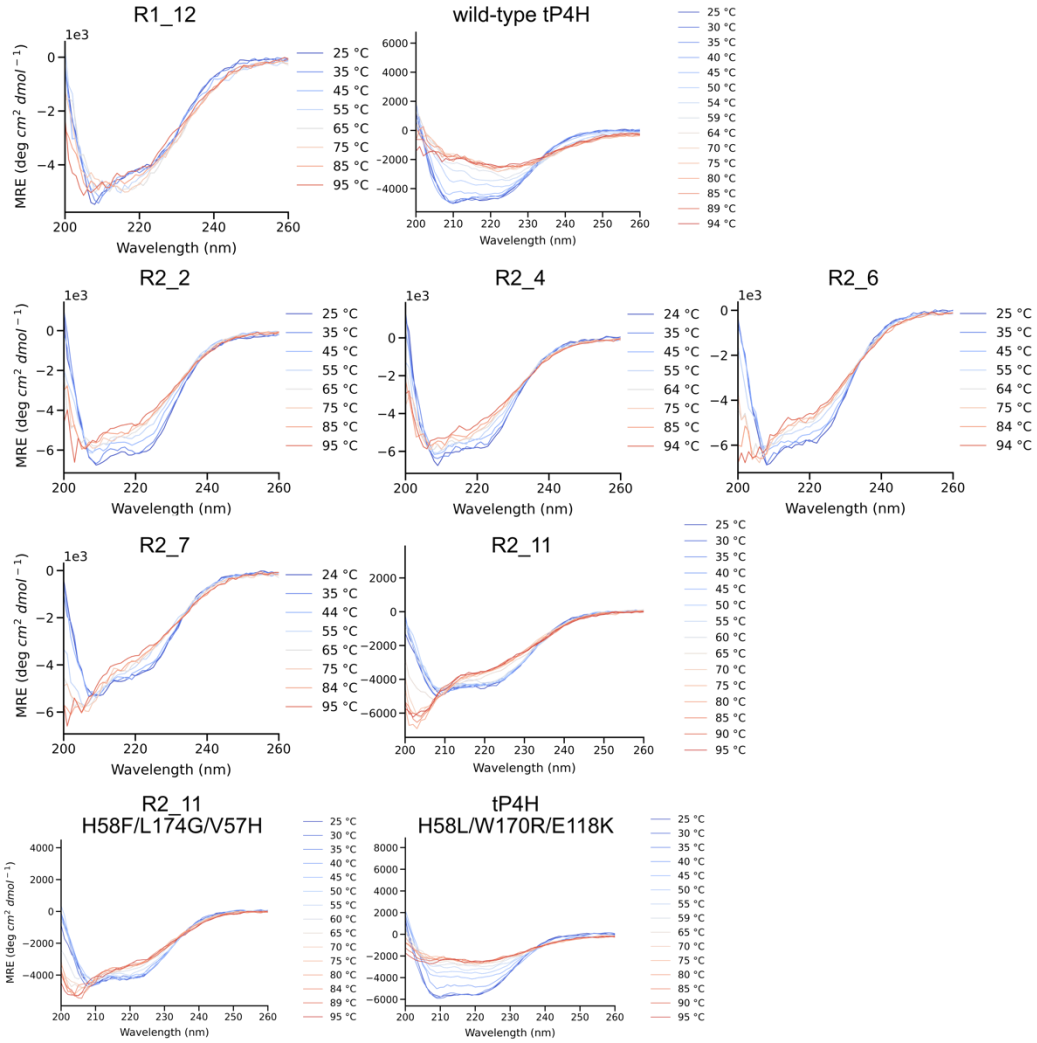


Figure S8. CD spectra for wild-type tP4H and redesigns, as well as tP4H and R2_11 triple mutants. CD spectra were collected at 25-95°C in 5 or 10 °C increments. Protein was prepared at 0.4 mg/mL in potassium phosphate buffer (50 mM, pH 7.0).

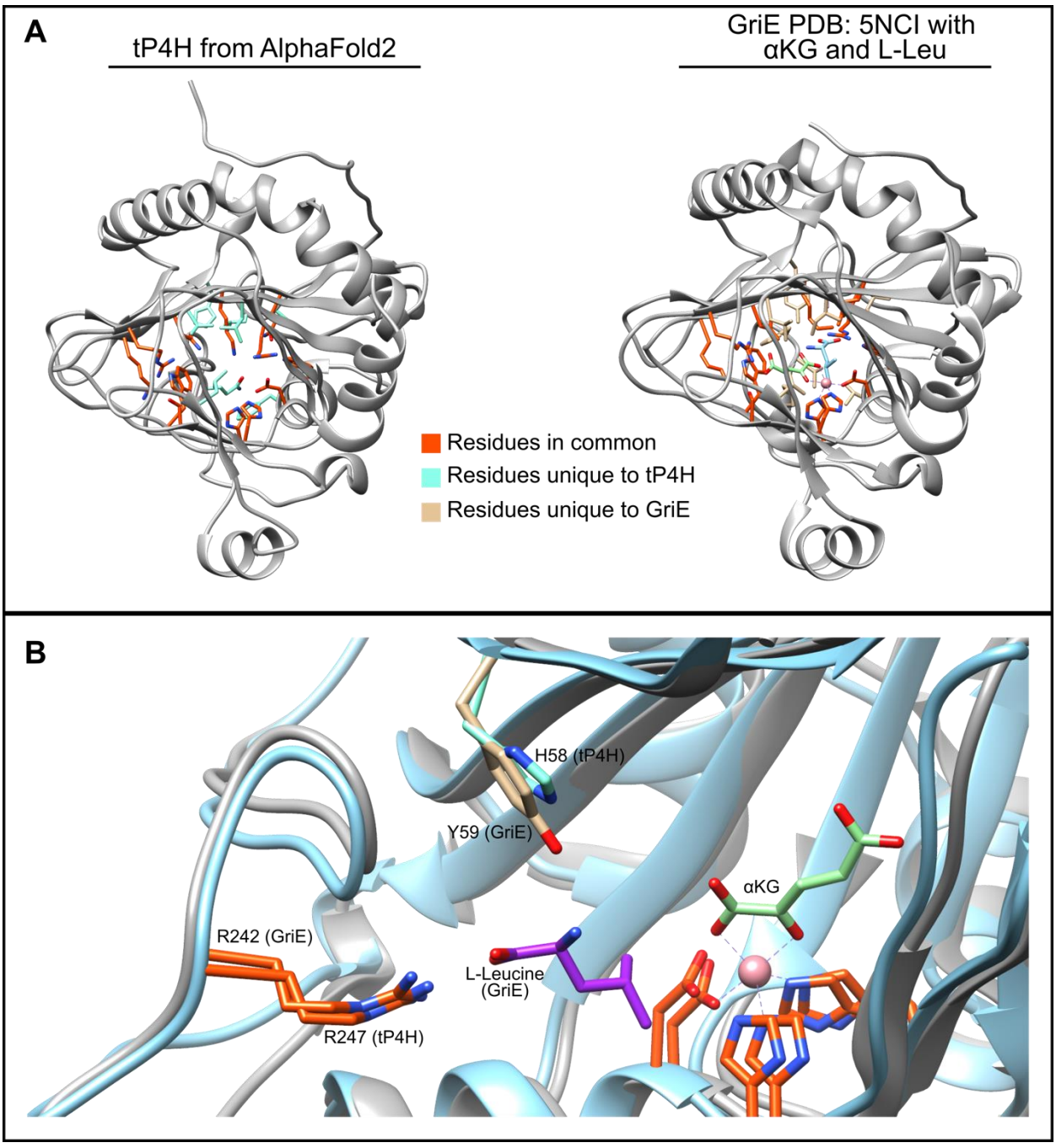


Figure S9. A) Structural model of tP4H from AlphaFold2 compared to a crystal structure of Fe(II)/αKG GriE (PDB ID: 5NCI).[3] Both common and unique features are shown via colored side-chains. The GriE structure contains bound cobalt (metal that GriE was crystallized with, light pink), αKG (light green), and the L-Leucine substrate (purple). B) Active site overlay of GriE (grey) and tP4H (light blue). If the L-pipecolic acid substrate in tP4H binds similarly to the orientation of L-Leucine in GriE, then the carboxylate group would likely interact with the tP4H R247 residue, and the amine group would likely interact with tP4H H58.

Table S6. Wild-type Fe(II)/αKG enzyme turnover with native substrates in whole-cell reactions.

Fe(II)/αKG	Substrate	% conversion to hydroxylated product[a,b]
PolL[15]	L-leucine	59
LdoA[15]	L-leucine	<5
IDO[15]	L-leucine	92
GriE[2]	L-leucine	84
Gox[2]	L-leucine	93
cP3H[1]	L-proline	45
cP4H[1]	L-proline	19
GloF[9]	L-proline	59
HtyE[9]	L-proline	68
tP4H[c][1]	L-pipecolic acid	17
GetF[16]	L-pipecolic acid	>99
PiFa[16]	L-pipecolic acid	52

[a] All reactions were carried out in whole-cell with 20 mM substrate, 40 mM αKG, 1 mM ferrous ammonium sulfate and 1 mM ascorbic acid for 24 hours at 25 °C. These conditions are comparable to reaction conditions previously reported for these enzymes.[1,2,9,15,16] Frozen and thawed cell pellets were resuspended in buffer (MOPS 50 mM pH 7.0) at 5% of the expression volume.
[b] Conversion to hydroxylated amino acid products was quantified by analytical HPLC-UV analysis after Fmoc-Cl derivatization of reaction mixtures.
[c] Previous reports with tP4H suggest that the enzyme needs to be co-expressed with a chaperone in order to have activity.[1] We found that chaperone co-expression was not necessary and enzyme activity with and without chaperone co-expression was comparable. Additionally, we found no difference in the initial rates of tP4H with L-pipecolic acid with and without chaperone co-expression.

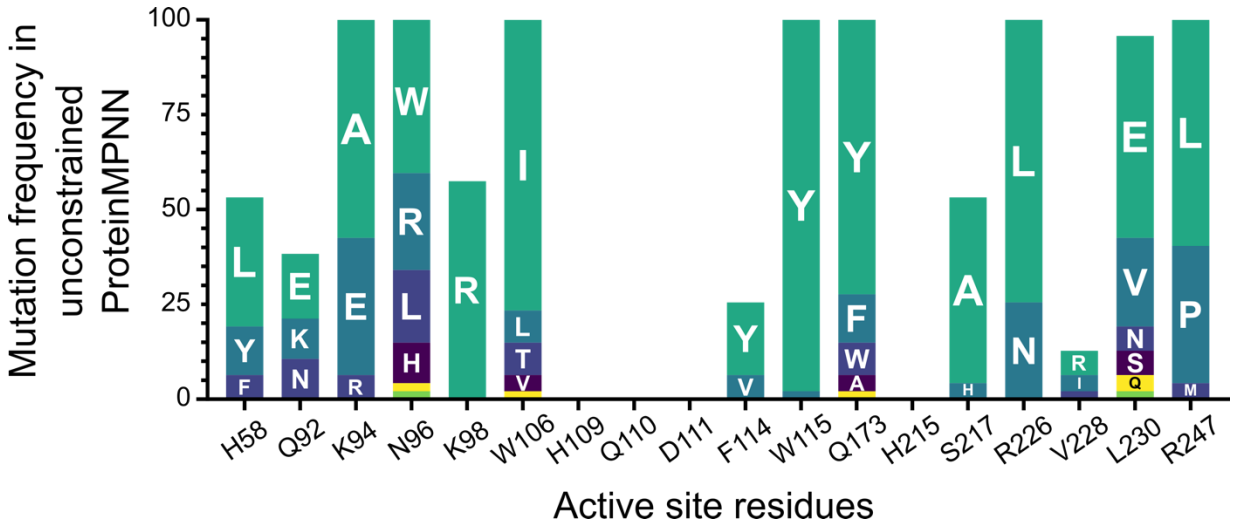


Figure S10. Unconstrained ProteinMPNN redesigns tP4H active site residues. The x-axis shows key tP4H active site residues that were fixed in typical tP4H ProteinMPNN redesigns (see Table S4). The y-axis shows the frequency of designed mutations at these sites when ProteinMPNN is unconstrained (i.e. no active site residues fixed). Specifically, the colored bars indicate the frequency across 48 output sequences for any amino acid other than the wild-type parent. We observed frequent mutations at functionally important sites. For example, position 94 is critical for αKG cofactor binding and is changed from the wild-type residue K in 100% of the unconstrained ProteinMPNN redesigns. Notably, unconstrained ProteinMPNN did not introduce mutations at Fe^{2+} ligands H109, D111, and H215 (Figure 2) or the nearby residue Q110. ProteinMPNN receives only protein information as input and does not explicitly account for the presence of the metal ion,[13] so this motif may be preserved due to its frequent presence in the original training set.

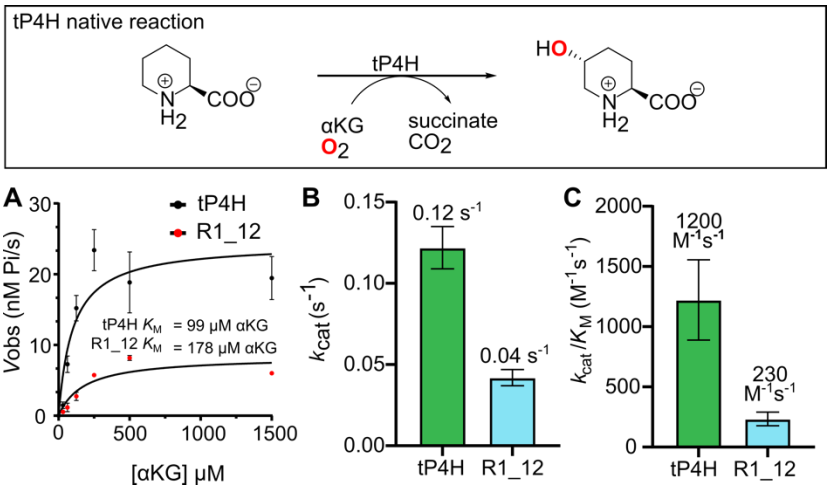


Figure S11. A) Michaelis-Menten analysis of wild-type tP4H and R1_12 with varying concentrations of αKG cofactor using the native reaction with 20 mM L-pipecolic acid and 200 nM enzyme. Initial rates were measured in triplicate using the PBP assay. Initial rates were plotted against cofactor αKG concentrations in GraphPad Prism and the non-linear Michaelis-Menten fit function was used to calculate k_{cat} and K_M values. The fit was generated using each individual replicate y-value as an individual point. The plot of V_{obs} vs. [concentration] is shown with each point as the average V_{obs} ± SD to simplify comparisons between different reactions. B) k_{cat} comparison of wild-type tP4H and ProteinMPNN design candidate R1_12. C) Catalytic efficiency (k_{cat}/K_M) comparison of wild-type tP4H and ProteinMPNN design candidate R1_12. In B and C, error bars reflect the standard error of the fit. To calculate the standard error of the fit for k_{cat}/K_M, an alternative form of the Michaelis-Menten equation was used: $V_{obs}=(k_{cat}/K_M)[E]_0[S]/(1+([S]/K_M))$.

Table S7. Michaelis-Menten parameters for active tP4H ProteinMPNN designs[a]

	wt[b]	R2_1	R2_2	R2_3	R2_4	R2_5	R2_6	R2_7	R2_8	R2_9	R2_11[b]
k_{cat} (s^{-1})	0.14 ±0.004	0.027	0.041	0.005	0.068	0.037	0.06	0.037	0.016	0.033	0.10± 0.006
K_M (mM)	0.65 ±0.08	41	1.3	<0.25	2.1	51	3.7	1.4	2.6	1	0.54± 0.15
k_{cat}/K_M ($M^{-1}*s^{-1}$)	220± 24	0.7	31	--	32	0.7	16	27	6.2	32	190± 44

[a] Active tP4H ProteinMPNN designs obtained from Methods 6-8 (Table S4). Initial rates (product vs. time curves) were measured using the PBP assay using 200 nM enzyme and varying concentrations of substrate L-pipecolic acid. First-derivatives of product vs time curves were plotted against substrate concentrations in GraphPad Prism and the non-linear Michaelis-Menten fit function was used to calculate k_{cat} and K_M values. For R2_3, observed rates exhibited saturation behavior for the lowest substrate concentration tested (0.25 mM). For this variant only a limit for K_M could be determined and k_{cat}/K_M could not be measured. Note that the wt tP4H K_M value in this table is for substrate and is not the same as the K_M value obtained for the αKG cofactor (Figure S11).

[b] Both wt tP4H and R2_11 kinetics were measured in triplicate and standard errors are reported for each parameter. The fits were generated using each individual replicate y-value as an individual point. All other designs were screened in single replicate.

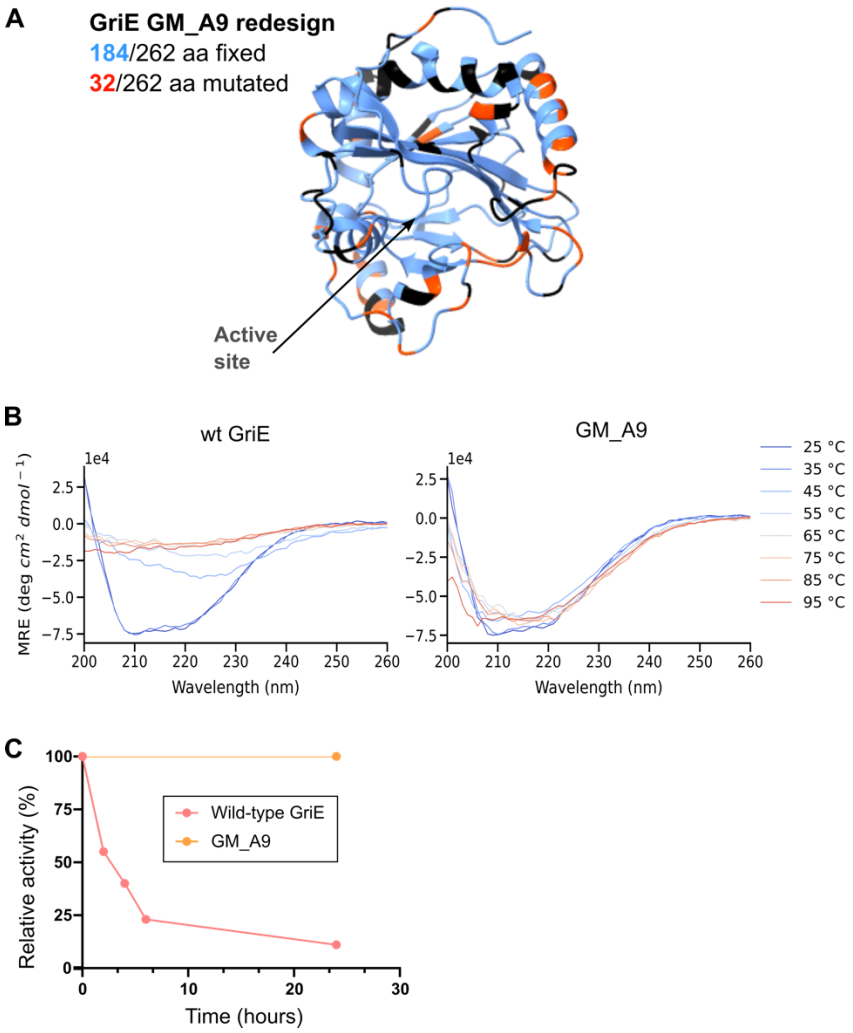


Figure S12. A) GriE structure (PDB 5NCI)[3] color-coded to show sites fixed in the design process (blue, Supplementary spreadsheet – ProteinMPNN sequences_metrics) and sites mutated in the ProteinMPNN GM_A9 redesign (orange-red). Sites colored black were neither fixed nor redesigned in the GM_A9 variant. B) Temperature dependent CD of wild-type GriE compared to ProteinMPNN redesign GM_A9. C) Activity-stability analysis of wild-type GriE compared to ProteinMPNN redesign GM_A9 at room temperature. Activity measurements were made using the PBP coupled assay described above in the Methods section.

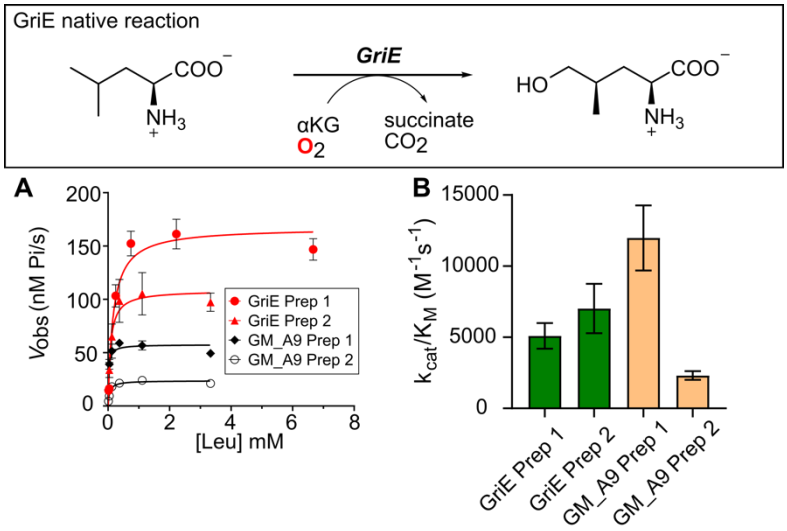
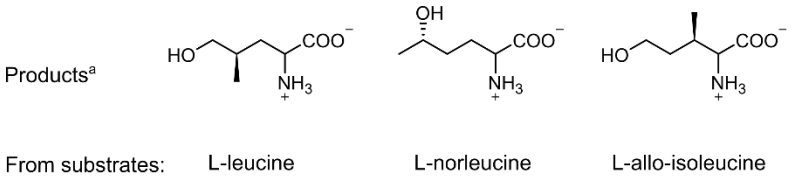


Figure S13. (A) Michaelis-Menten analysis of wild-type GriE and ProteinMPNN redesign GM_A9 with L-Leucine with 200 nM enzyme. Initial rates were measured in triplicate using the PBP assay. Initial rates were plotted against substrate concentrations in GraphPad Prism and the non-linear Michaelis-Menten fit function was used to calculate k_{cat} and K_M values. Fits were generated using each individual replicate y-value as an individual point. Plots of V_{obs} vs. [concentration] are shown with each point as the average V_{obs} ± SD to simplify comparisons between different reactions. For each enzyme, triplicate kinetic data were obtained from two separate purifications. Modest prep-to-prep variability was observed with these enzymes (Table S8). The kinetic parameters described in the main text are from fits to the data from both purifications. (B) Catalytic efficiency comparison of wild-type GriE and ProteinMPNN redesign GM_A9. To calculate the standard error of the fit for k_{cat}/K_M, an alternative form of the Michaelis-Menten equation was used: $V_{obs}=(k_{cat}/K_M)[E]_0[S]/(1+([S]/K_M))$.

Table S8. Kinetic parameters for reactions of wild-type GriE and ProteinMPNN redesign GM_A9.[a]

Enzyme	Source	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M ($M^{-1}s^{-1}$)
Wt GriE	Prep 1	0.84 ± 0.03	164 ± 33	(5.1 ± 0.9) × 10^3
	Prep 2	0.54 ± 0.03	77 ± 22	(7.0 ± 1.7) × 10^3
	Combined data	0.71 ± 0.04	137 ± 33	(5.2 ± 1.0) × 10^3
GM_A9	Prep 1	0.29 ± 0.01	24 ± 11	(1.2 ± 0.2) × 10^4
	Prep 2	0.12 ± 0.004	51 ± 4	(2.3 ± 0.3) × 10^3
	Combined data	0.20 ± 0.02	29 ± 16	(6.9 ± 3.4) × 10^3

[a] Kinetic parameters were obtained from fits to initial rate date (Figure S13). Standard errors are from nonlinear, least squares fits to the initial rate data.



Enzyme	TTN values		
GriE[b]	7800	1800	190
GM_A9[c]	700	420	n.d.
Fold-decrease rel. to wild-type	11x	4.3x	NA

Table S9. Substrate scope of ProteinMPNN redesign GM_A9 compared to wild-type GriE.
[a] Products were determined by comparing the retention times of M+16 products of GM_A9 with wild-type GriE samples.
[b] TTN values for wild-type GriE were taken from a previous publication.[2]
[c] TTN values for GM_A9 products were determined by taking the M+16 product area% value compared to substrate.

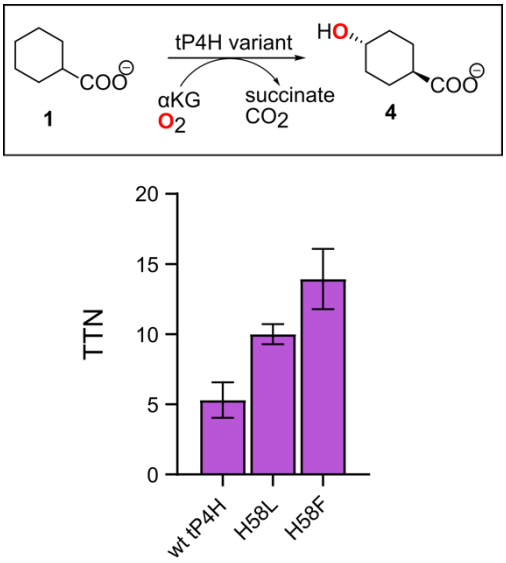


Figure S14. TTN effects for H58L and H58F mutations in wild-type tP4H. H58L is the round 1 winner for directed evolution from wild-type tP4H. H58F is the round 1 winner for directed evolution from the R2_11 variant. Reactions were carried out for 24 hrs at 25 °C using purified enzyme (20 μM) in MES buffer (50 mM, pH 6.8), with 20 mM cyclohexane carboxylic acid **1**, 40 mM αKG, 1 mM ferrous ammonium sulfate, and 1 mM ascorbic acid. Concentration of **4** in quenched reaction samples was quantified by analytical LC-MS analysis. Values are mean ± SD for three replicates.

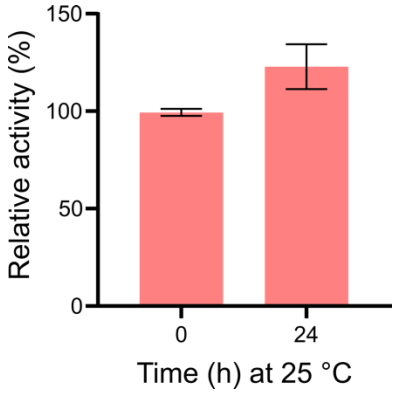


Figure S15. Activity-stability analysis of R2_11 H58F/L174G/V57H at room temperature. Relative activity was determined using PBP assay described above. Values are mean ± SD for three replicates.

2.8 References

[1] H. Renata, Z. J. Wang, F. H. Arnold, *Angew. Chem. Int. Ed.* **2015**, *54*, 3351–3367.
[2] C. Zeymer, D. Hilvert, *Annu. Rev. Biochem.* **2018**, *87*, 1–27.
[3] E. M. Meiering, L. Serrano, A. R. Fersht, *J. Mol. Biol.* **1992**, *225*, 585–589.
[4] B. K. Shoichet, W. A. Baase, R. Kuroki, B. W. Matthews, *Proc. Natl. Acad. Sci.* **1995**, *92*, 452–456.
[5] L. Giver, A. Gershenson, P.-O. Freskgard, F. H. Arnold, *Proc. Natl. Acad. Sci.* **1998**, *95*, 12809–12813.
[6] B. M. Beadle, B. K. Shoichet, *J. Mol. Biol.* **2002**, *321*, 285–296.
[7] R. A. Nagatani, A. Gonzalez, B. K. Shoichet, L. S. Brinen, P. C. Babbitt, *Biochemistry* **2007**, *46*, 6688–6695.
[8] N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, *PLoS Comput. Biol.* **2008**, *4*, e1000002.
[9] N. Tokuriki, C. J. Jackson, L. Afriat-Jurnou, K. T. Wyganowski, R. Tang, D. S. Tawfik, *Nat. Commun.* **2012**, *3*, 1257.
[10] M. Goldsmith, D. S. Tawfik, *Curr. Opin. Struct. Biol.* **2017**, *47*, 140–150.
[11] K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. M. Behbahani, J. J. Weinstein, N. M. Mangan, S. Ovchinnikov, G. J. Rocklin, *Nature* **2023**, *620*, 434–444.
[12] J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, *Proc. Natl. Acad. Sci.* **2006**, *103*, 5869–5874.
[13] Y. Gumulya, J.-M. Baek, S.-J. Wun, R. E. S. Thomson, K. L. Harris, D. J. B. Hunter, J. B. Y. H. Behrendorff, J. Kulig, S. Zheng, X. Wu, B. Wu, J. E. Stok, J. J. D. Voss, G. Schenk, U. Jurva, S. Andersson, E. M. Isin, M. Bodén, L. Guddat, E. M. J. Gillam, *Nat. Catal.* **2018**, *1*, 878–888.
[14] D. L. Trudeau, D. S. Tawfik, *Curr. Opin. Biotechnol.* **2019**, *60*, 46–52.
[15] W. Besenmatter, P. Kast, D. Hilvert, *Proteins: Struct., Funct., Bioinform.* **2007**, *66*, 500–506.
[16] R. D. Socha, N. Tokuriki, *FEBS J.* **2013**, *280*, 5582–5595.
[17] S. D. Stimple, M. D. Smith, P. M. Tessier, *AIChE J.* **2020**, *66*, DOI 10.1002/aic.16814.
[18] Y. Li, D. A. Drummond, A. M. Sawayama, C. D. Snow, J. D. Bloom, F. H. Arnold, *Nat. Biotechnol.* **2007**, *25*, 1051–1056.
[19] P. Heinzelman, C. D. Snow, I. Wu, C. Nguyen, A. Villalobos, S. Govindarajan, J. Minshull, F. H. Arnold, *Proc. Natl. Acad. Sci.* **2009**, *106*, 5610–5615.
[20] A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik, S. J. Fleishman, *Mol. Cell* **2016**, *63*, 337–346.
[21] M. Musil, H. Konegger, J. Hon, D. Bednar, J. Damborsky, *ACS Catal.* **2019**, *9*, 1033–1054.
[22] H. Lu, D. J. Diaz, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. R. Alexander, H. O. Cole, Y. Zhang, N. A. Lynd, A. D. Ellington, H. S. Alper, *Nature* **2021**, *604*, 662–667.
[23] M. S. Islam, T. M. Leissing, R. Chowdhury, R. J. Hopkinson, C. J. Schofield, *Annu. Rev. Biochem.* **2018**, *87*, 585–620.
[24] C. Q. Herr, R. P. Hausinger, *Trends Biochem Sci* **2018**, *43*, 517–532.
[25] A. Papadopoulou, F. Meyer, R. M. Buller, *Biochemistry* **2023**, *62*, 229–240.
[26] C. R. Zwick, H. Renata, *Nat. Prod. Rep.* **2020**, *37*, 1065–1079.
[27] E. Roduner, W. Kaim, B. Sarkar, V. B. Urlacher, J. Pleiss, R. Gläser, W. Einicke, G. A. Sprenger, U. Beifuß, E. Klemm, C. Liebner, H. Hieronymus, S. Hsu, B. Plietker, S. Laschat, *ChemCatChem* **2013**, *5*, 82–112.
[28] C. He, W. G. Whitehurst, M. J. Gaunt, *Chem* **2019**, *5*, 1031–1058.

[29] C. Liu, J. Zhao, J. Liu, X. Guo, D. Rao, H. Liu, P. Zheng, J. Sun, Y. Ma, *Appl. Microbiol. Biotechnol.* **2019**, *103*, 265–277.
[30] W. L. Cheung-Lee, J. N. Kolev, J. A. McIntosh, A. A. Gil, W. Pan, L. Xiao, J. E. Velásquez, R. Gangam, M. S. Winston, S. Li, K. Abe, E. Alwedi, Z. E. X. Dance, H. Fan, K. Hiraga, J. Kim, B. Kosjek, D. N. Le, N. S. Marzijarani, K. Mattern, J. P. McMullen, K. Narsimhan, A. Vikram, W. Wang, J. Yan, R. Yang, V. Zhang, W. Zhong, D. A. DiRocco, W. J. Morris, G. S. Murphy, K. M. Maloney, *Angew. Chem. Int. Ed.* **2024**, *63*, e202316133.
[31] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, *Science* **2022**, *378*, 49–56.
[32] K. H. Sumida, R. Núñez-Franco, I. Kalvet, S. J. Pellock, B. I. M. Wicky, L. F. Milles, J. Dauparas, J. Wang, Y. Kipnis, N. Jameson, A. Kang, J. D. L. Cruz, B. Sankaran, A. K. Bera, G. Jiménez-Osés, D. Baker, *J. Am. Chem. Soc.* **2024**, *146*, 2054–2061.
[33] J. Yang, F.-Z. Li, F. H. Arnold, *ACS Cent. Sci.* **2024**, *10*, 226–241.
[34] B. J. Wittmann, Y. Yue, F. H. Arnold, *Cell Syst.* **2021**, *12*, 1026-1045.e7.
[35] S. d'Oelsnitz, D. J. Diaz, W. Kim, D. J. Acosta, T. L. Dangerfield, M. W. Schechter, M. B. Minus, J. R. Howard, H. Do, J. M. Loy, H. S. Alper, Y. J. Zhang, A. D. Ellington, *Nat. Commun.* **2024**, *15*, 2084.
[36] K. Ding, M. Chin, Y. Zhao, W. Huang, B. K. Mai, H. Wang, P. Liu, Y. Yang, Y. Luo, *Nat. Commun.* **2024**, *15*, 6392.
[37] J. Yang, R. G. Lal, J. C. Bowden, R. Astudillo, M. A. Hameedi, S. Kaur, M. Hill, Y. Yue, F. H. Arnold, *bioRxiv* **2024**, 2024.07.27.605457.
[38] N. Thomas, D. Belanger, C. Xu, H. Lee, K. Hirano, K. Iwai, V. Polic, K. D. Nyberg, K. G. Hoff, L. Frenz, C. A. Emrich, J. W. Kim, M. Chavarha, A. Ramanan, J. J. Agresti, L. J. Colwell, *bioRxiv* **2024**, 2024.03.21.585615.
[39] C. Klein, W. Hüttel, *Adv. Synth. Catal.* **2011**, *353*, 1375–1383.
[40] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, *Nat. Methods* **2022**, *19*, 679–682.
[41] X. Hu, X. Huang, J. Liu, P. Zheng, W. Gong, L. Yang, *Acta Crystallogr. Sect. D* **2023**, *79*, 318–325.
[42] P. Lukat, Y. Katsuyama, S. Wenzel, T. Binz, C. König, W. Blankenfeldt, M. Brönstrup, R. Müller, *Chem Sci* **2017**, *8*, 7521–7527.
[43] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25*, 1605–1612.
[44] N. W. Goldberg, A. M. Knight, R. K. Zhang, F. H. Arnold, *J. Am. Chem. Soc.* **2019**, *141*, 19585–19588.
[45] C. R. Zwick, H. Renata, *J. Am. Chem. Soc.* **2018**, *140*, 1165–1169.
[46] S. Kille, C. G. Acevedo-Rocha, L. P. Parra, Z.-G. Zhang, D. J. Opperman, M. T. Reetz, J. P. Acevedo, *ACS Synth. Biol.* **2013**, *2*, 83–92.
[47] Y. Peleg, R. Vincentelli, B. M. Collins, K.-E. Chen, E. K. Livingstone, S. Weeratunga, N. Leneva, Q. Guo, K. Remans, K. Perez, G. E. K. Bjerga, Ø. Larsen, O. Vaněk, O. Skořepa, S. Jacquemin, A. Poterszman, S. Kjær, E. Christodoulou, S. Albeck, O. Dym, E. Ainbinder, T. Unger, A. Schuetz, S. Matthes, M. Bader, A. de Marco, P. Storici, M. S. Semrau, P. Stolt-Bergner, C. Aigner, S. Suppmann, A. Goldenzweig, S. J. Fleishman, *J. Mol. Biol.* **2021**, *433*, 166964.

[48] R. Shroff, A. W. Cole, D. J. Diaz, B. R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A. D. Ellington, R. Thyer, *ACS Synth. Biol.* **2020**, *9*, 2927–2935.
[49] M. P. Okoh, J. L. Hunter, J. E. T. Corrie, M. R. Webb, *Biochemistry* **2006**, *45*, 14764–14771.
[50] L. Luo, M. B. Pappalardi, P. J. Tummino, R. A. Copeland, M. E. Fraser, P. K. Grzyska, R. P. Hausinger, *Anal. Biochem.* **2006**, *353*, 69–74.
[51] J. C. Nolte, M. Schürmann, C.-L. Schepers, E. Vogel, J. H. Wübbeler, A. Steinbüchel, *Appl. Environ. Microbiol.* **2014**, *80*, 166–176.
[52] J. Mattay, S. Houwaart, W. Hüttel, *Appl. Environ. Microbiol.* **2018**, *84*, e02370-17.
[53] M. Remmert, A. Biegert, A. Hauser, J. Söding, *Nat. Methods* **2012**, *9*, 173–175.
[54] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589.
[55] T. Wang, X. Jin, X. Lu, X. Min, S. Ge, S. Li, *Front. Genet.* **2024**, *14*, 1347667.
[56] M. Davidson, M. McNamee, R. Fan, Y. Guo, W.-C. Chang, *J. Am. Chem. Soc.* **2019**, *141*, 3419–3423.
[57] J. Mattay, W. Hüttel, *Chembiochem* **2017**, *18*, 1523–1528.