Enhancing computational enzyme design by a maximum entropy strategy

Wen Jun Xie[a,1], Mojgan Asadi[a], and Arieh Warshel[a,1]

[a]Department of Chemistry, University of Southern California, Los Angeles, CA 90089-1062

Contributed by Arieh Warshel; received December 10, 2021; accepted January 3, 2022; reviewed by Johan Åqvist and Jenn-Kang Hwang

Although computational enzyme design is of great importance, the advances utilizing physics-based approaches have been slow, and further progress is urgently needed. One promising direction is using machine learning, but such strategies have not been established as effective tools for predicting the catalytic power of enzymes. Here, we show that the statistical energy inferred from homologous sequences with the maximum entropy (MaxEnt) principle significantly correlates with enzyme catalysis and stability at the active site region and the more distant region, respectively. This finding decodes enzyme architecture and offers a connection between enzyme evolution and the physical chemistry of enzyme catalysis, and it deepens our understanding of the stability–activity trade-off hypothesis for enzymes. Overall, the strong correlations found here provide a powerful way of guiding enzyme design.

enzyme design | maximum entropy | evolution | catalysis

Enzymes are extraordinary catalysts that play vital roles in nearly all biochemical processes. Designing efficient enzymes could help in solving threats to humankind, including the energy crisis, environmental pollution, and food shortages (1). The use of computational modeling for enzyme design is very promising (2–6). However, such approaches are still not at the stage where they can guide sufficiently reliable enzyme design (7–9). Thus, it is crucial to exploit additional options for improving the design predictability. This work will explore the potential of statistical analysis of enzyme homologous sequences for enhancing computational enzyme design prediction.

Naturally evolving enzymes can speed up chemical reactions by many orders of magnitude (e.g., Fig. 1 *A* and *B*). Such a great catalytic power reflects a very long evolutionary process that started at the emergence of life. In principle, it is tempting to study the origin of the catalytic power of enzymes using physics-based models. However, machine learning methods may provide an invaluable guide. The maximum entropy (MaxEnt) principle (10) offers the least-biased model for the sequence distribution by maximizing information entropy subjected to the statistics obtained from multiple sequence alignment (MSA). The MaxEnt model taking epistasis into account has been proposed to distill evolutionary information within a protein family, which was then correlated with residue–residue contact (11–13) and fitness (14, 15), partly leading to the breakthrough of protein structure prediction (16). For enzymes, a high correlation between the statistical energy derived from the MaxEnt model and enzyme efficiency for beta-lactamase was found, but it did not seem to work for trypsin and dihydrofolate reductase (DHFR) (14). This generative model has been recently used to design enzymes (17). The MaxEnt model on its current form can classify designed sequences in a binary way as functional or nonfunctional based on a regression model trained with the statistical energy and additional high-throughput experimental data. However, designed sequences chosen for biochemical analysis do not show improved catalytic power compared with natural sequences (17). Therefore, the MaxEnt approach has not reached the stage of rational enzyme design, where one should be able to accurately predict the effect of mutations on catalytic power and design better enzymes. This might not be that surprising considering the complex interplay among various selection pressures applied to enzyme evolution (18). In particular, enzyme stability and activity may trade off with each other (19–22).

This work explores the hypothesis that the enzyme catalytic center involved in the catalysis and transition-state stabilization directly correlates with the selection pressure of enzyme efficiency in a way that can be captured by the MaxEnt model. This idea is confirmed by finding a significant correlation for the catalytic center between enzyme catalysis and statistical energy derived by applying the MaxEnt model (Fig. 1*C*). In contrast, the statistical energy correlates well with protein stability for remote regions (referred to here as enzyme surface), suggesting that a stable enzyme surface may be needed for optimal enzyme function. Therefore, the results here show that evolutionary information can be used to decode enzyme architecture and understand biocatalysis. Furthermore, we demonstrate that the widely used consensus design is a special case of the MaxEnt model. The correlations and insights thus offer a powerful way to guide enzyme design.

The MaxEnt Model

Homologous enzyme sequences from different species share the same evolutionary origin (23). The natural sequence variation within an enzyme family is constrained by different factors,

Significance

Designing efficient enzymes could contribute to a sustainable future. Current computational approaches, including physics-based and machine learning–based design, have not led to a robust enzyme design. Predicting enzyme catalytic power is the crucial step for enzyme design. Here, we found that the properties of enzymes are correlated in a nontrivial way with their evolutionary information. For the active site region and the more distant region, the statistical energy obtained from the maximum entropy model for enzyme homologs is strongly correlated with enzyme catalytic power and stability, respectively. The findings here could be used to understand enzyme catalysis and evolution. Combining the present approach with physics-based computer modeling can provide a potent tool for enzyme design.

Author contributions: W.J.X. and A.W. designed research; W.J.X., M.A., and A.W. performed research; W.J.X. and M.A. contributed new reagents/analytic tools; W.J.X. and A.W. analyzed data; and W.J.X. and A.W. wrote the paper.

Reviewers: J.Å., Uppsala Universitet; and J.-K.H., National Yang Ming Chiao Tung University.

Competing interest statement: W.J.X. and A.W. filed a provisional patent application by the University of Southern California on enzyme engineering in July 2021 (US application no. 63/234,099).

[1]To whom correspondence may be addressed. Email: xwj123@gmail.com or warshel@usc.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2122355119/-/DCSupplemental.

Published February 8, 2022.

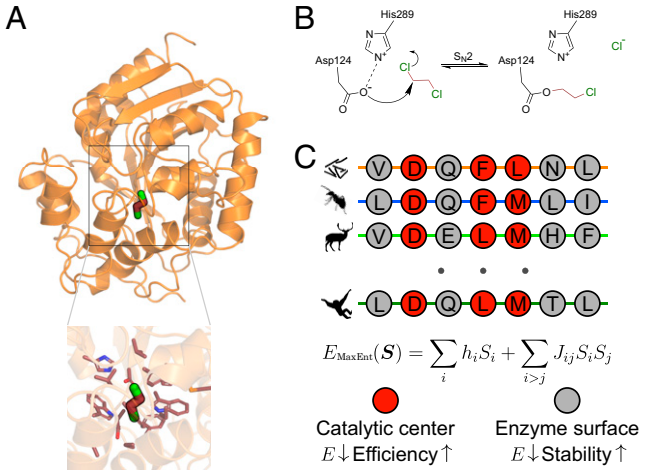

Fig. 1. The MaxEnt model for enzyme sequences connects enzyme evolution and function. (*A* and *B*) The enzyme accelerates chemical reaction by lowering the activation energy using mainly the residues in the catalytic center. Haloalkane dehalogenase (PDB ID code 2dhc) is used as an example to illustrate enzyme catalysis and the reaction mechanism. (*A*) The residues within a distance of 7.0 Å from the substrate are highlighted. (*B*) The scheme of the substitution nucleophilic (S_N2) step is illustrated using the substrate of 1,2-dichloroethane. (*C*) The MaxEnt model connects enzyme evolution to the physical chemistry of enzyme catalysis. A pairwise MaxEnt model is learned from the MSA, and each protein sequence ($\boldsymbol{S}$) is associated with statistical energy (E_{MaxEnt}) following the Boltzmann distribution. We found that decreasing the statistical energy significantly correlates with increasing enzyme efficiency and stability in the catalytic center and enzyme surface, respectively.

including its physical chemistry (24). Therefore, distilling evolutionary information from MSA of an enzyme family could shed light on enzyme three-dimensional structure and function. Due to limited homologous sequences and high computational cost, the MaxEnt model is usually truncated to consider pairwise epistatic effect. The MaxEnt model provides a Boltzmann distribution $P(\boldsymbol{S}) = e^{-E_{\text{MaxEnt}}(\boldsymbol{s})}/Z$ for each sequence $\boldsymbol{S}$, where E_{MaxEnt} is the statistical energy with effective temperature as unity and Z is the partition function:

$$E_{\text{MaxEnt}}(\boldsymbol{S}) = \sum_i h_i S_i + \sum_{i>j} J_{ij} S_i S_j,$$

$$Z = \sum_{\boldsymbol{S}} e^{-E_{\text{MaxEnt}}(\boldsymbol{s})}.$$

The parameters h_i and J_{ij} are site energy and pairwise coupling between amino acids at two different residue sites, respectively. The E_{MaxEnt} is shifted by a constant so that the wild type (WT) has a zero value, which will not affect any results due to gauge invariance. A lower E_{MaxEnt} for a sequence indicates a higher probability to appear during evolution and might reflect a particular evolutionary advantage.

The statistical energy E_{MaxEnt} corresponds to a spin-glass Hamiltonian, which has enormous local frustrations (25). The parameterization, which requires extensive sampling of the model, is thus highly nontrivial, especially for large proteins with hundreds of residues. Instead of using the popular pseudolikelihood (PLL) approximation (13, 14), we have previously developed an efficient code that marries different computational advancements to sample the Hamiltonian rigorously (26). The derivation and parameterization details of the MaxEnt model can be found in *SI Appendix*, *Text*. The PLL approximation is unable to reproduce the statistics of natural sequences (27). Here, the excellent reproduction and prediction of natural MSA statistics validate our implementation (*SI Appendix*, Fig. S1).

Results

A critical obstacle to examining the enzyme evolution–catalysis relationship is the lack of enzyme catalytic data covering sufficient mutants for the target enzyme. Although deep mutational scanning can measure the consequence of mutation at scale, the relation between its readout and enzyme physical properties is uncertain (28). Directly measuring the catalytic parameters (k_{cat}, $k_{\text{cat}}/K_{\text{M}}$, and k_{obs}) requires laborious biochemical assays (29). The experimental data are thus relatively sparse. To this end, we first manually curated a database for enzyme efficiency upon mutation from published literature (*SI Appendix*, Tables S1–S9), focusing on the systems with many mutations either in the catalytic center (here defined as within 7.0 Å from the substrate) or the enzyme surface (here defined as beyond 9.0 Å from the substrate). Many of the enzymes here are model systems in computational chemistry studies. The database contains 12 enzyme–substrate pairs, and for each pair, at least seven mutations measured in similar conditions (pH and temperature) are collected. The protein stability data were also included whenever available; the database includes many higher-order mutations (up to the 10th order). Meanwhile, we made sure that each enzyme has thousands of homologous sequences in the MSA to get statistically meaningful evolutionary information (*SI Appendix*, Table S10). The enzymes studied here cover various types of reactions identified by their different Enzyme Commission class number (*SI Appendix*, Table S10).

We started with the haloalkane dehalogenase from *Xanthobacter autotrophicus* that catalyzes the conversion of toxic haloalkanes to alcohols (Fig. 1 *A* and *B*) (30). We evaluated the correlation between the statistical energy E_{MaxEnt} and the observed enzyme catalytic power (expressed by both $\log k_{\text{cat}}/K_{\text{M}}$ and $\log k_{\text{cat}}$). All the mutations are located in the catalytic center with a mean distance of 3.4 Å from the substrate (Fig. 2*A*). Except for one double mutation, the other six are single mutations. The enzymatic rates span more than six orders of magnitude, posing great challenges for prediction methods. Nevertheless, as seen from Fig. 2*B*, the E_{MaxEnt} shows impressive Pearson correlations with $\log k_{\text{cat}}/K_{\text{M}}$ and $\log k_{\text{cat}}$ with values of −0.87 and −0.95, respectively.

We then explored the catalytic center of chorismate mutase, which is widely used in enzyme mechanism and design studies (31). This enzyme transforms chorismite to prephenate in the pathway to produce tyrosine and phenylalanine, essential for plants, fungi, and bacteria (29). The enzyme mutations for chorismate mutase from *Escherichia coli* are 3.7 Å from the substrate on average (Fig. 2*C*). Here again, the correlations are significant, and the E_{MaxEnt} has a correlation value of –0.68 with $\log k_{\text{cat}}/K_{\text{M}}$ (Fig. 2*D*). The A32S mutation stands out as the only mutant with increased efficiency relative to the WT; our approach also detected such a unique experimental result. The MaxEnt model has been recently applied to chorismite mutase to explore the sequence space of the whole enzyme (17). However, the E_{MaxEnt} for the whole enzyme is not that informative for catalytic power; only after combining with fitness data from high-throughput experiments can it train a logistic regression model to binary classify whether a designed sequence is functional or not (17).

For both haloalkane dehalogenase and chorismite mutase, the mutants are mainly single mutations. One may wonder what the performance of the MaxEnt model on higher-order mutations is. We then considered alcohol dehydrogenase from *Starmerella magnoliae*; nine of the 20 mutants having experimental kinetic data are higher-order mutations up to the 10th order (32). The average distance between the mutations and substrate is 6.2 Å (Fig. 2*E*). Here, the cofactor $NADP^+$ and catalytic triad were used as the reference point in calculating the distance because of the absence of substrate in the Protein Data Bank (PDB) structure; note that not every residue involved in this case is very close to the substrate. The

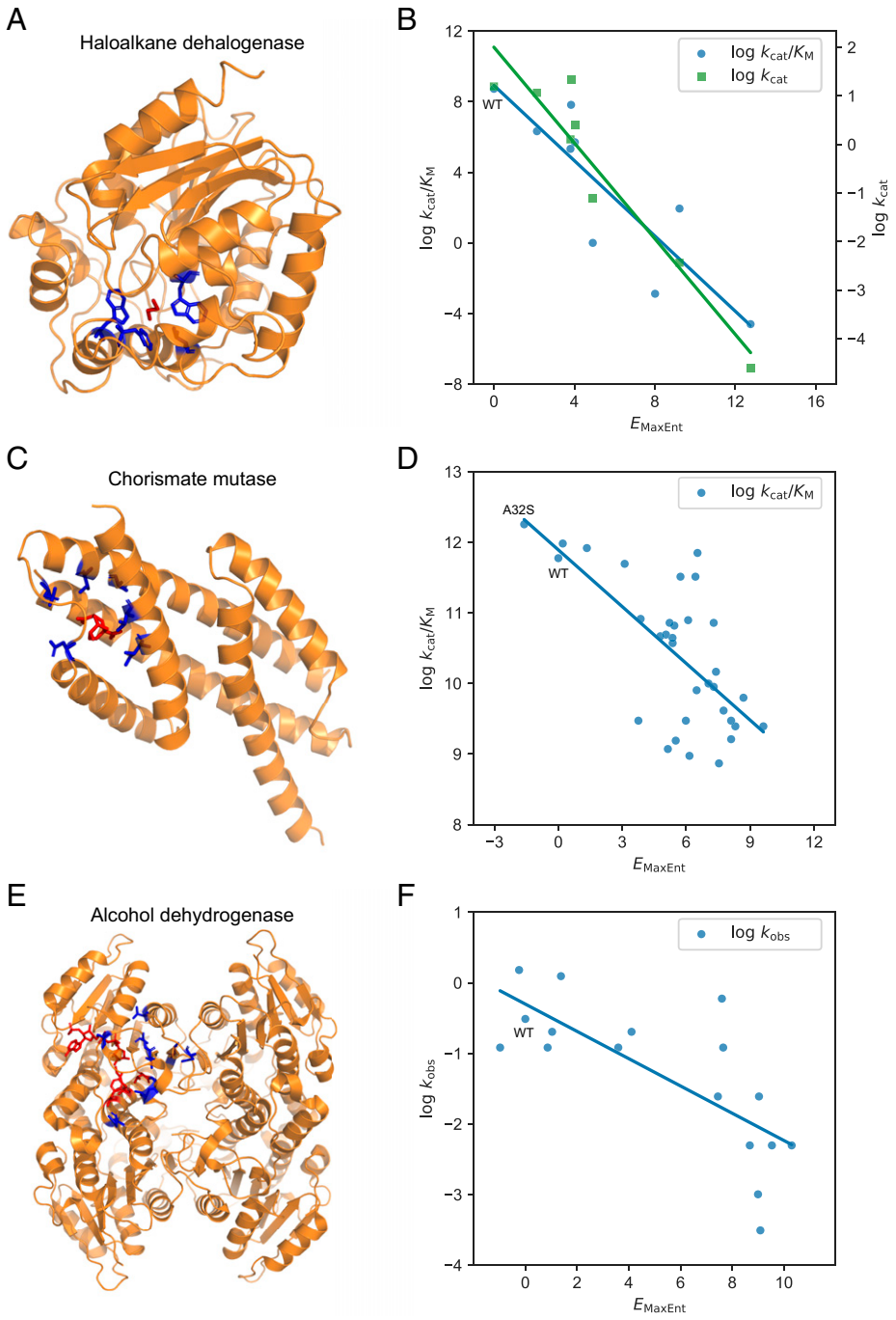

Fig. 2. The MaxEnt model for enzyme sequences correlates with enzyme efficiency at the catalytic center. (*A* and *B*) Haloalkane dehalogenase. (*C* and *D*) Chorismate mutase. (*E* and *F*) Alcohol dehydrogenase. (*A*, *C*, and *E*) Substrates and mutated residues in the dataset are shown in red and blue, respectively; only one unit of the dimeric chorismate mutase and the tetrameric alcohol dehydrogenase is highlighted. For alcohol dehydrogenase, the cofactor $NADP^+$ and catalytic triad are colored in red because of the absence of substrate. PDB ID codes used in rendering the structures are (*A*) 2dhc, (*C*) 1ecm, and (*E*) 6tq5. Substrates are (*A*) 1,2-dichloroethane, (*C*) chorismate, and (*E*) cyclohexanol. (*B*, *D*, and *F*) Correlations between E_{MaxEnt} and experimental catalytic power. The least-squares regression line is plotted for each enzyme; the WT enzyme has a zero value of E_{MaxEnt}.

E_{MaxEnt}–logk_{obs} correlation is −0.74 (Fig. 2*E*). The successful prediction for such higher-order mutations underscores the potential of our approach. Interestingly, the E_{MaxEnt} strongly correlates with T_m (correlation value of 0.91) but with the opposite trend as the catalytic efficiency, supporting the activity–stability trade-off proposal (19–22). In this case, the independent model without epistasis shows opposing trends as the MaxEnt model (*SI Appendix*, Table S12). If we further dissect the dataset into two subdatasets, one contains all single mutations, and the remaining are in the other. The independent model shows opposite trends between the two subdatasets. The results demonstrate the importance of considering epistasis in extracting evolutionary infromation.

Next, we examined the generality of our finding by considering an extensive set of enzymes summarized in Table 1. For all the mutants in the catalytic center, we observed a strong correlation between the MaxEnt model and the catalytic effect. Although the collected data for some enzymes has a biased

Table 1. Correlation between the MaxEnt model and enzyme efficiency/T_m

Enzyme	E_{MaxEnt}–logk_{cat}/K_M	E_{MaxEnt}–logk_{cat}	E_{MaxEnt}–T_m	Distance to substrate (Å)	Mutation region	Substrate
Haloalkane dehalogenase	−0.87	−0.95		3.4 ± 0.4	Catalytic center	1,2-dichloroethane
Chorismate mutase	−0.68	−0.51	−0.20	3.7 ± 0.8	Catalytic center	Chorismate
Alcohol dehydrogenase*	−0.74		0.91	6.2 ± 3.8	Catalytic center	Cyclohexanol
Triosephosphate isomerase	−0.62	−0.69		3.9 ± 1.6	Catalytic center	Dihydroxyacetone phosphate
Ketosteroid isomerase	−0.74	−0.75		2.1 ± 0.5	Catalytic center	5(10)-estrene-3,17-dione
	−0.81	−0.77		3.4 ± 1.3	Catalytic center	5-androstenedione
DHFR†		−0.74		5.4 ± 2.3	Catalytic center	NADPH
		0.08		11.4 ± 3.1	Enzyme surface	
DHFR‡	−0.42	0.16	−0.65	10.6 ± 3.8	Enzyme surface	NADPH
Beta-lactamase	−0.57	−0.48	−0.68	10.4 ± 5.8	Enzyme surface	6-furylacrylpenicillanic acid
	−0.37	0.04		16.4 ± 4.8		Ampicillin
Trypsin	0.06	−0.03	−0.65	10.1 ± 5.2	Enzyme surface	Suc-Ala-Ala-Pro-Lys-PNA

The correlation values throughout this manuscript are the Pearson correlation coefficients. The two regions highlighted in the table are either catalytic center (highly correlated with enzyme catalysis) or enzyme surface (highly correlated with enzyme stability). *P* values from two-tailed test for the correlations in the highlighted region are shown in *SI Appendix*, Table S14.

*The enzyme kinetics is measured as logk_{obs} (*SI Appendix*, Table S3) and collected from ref. 32.

†Data are collected from ref. 33 (*SI Appendix*, Table S6).

‡Data are collected from ref. 34 (*SI Appendix*, Table S7).

choice of mutants in experiment, the correlations are consistently strong. The correlation seems insensitive to substrates for ketosteroid isomerase. However, for DHFR, the strong correlation disappears when moving from the catalytic center to the enzyme surface. Such results confirm our hypothesis that the catalytic center evolved under the selection pressure of optimizing enzyme catalysis.

In addition, the correlation obtained here using the rigorous sampling of the Hamiltonian is slightly stronger than those using PLL approximation, but both of them are better than the independent model (*SI Appendix*, *Text* and Tables S12 and S13). The consistent results obtained from the PLL approximation again confirm our findings.

For the enzyme surface regions, which are at least 9.0 Å away from the substrate (enzyme surface), the correlation between E_{MaxEnt} and enzyme efficiency is not that strong or systematic, although in general, there seems to be a negative correlation. Using beta-lactamase as an example (and discarding the substrate difference in two enzyme–substrate pairs), E_{MaxEnt} has a stronger correlation with enzyme catalysis for mutations closer to the substrate. This is again consistent with our above finding for the catalytic center. The rationale is that the surface region is not directly responsible for the evolution pressure of enzyme catalysis.

To better understand the physical nature of E_{MaxEnt}, we also considered in Table 1 the correlation between E_{MaxEnt} and the observed T_m (which is inversely related to the protein folding energy). As seen from Table 1, we have a systematically negative correlation between E_{MaxEnt} and T_m for the enzyme surface, indicating that the MaxEnt model does reflect the protein stability for regions far away from where catalysis happens. It is reasonable since the MaxEnt model reflects the contact probability (11–13), which can be considered as a generalized free energy function for protein folding (35).

It appears that the catalytic center and enzyme surface face different selection pressures. The statistical energy inferred from MSA strongly inversely correlated with enzyme efficiency and enzyme stability in the catalytic center and enzyme surface, respectively. The finding that a more stable enzyme surface might help in promoting enzyme catalysis could also rationalize the growing evidence that it is possible to engineer remote mutations to improve catalysis (Table 1) (33, 36).

To demonstrate our approach to enzyme design, we redesigned the catalytic center of haloalkane dehalogenase after parameterization of the MaxEnt model; 37% of the designed sequences have lower E_{MaxEnt} than the WT (*SI Appendix*, Fig. S2*A*), suggesting possible enhanced catalysis. Interestingly, one of the top five designs is a consensus design, where the residue is replaced by the most frequently observed amino acid in the natural MSA (*SI Appendix*, Fig. S2 *B–D*). Consensus design has already been shown to be effective in protein engineering (37, 38), and it turns out to be a special case of the MaxEnt model where epistasis is considered. Our results thus provide a statistical basis for the consensus design and suggest that the consensus amino acids near the substrate are likely to improve enzyme catalytic power.

Discussion

This work explored the relationship between enzyme evolution and catalysis by correlating E_{MaxEnt} obtained from natural homologous sequences with the catalytic power of different enzymes. It is found that the correlation is significant for the catalytic center, and adopting our finding to guide enzyme design is straightforward. The catalytic center and enzyme surface face different selection pressures. Therefore, it is more likely to improve enzyme catalysis by optimizing the catalytic center instead of the enzyme surface using evolutionary information. This also explains why there are no consistent correlations between catalytic power and evolutionary information in previous studies (14, 17); the complex physical constraints make it highly nontrivial to predict enzyme efficiency from sequence data. For trypsin and DHFR, the mutations in the dataset (14) are on the enzyme surface, which is not likely to show a strong correlation between E_{MaxEnt} and catalytic power; such an explanation also applies to cases when the whole enzyme is studied (17). Adopting machine learning to protein studies is promising (14, 15, 17, 39–46); here, we incorporate the understanding of the physical chemistry of enzymes into machine learning and thus, could obtain a consistent prediction for biocatalysis. For simple protein (e.g., protein without domains for catalysis and other complex functions), evolutionary information may be simply correlated with protein stability.

While we can use the correlation to estimate enzyme catalytic power, it is interesting to look for some possible rationalization. Thus, we checked the correlation between E_{MaxEnt} and T_m. Alcohol dehydrogenase provides strong evidence for the enzyme activity–stability trade-off. However, for enzyme surface, E_{MaxEnt} and T_m are inversely correlated, while T_m is roughly correlated with enzyme efficiency. This seems to contradict the idea that catalytic preorganization costs folding

energy (19, 20). However, the folding energy as expressed by T_m is related to the stability of the entire enzyme, and the preorganization can be determined by the folding of a limited part of the enzyme (31). As shown in our previous study on DHFR, reducing the reorganization energy may or may not reduce the protein stability since it requires protein restraints in specific directions along the reaction coordinate and not necessarily restraints in all directions (20). Interestingly, the role of the protein surface in helping catalysis found here may not directly be related to the reduction in the reorganization energy (47).

In addition to striving to understand the nature of the correlation found in this work, we can also take a very pragmatic "engineering" approach employing the correlation between E_{MaxEnt} and k_{cat}. That is, regardless of whether the correlation is negative or positive, we can generate mutants, determine their E_{MaxEnt}, choose those with increased k_{cat}, and further screen them with the empirical valence bond (EVB) calculation (2, 7) for design experiments.

We anticipate that the emerging evolutionary information, with the rapidly accumulating genomic sequence data, will facilitate studies of more enzyme families. For cases where such information is not sufficient or for new catalytic reactions, the performance of the MaxEnt model might not be sufficient. In such cases, we can further combine with EVB calculations (2, 7) to model the catalytic power and further screen the design to increase the success rate. The findings here seem to be a general principle for enzymes but require a thorough examination of more enzymes. We indeed found more enzymes to support the conclusions present here while this manuscript was under review; the results will be presented in a forthcoming manuscript. Whether the enzyme architecture can be further decoded into more categories with evolutionary information also needs to be investigated. Furthermore, enzymes may be far more intricate than we currently know; it would be exciting to understand the coupling between different enzyme parts from evolutionary information.

The great potential of laboratory-directed evolution has been demonstrated (48). In this respect, we believe that our approach can help in extending the design by directed evolution. Moreover, our approach can be used to trace the moves in directed evolution by following the prediction of the MaxEnt model to understand how enzymes evolve. It could also be applied to understand natural evolution, considering ancestral sequences are homologs of the extant sequences.

In summary, we decoded enzyme architecture using evolutionary information and connected enzyme evolution with enzyme catalysis. Such a connection can help to bridge evolutionary biology and enzymology. The high-throughput and predictability from the MaxEnt model (or other generative models) combined with experimental validation and computational modeling could push enzyme studies to a systems level. Significantly, the results here call attention to integrating domain knowledge in physical chemistry into machine learning models for protein engineering.

Data Availability. All study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. This work was supported by NIH Grant R35 GM122472 and NSF Grant MCB 1707167. We thank the University of Southern California High Performance Computing and Communication Center for computational resources.

1. U. T. Bornscheuer *et al.*, Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
2. A. Warshel, Multiscale modeling of biological functions: From enzymes to molecular machines (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* **53**, 10020–10031 (2014).
3. G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker, K. N. Houk, Computational enzyme design. *Angew. Chem. Int. Ed. Engl.* **52**, 5700–5725 (2013).
4. V. Vaissier Welborn, T. Head-Gordon, Computational design of synthetic enzymes. *Chem. Rev.* **119**, 6613–6630 (2019).
5. S. Hammes-Schiffer, Catalysts by design: The power of theory. *Acc. Chem. Res.* **50**, 561–566 (2017).
6. H. K. Privett *et al.*, Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3790–3795 (2012).
7. D. Mondal, V. Kolev, A. Warshel, Combinatorial approach for exploring conformational space and activation barriers in computer-aided enzyme design. *ACS Catal.* **10**, 6002–6012 (2020).
8. R. B. Leveson-Gower, C. Mayer, G. Roelfes, The importance of catalytic promiscuity for enzyme design and evolution. *Nat. Rev. Chem.* **3**, 687–705 (2019).
9. W. S. Mak, J. B. Siegel, Computational enzyme design: Transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).
10. E. T. Jaynes, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
11. D. S. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
12. F. Morcos *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
13. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013).
14. T. A. Hopf *et al.*, Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
15. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Mol. Biol. Evol.* **33**, 268–280 (2016).
16. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
17. W. P. Russ *et al.*, An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
18. D. Davidi, L. M. Longo, J. Jabłońska, R. Milo, D. S. Tawfik, A bird's-eye view of enzyme evolution: Chemical, physicochemical, and physiological considerations. *Chem. Rev.* **118**, 8786–8797 (2018).
19. A. Warshel, Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 5250–5254 (1978).
20. M. Roca, H. Liu, B. Messer, A. Warshel, On the relationship between thermal stability and catalytic power of enzymes. *Biochemistry* **46**, 15076–15088 (2007).
21. B. K. Shoichet, W. A. Baase, R. Kuroki, B. W. Matthews, A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 452–456 (1995).
22. B. M. Beadle, B. K. Shoichet, Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.* **321**, 285–296 (2002).
23. E. V. Koonin, Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).
24. M. J. Harms, J. W. Thornton, Evolutionary biochemistry: Revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571 (2013).
25. Y. Roudi, E. Aurell, J. A. Hertz, Statistical physics of pairwise probability models. *Front. Comput. Neurosci.* **3**, 22 (2009).
26. W. J. Xie, B. Zhang, Learning the formation mechanism of domain-level chromatin states with epigenomics data. *Biophys. J.* **116**, 2047–2056 (2019).
27. M. Figliuzzi, P. Barrat-Charlaix, M. Weigt, How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35**, 1018–1027 (2018).
28. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nat. Methods.* **11**, 801–807 (2014).
29. D. L. Nelson, M. M. Cox, *Lehninger Principles of Biochemistry* (W. H. Freeman, New York, NY, ed. 4, 2005).
30. D. B. Janssen, Evolving haloalkane dehalogenases. *Curr. Opin. Chem. Biol.* **8**, 150–159 (2004).
31. A. Warshel *et al.*, Electrostatic basis for enzyme catalysis. *Chem. Rev.* **106**, 3210–3235 (2006).
32. F. S. Aalbers *et al.*, Approaching boiling point stability of an alcohol dehydrogenase through computationally-guided enzyme engineering. *eLife* **9**, e54639 (2020).
33. J. Lee, N. M. Goodey, Catalytic contributions from remote regions of enzyme structure. *Chem. Rev.* **111**, 7595–7624 (2011).
34. S. Bershtein, W. Mu, E. I. Shakhnovich, Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4857–4862 (2012).
35. F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, P. G. Wolynes, Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12408–12413 (2014).
36. M. Wilding, N. Hong, M. Spence, A. M. Buckle, C. J. Jackson, Protein engineering: The potential of remote mutations. *Biochem. Soc. Trans.* **47**, 701–711 (2019).
37. B. T. Porebski, A. M. Buckle, Consensus protein design. *Protein Eng. Des. Sel.* **29**, 245–251 (2016).
38. M. Sternke, K. W. Tripp, D. Barrick, Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11275–11284 (2019).
39. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
40. P. Tian, J. M. Louis, J. L. Baber, A. Aniana, R. B. Best, Co-evolutionary fitness landscapes for sequence design. *Angew. Chem. Int. Ed. Engl.* **57**, 5674–5678 (2018).

41. X. Ding, Z. Zou, C. L. Brooks III, Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **10**, 5644 (2019).
42. Z. Wu, K. E. Johnston, F. H. Arnold, K. K. Yang, Protein sequence design with deep generative models. *Curr. Opin. Chem. Biol.* **65**, 18–27 (2021).
43. T. Bepler, B. Berger, Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021).
44. V. Frappier, A. E. Keating, Data-driven computational protein design. *Curr. Opin. Struct. Biol.* **69**, 63–69 (2021).
45. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
46. J. E. Shin *et al.*, Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).
47. G. V. Isaksen, J. Åqvist, B. O. Brandsdal, Enzyme surface rigidity tunes the temperature dependence of catalytic rates. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7822–7827 (2016).
48. P. A. Romero, F. H. Arnold, Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).