ACS SyntheticBiology

pubs.acs.org/synthbio

Research Article

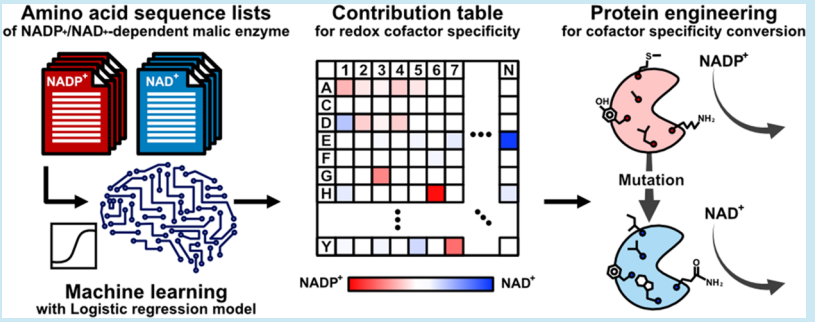
Logistic Regression-Guided Identification of Cofactor Specificity-Contributing Residues in Enzyme with Sequence Datasets Partitioned by Catalytic Properties

Sou Sugiki, Teppei Niide,* Yoshihiro Toya, and Hiroshi Shimizu*

Cite This: *ACS Synth. Biol.* 2022, 11, 3973–3985

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Changing the substrate/cofactor specificity of an enzyme requires multiple mutations at spatially adjacent positions around the substrate pocket. However, this is challenging when solely based on crystal structure information because enzymes undergo dynamic conformational changes during the reaction process. Herein, we proposed a method for estimating the contribution of each amino acid residue to substrate specificity by deploying a phylogenetic analysis with logistic regression. Since this method can estimate the candidate amino acids for mutation by ranking, it is readable and can be used in protein engineering. We demonstrated our concept using redox cofactor conversion of the *Escherichia coli* malic enzyme as a model, which still lacks crystal structure elucidation. The use of logistic regression with amino acid sequences classified by cofactor specificity showed that the $NADP^+$-dependent malic enzyme completely switched cofactor specificity to NAD^+ dependence without the need for a practical screening step. The model showed that surrounding residues made a greater contribution to cofactor specificity than those in the interior of the substrate pocket. These residues might be difficult to identify from crystal structure observations. We show that a highly accurate and inferential machine learning model was obtained using amino acid sequences of structurally homologous and functionally distinct enzymes as input data.

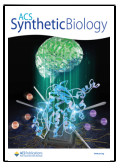
KEYWORDS: *machine learning, enzyme engineering, cofactor specificity conversion, consensus sequence*

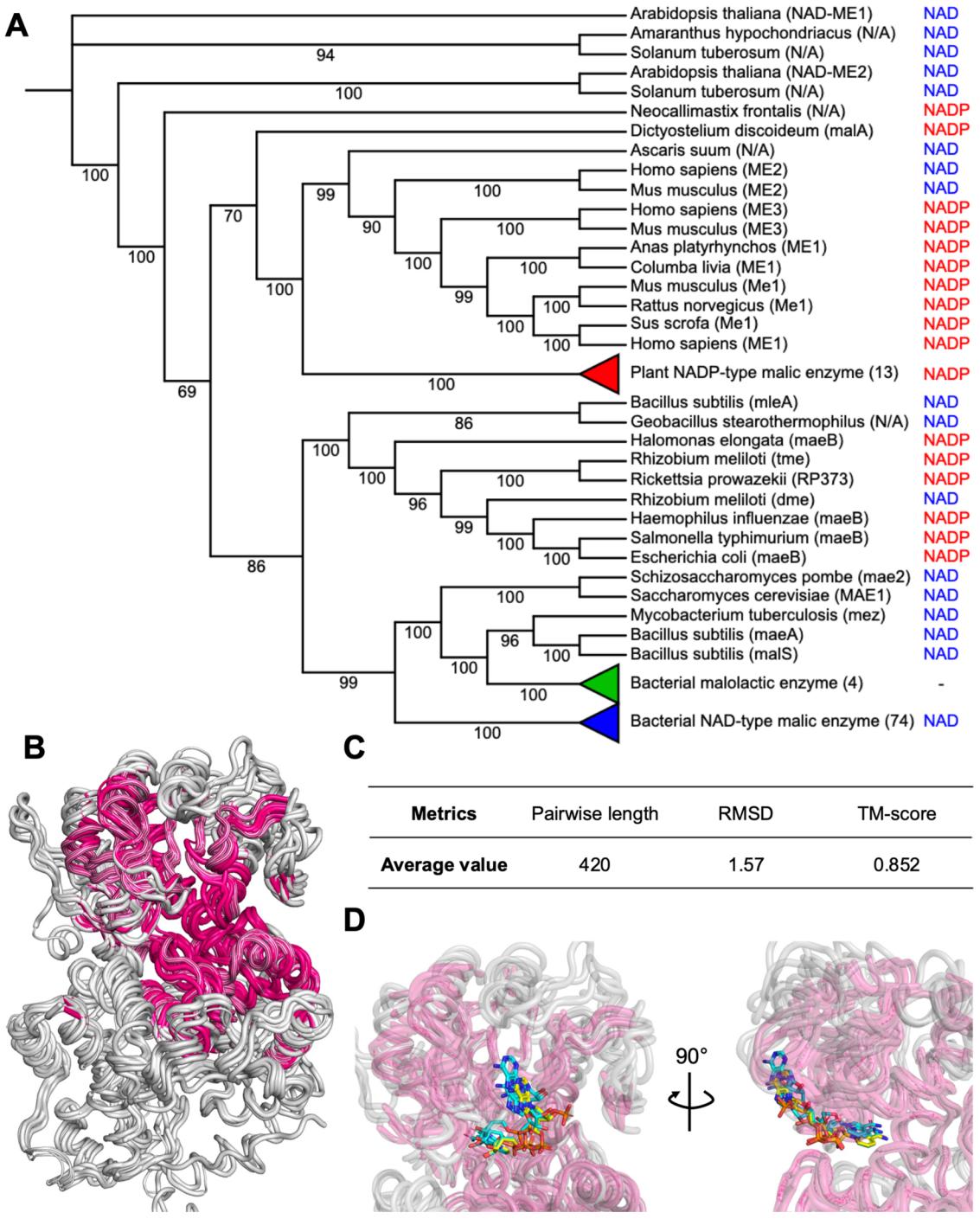
INTRODUCTION

Since the origin of life, proteins with diverse structures and functions have evolved through repeated cycles of mutations and selection. Although natural proteins have been used from basic science to industry, proteins are not expected to be utilized outside the host organism or its living environment. Thus, it is often necessary to adjust their functions to the desired environmental conditions by improving their thermal stability, catalytic activity, and substrate/cofactor specificity conversion. The desired function of a protein can be replicated by artificially accumulating the mutations that affect its function in a fashion similar to that occurring in nature over long periods of time.[1] However, introducing mutations into a protein is affected by previous mutations; therefore, each mutation makes a nonlinear contribution to the function.[2−5] Despite the mutations being necessary, they are often disadvantageous for function and thermostability and are likely to drop out of leading candidates through the artificial evolution process. Moreover, some functions are produced by the combined accumulation of unfavorable mutations in what is known as sign epistasis, which makes the artificial evolution of proteins more difficult.[6]

Artificial designing with the aim of altering substrate and cofactor specificity is particularly challenging. Altering substrate and cofactor specificity generally requires the introduction of multiple mutations,[7−9] and distant (>10 Å) mutations can markedly affect catalytic function.[10−13] More-

Received: June 13, 2022
Published: November 2, 2022



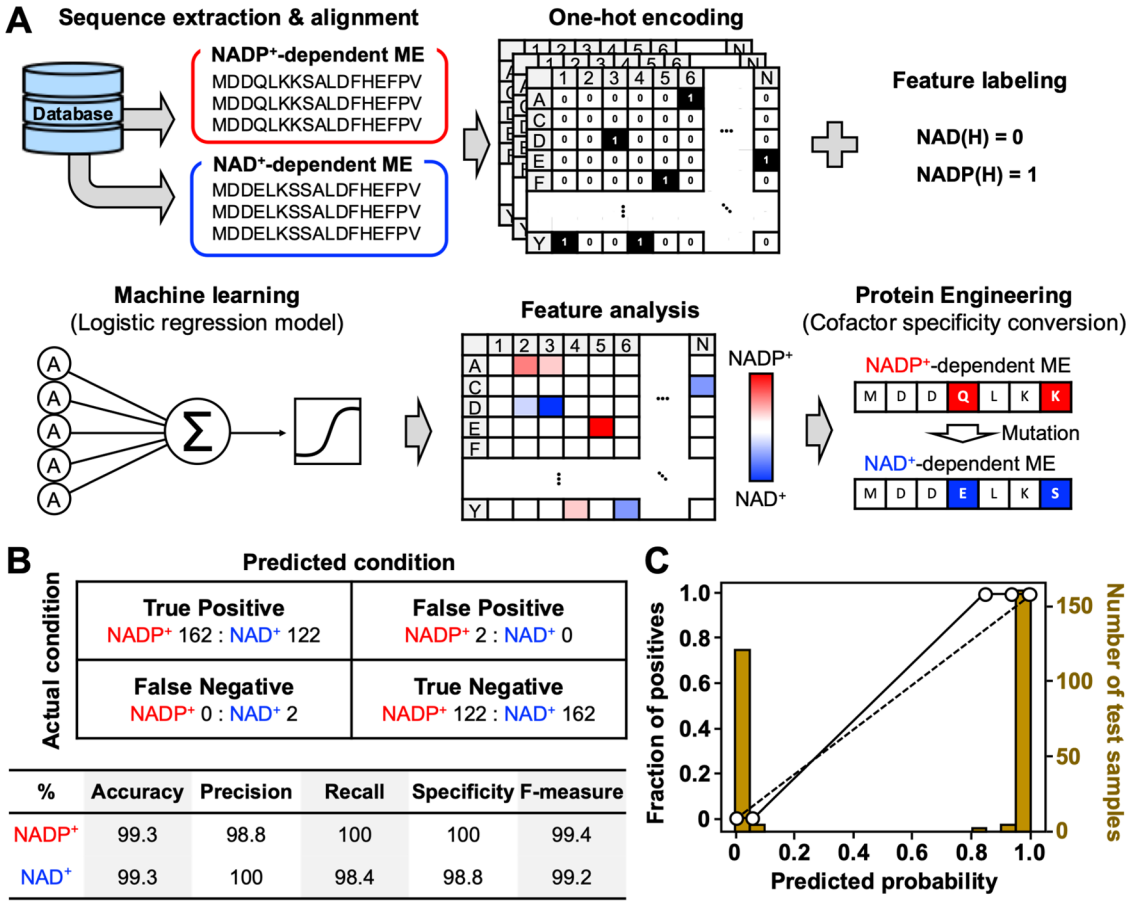


Metrics	Pairwise length	RMSD	TM-score
Average value	420	1.57	0.852

Figure 1. (A) Molecular phylogeny of ME. The leaf nodes are labeled with the species name, gene name in parentheses, and their preference for redox cofactors. Those that have not yet been assigned a gene name are N/A. The reduced leaf nodes are labeled $NADP^+$-dependent plant ME (red), bacterial malolactic enzyme (green), and NAD^+-dependent bacterial ME (blue). Numbers in parentheses in the reduced leaf denote the number of sequences in each clade. Branches are labeled with bootstrap probability. For unreduced phylogeny, see Figure S1. (B) Superimposed images of seven identified MEs (PDB IDs: 1DO8, 1GQ2, 1LLQ, 3WJA, 5CEE, 5OU5, and 6ZN4). Regions of high structural conservation where there are no gaps and the RMSD of the α carbons are within 4 Å are colored magenta. (C) Table of average metrics values from the superimposed structure of B. All metrics values for each structure are in the Supporting Information and Tables S1 and S2. (D) Superimposed images of ME cocrystal structures with NADH or NADPH: PDB IDs 1DO8, 1GQ2, 1LLQ, and 5CEE. $NADP^+$ or NAD^+ are indicated by sticks.

over, we can realistically screen samples of only a few thousand mutants in size, which is a small fraction of the total search space required for activity-based mutant selection. Although repeatedly utilizing saturation mutagenesis in directed evolution improved enzyme activity and stability,[14−16] high-throughput screening must be established to obtain the desired mutants. Computational enzyme design can overcome the limitations of experimentally oriented evolution methods since the evolutionary trajectories do not constrain the design process. The retro-aldol reaction was generated with non-natural substrates by embedding transition states generated by quantum chemical calculations in scaffold structures, including

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985



%	Accuracy	Precision	Recall	Specificity	F-measure
NADP+	99.3	98.8	100	100	99.4
NAD+	99.3	100	98.4	98.8	99.2

Figure 2. (A) Key steps illustrating the ML-based enzyme design for redox cofactor specificity conversion. ME containing both NADP+- and NAD+-dependents were collected and classified by cofactor specificities. Each ME was then converted to a one-hot encoding and labeled by the cofactors. ML was conducted using the one-hot encoding vectors made from MEs. The coefficients in the logistic function are the contribution to cofactor recognition for each residue and were described as a heat map. The protein engineering step was implemented on the ME from *E. coli* based on partial regression coefficients. (B) Confusion matrix and performance indexes. To evaluate the classification performance, 30% of the dataset was used. (C) Calibration plot for estimating prediction accuracy. The *x*-axis represents the average predicted probability from the learned model. The left and right *y*-axes indicate the ratio of positives and the sample number from the labeled data, respectively. A value of one and zero indicates an NADP+-dependent- and NAD+-dependent enzyme, respectively.

jelly-roll and TIM barrel.[17] Subsequently, the development of designer enzymes catalyzing Kemp elimination[18] and the Diels−Alder reaction[19] shows that a variety of enzymes can be designed by exploiting the high plasticity of the protein scaffold. Although the computational enzyme designs that convert these substrate specificities are useful and highly successful, their use is limited because they require a precise intermediate reaction structure of the substrate with residues and result in low enzymatic activity.

Recently, machine learning (ML) was utilized to search an enormous space and increase the possibility of selecting desired variants to support directed evolution procedures.[20−25] ML predicts the correspondence between sequences and functions from the data independent of the physical model and potential function. An early evolutionary engineering campaign with a ML linear model reported obtaining the halohydrin dehalogenase mutant that could increase volumetric productivity by approximately 4000-fold.[22] Furthermore, Gaussian process regression used to handle uncertainty in experimental data has been used to improve the thermal stability of cytochrome P450s,[23] the optical properties of rhodopsin,[24] and the modification of fluorescent proteins.[25] These models were prepared by analyzing protein function data from initially mutated residues for the following selection rounds. However, even if it could construct a high-throughput assay system to assess enzyme function, the experimentally predicted landscape would be systematically biased against the dynamics of adaptation because the experimental data would only cover a tiny fraction of the actual landscape.[26] In addition, the amino acid residues that contribute to substrate specificity are known to be distributed throughout the structure, not just around the substrate pocket,[27] making the identification of function-determining amino acid residues even more challenging. Therefore, identifying the location and number of those impactful residues are key to both conventional and ML-assisted protein evolution.

Herein, we propose a methodology for identifying amino acid residues involved in cofactor specificity by combining a logistic regression model with an amino acid sequence dataset having the same fold structure but different cofactor specificity. We hypothesized that conserved residues between structurally homologous enzymes possessing different substrate/cofactor specificities are interchangeable and can potentially alter their substrate/cofactor specificities. Utilizing the logistic regression model would allow preparing a ranking list of amino acids that correspond to enzymes with complex features but without an

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

elucidated crystal structure and specifying the mutational positions, which would allow for preferential mutation design and efficiently limit the search space. This study tested our hypothesis by changing the cofactor specificity of the *Escherichia coli* malic enzyme (ME) from the $NADP^+$-dependent- to the NAD^+-dependent form. ME is an enzyme that decarboxylates malate, an intermediate in the tricarboxylic acid cycle, to pyruvate, the final product of glycolysis, accompanied by NAD^+ or $NADP^+$. Most organisms, from prokaryotes to eukaryotes, harbor ME; thus, an abundance of sequence data is available for both NAD^+- and $NADP^+$-dependent MEs. By obtaining ME amino acid sequences from various species, we aimed to obtain the residues responsible for cofactor specificity.

Protein engineering to switch cofactor specificity was primarily developed in metabolic engineering research.[28−32] However, many redesigned enzymes are only promiscuous modifications, and a basic understanding of cofactor specificity switching is inadequate. In brief, this study used a logistic regression model to identify the amino acids behind NAD^+- and $NADP^+$-dependence for each position. Replacing the amino acids in the order of the most significant differences in features resulted in cofactor specificity switching. Interestingly, introducing mutations in tens of units proposed by the developed ML model did not produce any fatal negative effects on the structure. The results highlight the potential of combining ML with phylogenetic analysis for enzymatic design with the aim of altering cofactor specificity.

RESULTS

Structural Conservation of ME. Molecular phylogenetic analysis was initially performed to determine the similarity of amino acid sequences of NAD^+- and $NADP^+$-dependent MEs in terms of sequence homology. The amino acid sequences of 123 MEs from various species were obtained by performing a Basic Local Alignment Search Tool (BLAST) search using the UniProtKB/Swiss-Prot database annotated with various properties of three NAD^+- and $NADP^+$-dependent MEs as queries, respectively. A phylogenetic tree was constructed using the maximum likelihood method after multiple sequence alignments and trimming. Phylogenetic analysis revealed that MEs are widely distributed in animal, plant, and bacterial phyla and that NAD^+- and $NADP^+$-dependent MEs coexist within the phylum (Figure 1A). This indicates that it is possible to extract amino acid sequences of both $NADP^+$- and NAD^+-dependent MEs in a species-independent manner without introducing bias. Most plants have $NADP^+$-dependent MEs, while bacteria tend to have NAD^+-dependent MEs, and some bacteria have malolactic enzymes that do not utilize coenzymes.

The structural similarity between cofactor and species difference of MEs was performed using pairwise structural alignment of known structures: three NAD^+-dependent MEs from human mitochondria, phytoplasmas, and nematodes and four $NADP^+$-dependent MEs from human, pigeon, maize, and *Bdellovibrio* bacteria were obtained from the Protein Data Bank (PDB), and their structures were superimposed (Figure 1B). The magenta-colored regions are amino acid residues that do not contain gaps and have a maximum root-mean-square deviation (RMSD) of 4 Å or less between the paired carbon atoms. Each ME consists of 380−630 amino acids, with an average pairwise length of 420 residues. The average RMSD and TM-scores[33] calculated from the pairwise amino acid residues were 1.57 and 0.852, respectively (Figure 1C). The TM-score has a range of 0−1; since it was significantly greater than 0.5, this indicated that NADH or NADPH has the same fold structure. Magenta represents the cofactor binding site in the conserved region of the four ME crystal structures containing NAD(H) or NADP(H) (Figure 1D). These results indicate that the ME sequence structures of various species are highly conserved, especially in the vicinity of the active site. The amino acid sequences of MEs from various organisms in the database are expected to have homologous folds to each other. The results led us to believe that functional conversion by sequence exchange was achieved if we could extract features of the NAD^+ and $NADP^+$-dependents.

Building a Machine Learning Model. The scheme leading from machine learning to cofactor specificity conversion by protein engineering is shown in Figure 2A. We randomly collected sequences of both $NADP^+$- and NAD^+-dependents of MEs from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to investigate mutation positions and candidate amino acids for cofactor preference conversion. The KEGG database encompasses information from various databases such as Swiss-Prot, TrEMBL, RefSeq, and PDB; as a result, we thought that it is possible to collect broad ME amino acid sequences from diverse species. One thousand ME amino acid sequences were collected, and 952 (448 NAD^+-dependent and 504 $NADP^+$-dependent) unique sequences were obtained by deleting duplicates. The amino acid sequences in the dataset ranged between 289 and 1042 residues in length (Figure S2). The unique sequences were aligned using Clustal Omega to detect point mutations, insertions, and deletions. It is also advantageous for ML if the sequence lengths are identical, including the gaps. Finally, 286 sequence data (122 NAD^+-dependent and 164 $NADP^+$-dependent), or 30% of the dataset, were used as the test set for model validation and the remainder as the training set.

The dataset was trained on a logistic regression model. The collected ME sequences were expressed as a one-hot vector in $M \times N$ (eq 1), and $NADP^+$- and NAD^+-dependents were set to one and zero, respectively, to treat them as teacher labels. M is the type of amino acid (20 types), and N is the length of the amino acid residues, including gaps.

$$x = \begin{pmatrix} x_{1,1} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{M,1} & \cdots & x_{M,N} \end{pmatrix} \in \{0, 1\}^{MN} \tag{1}$$

The M × N dimensional one-hot vectors from the amino acid sequences were then transformed into a linear polynomial with an intercept parameter β_0 and coefficient parameters $\beta_{i,j}$ (i = 1, 2, 3, ⋯, M and j = 1, 2, 3, ⋯, N) (eq 2). The equation was substituted into the logistic function to express the features that determine $NADP^+$- and NAD^+-dependents from one to zero (eq 3). Values closer to zero represent the NAD^+-dependent form, while values closer to one represent the $NADP^+$-dependent sequence.

$$f(x) = \beta_0 + \beta_{1,1}x_{1,1} + \beta_{1,2}x_{1,2} + \cdots + \beta_{M,N}x_{M,N} \tag{2}$$

$$\Phi(x) = \frac{1}{1 + e^{-f(x)}} \tag{3}$$

A logistic regression model based on the training set was used to discriminate the 286 MEs in the test set. Figure 2B presents the confusion matrix and its performance indexes.

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

Table 1. Kinetic Parameters of MaeB, trcMaeB, and trcMaeB Variants[a]

	K_M (μM)		k_{cat} (s^{-1})		k_{cat}/K_M (s^{-1} mM^{-1})	
	NAD^+	$NADP^+$	NAD^+	$NADP^+$	NAD^+	$NADP^+$
MaeB (full length)	n.d.	62 ± 8	n.d.	37.7 ± 0.50	n.d.	608
trcMaeB	n.d.	173 ± 28	n.d.	7.35 ± 0.08	n.d.	42.5
trcMaeB10	210 ± 50	100 ± 30	0.39 ± 0.02	0.20 ± 0.02	1.9	2.0
trcMaeB20	77 ± 12	110 ± 30	0.55 ± 0.01	0.14 ± 0.01	7.1	1.2
trcMaeB30	78 ± 7	n.d.	0.66 ± 0.01	n.d.	8.4	n.d.
trcMaeB40	80 ± 14	n.d.	0.47 ± 0.01	n.d.	5.9	n.d.
trcMaeB50	76 ± 2	n.d.	0.73 ± 0.00	n.d.	9.6	n.d.
trcMaeB60	190 ± 30	n.d.	0.43 ± 0.01	n.d.	2.3	n.d.
trcMaeB70	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
trcMaeB K237Q	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.

[a]"n.d." not determined due to the inability to fit.

The training model mistakenly recognized only two $NADP^+$-dependent MEs as NAD^+-dependent MEs but achieved the classification of cofactor specificity of MEs with 99.3% accuracy. Moreover, as a classification index, *F*-measures were given from precision and recall for $NADP^+$- and NAD^+-dependent MEs and were 99.4 and 99.2%, respectively, confirming that the classifier was able to discern cofactor specificity from the differences in amino acid sequence. Figure 2C demonstrates the calibration plot used to evaluate the logistic regression model. It evaluates how well the model's predicted probabilities match the actual ratios. The bar graph is a histogram where the horizontal *x*-axis is the probability of $NADP^+$-dependent ME being predicted by the model divided into multiple intervals, and the vertical right *y*-axis is the number of test samples. The calibration plot contains the same horizontal axis as that of the histogram. The vertical axis (left *y*-axis) is the percentage of actual $NADP^+$-dependent ME samples in each histogram interval. When the score and the actual percentage match, the calibration plot fits on the 45° line (dashed line); therefore, the closer the calibration plot is to the 45° line, the better the model's prediction performance. In this ML model, the prediction did not significantly deviate from the 45° line in the calibration plot. Moreover, *k*-fold cross-validation was used to evaluate the generalization performance of the test and training sets for the logistic regression and other ML models, including linear regression, group lasso, and decision tree models (Table S3). The logistic regression model had the highest score, and its prediction rate was 99.7%. The cross-validation results proved that overfitting has not occurred during model-building. Although the decision tree model also achieved high accuracy (98.8%), it is not easy to focus on a single residue due to its properties. Therefore, we adopted the logistic regression model for the following enzyme design since it can evaluate the importance of each residue and position contributing to cofactor specificity with higher resolution.

Contribution of ME Amino Acid Residues to Cofactor Specificity. The partial regression coefficient (β) of the logistic regression model after training is a weight parameter for the cofactor specificity of each amino acid residue and is synonymous with the contribution of each amino acid residue to cofactor specificity. In this study, the cofactor specificity of ME from *E. coli* strain K12 (MaeB) was converted from $NADP^+$- to NAD^+-dependent. This is a suitable model to explain the concept of this study because the three-dimensional structure of MaeB is not elucidated. The PTA domain from 427 to 759 consists of a phosphate acetyltransferase without catalytic activity and is a suggested acetyl-CoA sensor for allosteric regulation.[34,35] MEs of many species do not have a PTA domain; therefore, only the ME domain was used for ML in this study. Henceforth, the ME domain of MaeB from *E. coli* is referred to as trcMaeB. trcMaeB did not recognize NAD^+ as well as full-length MaeB but expressed catalytic function in the presence of $NADP^+$ (Table 1), which is consistent with a previous study.[36]

A heat map of the contribution of each trcMaeB amino acid residue to cofactor specificity based on ML-optimized partial regression coefficients is shown in Figure 3. The top line shows the predicted secondary structure information of trcMaeB based on AlphaFold2[37] and RoseTTAFold[38] (Figure S3), the middle line shows the contribution of each amino acid residue of trcMaeB to cofactor specificity, and the bottom lines show the contribution of each of the 20 amino acid side chains toward cofactor specificity. Larger partial regression coefficients $\beta_{i,j}$ obtained from this model contribute to the $NADP^+$-dependent form, while smaller coefficients contribute to $NADP^+$-type residues. Equation 4 was set up to identify the amino acid residues with the highest contribution to cofactor specificity.

$$\text{score}(s_j) = |\max(\beta_j)| + |\min(\beta_j)| \quad (4)$$

$\text{Score}(s_j)$ is the contribution of cofactor specificity, and $|\max(\beta_j)|$ and $|\min(\beta_j)|$ are the absolute values of the maximum and minimum partial regression coefficients at the residue position, respectively. Since $\text{score}(s_j)$ values represent the strength of the effect on cofactor specificity, a mutation ranking was created according to this value, and this information was used to introduce mutations into trcMaeB.

Assessing Catalytic Properties of Designed ME Variants. The ML model introduced up to 100 mutations into trcMaeB (ten at a time) to create trcMaeB mutants (trcMaeB10−100; Figure S4). If the natural trcMaeB already had amino acid residues characteristic of $NADP^+$, the original residues were retained. The ML-designed trcMaeB mutants were prepared using an *E. coli* expression system. Purified, soluble mutants were obtained for trcMaeB10−70, which introduced mutations ranked 10−70 (Figure 4A), while trcMaeB80−100 mutants could not be purified. The introduction of mutations accumulated load to the structure since amino acid sequence conservation decreases with lower ranks.

Conversion of malate to pyruvate in the presence of $NADP^+$ or NAD^+ was performed on trcMaeB10−70 expressed in the

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

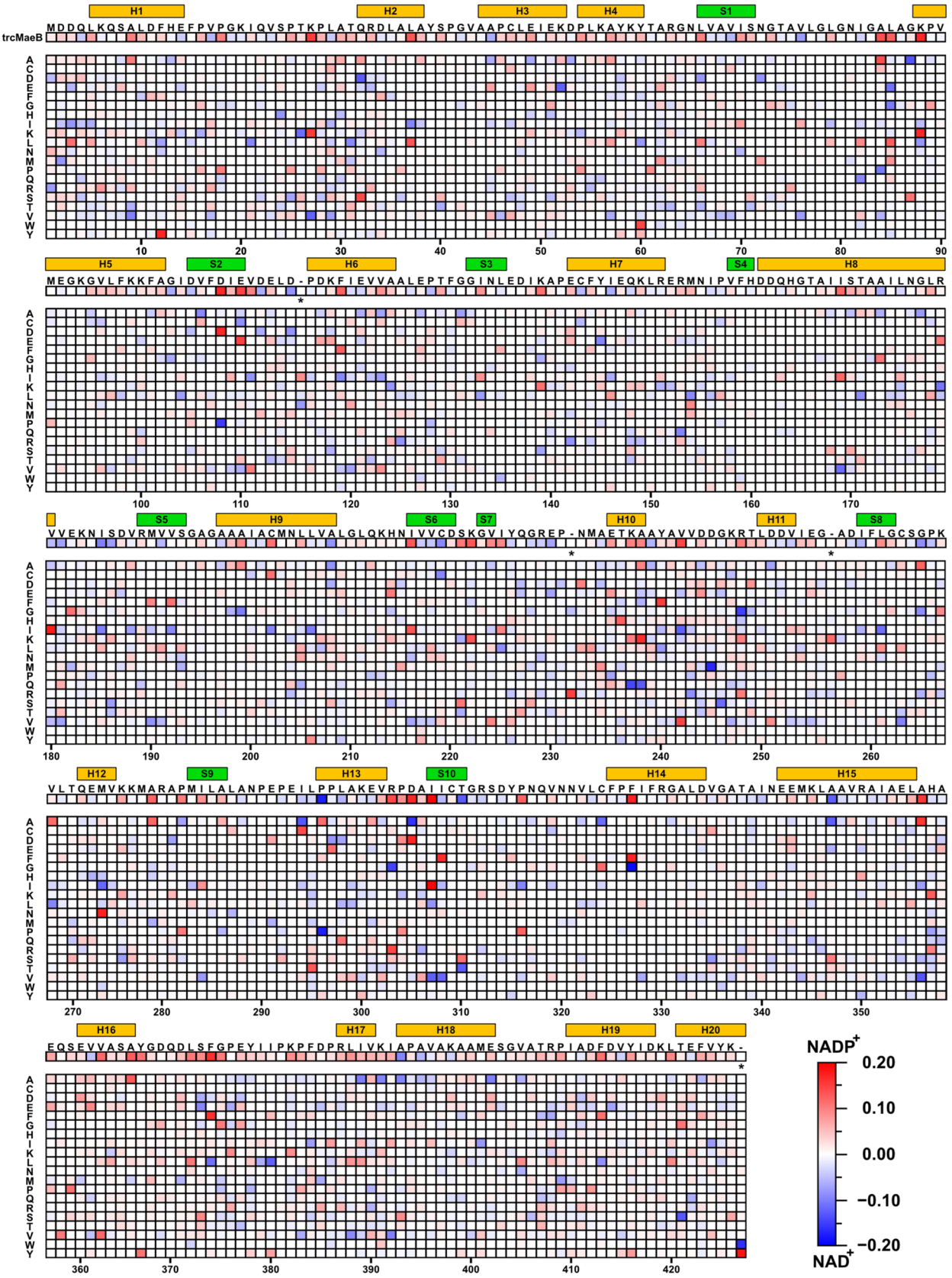
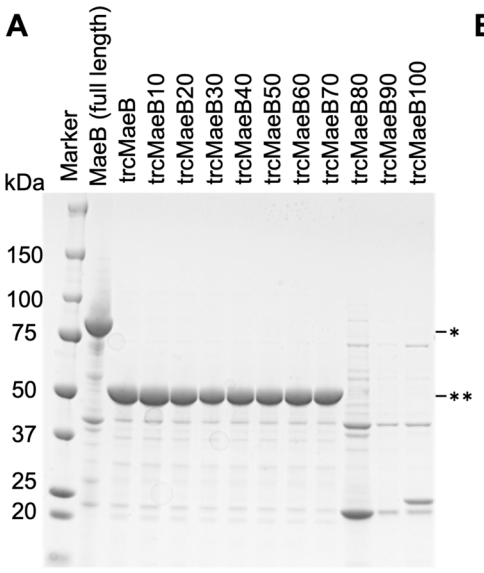


Figure 3. Aligned heat map showing properties of amino acid residues with relationship scores for redox cofactors. Red and blue indicate a high correlation for $NADP^+$ and NAD^+, respectively. The contribution of each amino acid residue comprising trcMaeB to cofactor specificity is also shown as a heat map. Hyphenated columns in the trcMaeB sequence indicate newly inserted amino acids. In addition, the top line of the heat map shows the secondary structure information of trcMaeB predicted by AlphaFold2 and RoseTTAFold. The letters H and S represent the helix and sheet structure, respectively. For the predicted three-dimensional structures, see Figure S3.

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985



B

enzyme	Yield (mg/L)	% fold change (design/wild type)
trcMaeB	128.2	100.0
trcMaeB10	55.2	43.1
trcMaeB20	36.9	28.8
trcMaeB30	93.5	73.0
trcMaeB40	69.4	54.2
trcMaeB50	27.7	21.6
trcMaeB60	34.5	26.9
trcMaeB70	38.5	30.0
trcMaeB80		
trcMaeB90		
trcMaeB100		

Figure 4. SDS-PAGE image (A) and yields (B) of MaeB, trcMaeB, and trcMaeB variants after purification by metal affinity chromatography. In panel (A), one asterisk indicates band positions of MaeB (full length), and two asterisks indicate band positions of trcMaeB and its mutants. In panel (B), the variants that could not be purified from soluble fractions are indicated by sloping lines.

soluble fraction to investigate cofactor selectivity (Table 1 and Figure S5). Wild-type trcMaeB was completely inactive against NAD^+, whereas trcMaeB10 expressed catalytic activity for NAD^+. Moreover, the affinity and turnover number for $NADP^+$ greatly decreased, the K_M value increased 2.8-fold, and k_{cat} value decreased to 1/5, resulting in a promiscuous state with activity against both NAD^+ and $NADP^+$. trcMaeB20, which added more than 10 mutations to trcMaeB10, showed increased and decreased affinity and turnover number for NAD^+ and $NADP^+$, respectively. Further mutation accumulation, according to the ranking, indicated a shift to a more NAD^+-dependent state. trcMaeB30 lost the affinity for $NADP^+$ and gained NAD^+ specificity. Our ML model found that 20−30 mutations were sufficient to switch the cofactor specificity of trcMaeB, and the expression level was not greatly affected. trcMaeB50 showed the best k_{cat}/K_M value for NAD^+. trcMae60 had decreased enzyme function, and trcMaeB70 showed no recognition of either cofactor, which indicated that mutations with lower ML rankings have weaker contributions to NAD^+ function and may harm catalytic activity. Substituting dozens of amino acids in a protein and converting substrate and cofactor specificity from sequence information alone is beyond the reach of conventional protein engineering, which relies on random mutations or point mutations based on the curation. The ML model in this work may narrow down the mutation positions and amino acid candidates that cannot be found based on structural information alone.

Molecular Modeling Simulations for Structural Comparison and Visualization of Mutation Sites. Comparative models were generated using Rosetta to identify the mutation position and effect of mutagenesis on structure (Figure 5A). Mutations were scattered throughout the ranking-independent structure, and mutated residue clusters were observed at various locations. These mutations and clusters were suggested to induce structural distortions; however, no significant decrease in expression levels and no large increase in energy scores were observed in homology models (Figures 4B and S6). These results suggested that the ML model effectively selected pairs of covariant residues, preventing stability loss and inactivation. In particular, we focused on trcMaeB10−30 with switched cofactor specificity and investigated the mutated residues around the substrate pocket (Figure 5B). Except for A238Q and K237Q, we observed a tendency for hydrophobic amino acids to appear around the substrate pocket. In trcMaeB10, R302G and D304A mutations were observed. Since these mutations were in the form of hydrophilic amino acids with large side chains that were converted to hydrophobic small amino acids, they may influence the structure and/or dynamics of trcMaeB. The observed substitutions around the pocket imply a slight change in internal structure and dynamics, which affects cofactor selectivity.

On the other hand, mutations in residues away from the substrate pocket were also observed, which may affect the structure of paired ME domains. The MaeB structure of *B. bacteriovorus* HD100, which has a 61.8% homology with the MaeB of *E. coli*, demonstrates that the N-terminal crossover ME region of this enzyme is composed of an α-helix and a random coil and is essential for dimeric ME domain formation (Figure S7). Thus, this indicates that the residues far from the substrate pocket may affect the multimer formation and induce slight conformational changes.

Furthermore, the MEs from *Aster yellows witches'-broom phytoplasma* (AYWB) have a 47.9% homology to trcMaeB. The amino acid residues in the known substrate pocket from the crystal structure of $NADP^+$-dependent ME from *B. bacteriovorus* HD100 and NAD^+-dependent ME from AYWB were compared. The phosphate group of $NADP^+$ interacts with lysine at position 250 of ME from *B. bacteriovorus* HD100, while glutamine is present at the same position (position 231) in NAD^+-dependent ME from AYWB (Figure 5C,D). The trcMaeB mutant lysine at position 237 was mutated to glutamine and was ranked 21st (Figure 5E,F). No further mutations to the substrate pocket were identified up to the 70th ranking of accumulating mutations with an observation of our homology models. These results suggest that the amino acid residue at position 237 strongly influences cofactor specificity. However, it is not apparent whether a single mutation at K237Q would be able to affect cofactor specificity

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

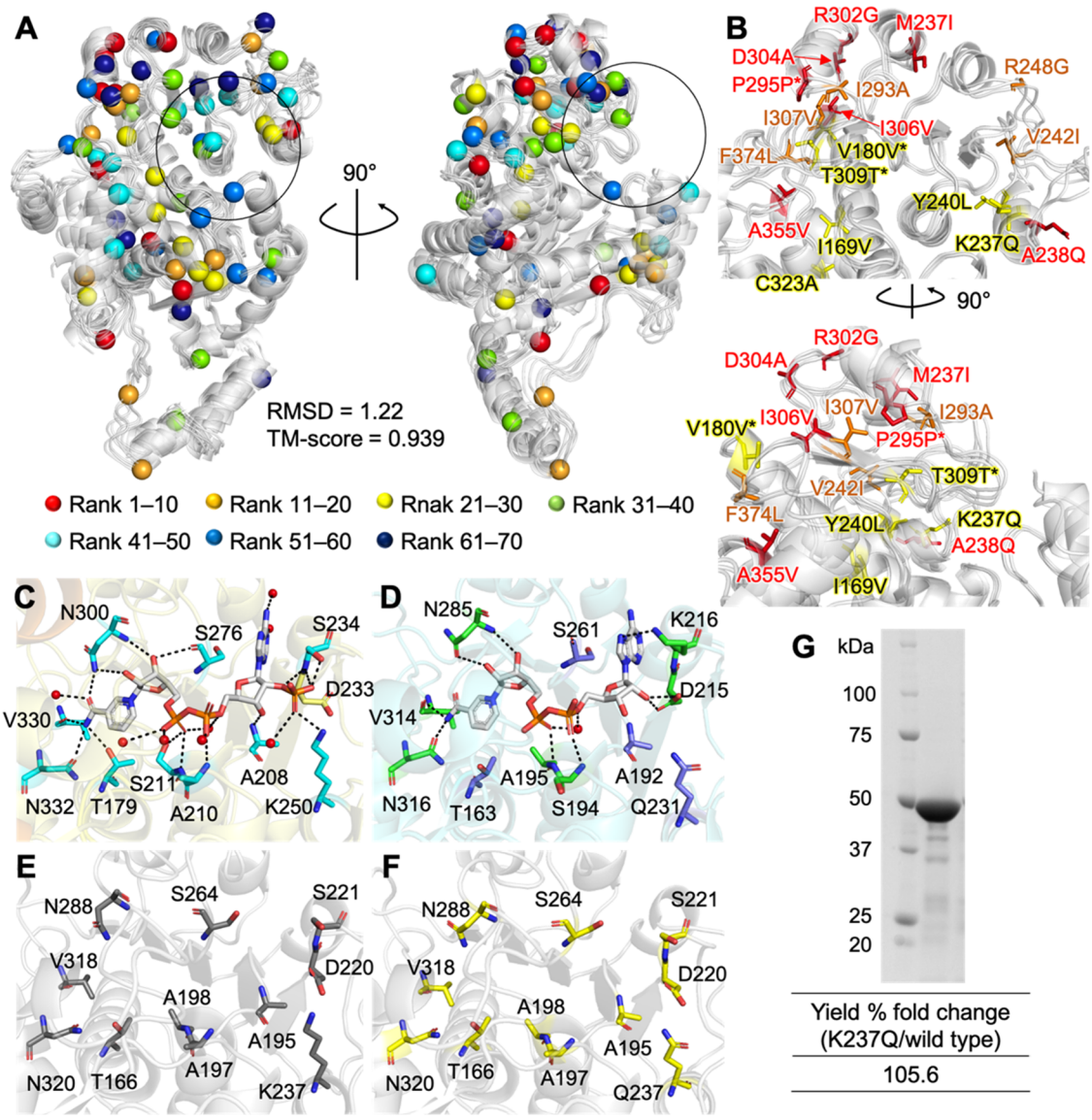


Figure 5. (A) Superimposed images of the trcMaeB10−70 homology model created with the RosettaCM protocol. The colored spheres indicate the mutation sites based on the contribution ranking for cofactor specificity provided by ML. Black circles indicate the area around the substrate pocket. (B) Superimposed images of regions around the substrate pockets of the trcMaeB10−30 homology models. Red, orange, and yellow sticks indicate residues in Ranks 1−10, 11−20, and 21−30, respectively. Amino acid residues marked with an asterisk indicate that the mutations have not occurred because trcMaeB already has residues. (C) Close-up look into the NADPH-binding site of ME from *Bdellovibrio bacteriovorus* HD100 (PDB ID: 6ZN7). (D) Close-up look into the NADH-binding site of ME from AYWB (PDB ID: 5CEE). (E) Predicted cofactor binding site of trcMaeB from the homology model. (F) Predicted cofactor binding site of trcMaeB30 from the homology model. (G) SDS-PAGE image and yields of the trcMaeB K273Q variant after purification by immobilized metal-ion affinity chromatography.

switching. Thus, we tested the need for substituting the top-ranking residues higher than K237Q on 21st in the ranking by creating trcMaeB K237Q and investigating its catalytic function.

trcMaeB K237Q was prepared by trcMaeB point mutation, expressed in *E. coli* strain AG1, and purified by immobilized metal-ion affinity chromatography as with previous trcMaeB proteins to evaluate the effect on cofactor selectivity. trcMaeB K237Q was expressed at the same level as trcMaeB (Figure 5G) but did not react with NAD^+ (Table 1). Furthermore, the original activity against $NADP^+$ was lost. These results indicate that the mutation of only the substrate-interacting residues in trcMaeB is insufficient for cofactor specificity and suggests that small structural changes around the substrate pocket are essential. The substituted candidate residues proposed by the logistic regression model indirectly interact with the cofactor, resulting in a substrate pocket that influences structure and dynamics and switches cofactor specificity.

DISCUSSION

We proposed a highly readable ML model with high prediction accuracy by restricting the input data to enzymes belonging to the same structural family and demonstrated its application to protein engineering. In recent years, protein science has made many attempts to conduct deep learning to grasp the complex features of proteins using a large amount of miscellaneous data. For example, predictions of protein structures,[37,38] EC numbers,[39] interaction sites with other biomolecules,[40] and unstable regions in a protein core[41] were performed to statistically represent this global knowledge of proteins. ML-

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

based protein research is now in its infancy, with the successful identification of protein features. The more one tries to capture complex features using various data, the more complex the model becomes and the more difficult inference becomes; however, prediction accuracy may improve. Inference difficulty is a big issue when results from complex models are applied to protein engineering. Furthermore, the quantity of data is critically important to capture features from many explanatory variables; however, it is not easy to prepare a large amount of high-fidelity experimental data that withstand deep learning specifications.

The emphasis of our study is that protein function can be predicted with high accuracy and readability using classified input data and a simple multiple regression model. Although this method requires the creation of an optimized model for each protein, it allows us to classify and engineer enzymes with a high degree of accuracy. Therefore, this method can help estimate the residues that control individual protein function and modify the substrate and cofactor specificity without changing the framework. We have classified the cofactor specificities of MaeB and predicted the degree of conservation for each position in the amino acid sequence. Our model's cofactor specificity discrimination accuracy was more than 99.3% in MEs (Figure 2B and Table S3), and the introduction of mutations up to the 30th ranking completely switched the cofactor specificity from $NADP^+$-dependent to NAD^+-dependent (Table 1). Although ML was used to predict substrate specificity in the past, our proposed cofactor specificity modification focusing on individual amino acid residues will further expand the use of ML. Moreover, a glimpse of the evolutionary process in nature was observed by converting cofactor specificity. The evolution of enzymes in nature is thought to occur via a promiscuous state in which a secondary activity distinct from the original primary activity is expressed, and at least two substrates are recognized.[10,42] Our results also followed a pathway from a promiscuous state recognizing NAD^+ and $NADP^+$ to a completely NAD^+-dependent state due to the accumulation of mutant residues (Table 1). Since multiple mutagenesis is necessary to alter substrate and cofactor specificity, we believe that ML can be an effective tool in identifying mutation positions and suggesting candidate amino acids.

It is noteworthy that the trcMaeB70 variant with a mutation ratio of 14.5% (up to the 70th rank) was expressed in the soluble fraction (Figure 4A). The difference in free energy between the stable and unfolded states of natural proteins is approximately 5−10 kcal/mol, and many proteins unfold with only a slight loss of hydrogen bonds.[43] Our ML-based protein engineering strategy is based on multiple sequence alignment similar to the consensus sequence design method.[44,45] In this consensus sequence design method, nonconserved amino acids in the protein are replaced with highly conserved residues extracted from the homologous protein sequence to improve protein stability while preserving biological activity. There is also a structural stabilization method called PROSS, which expands the consensus sequence design method combined with rational design.[46,47] Meanwhile, our logistic regression method aims to switch substrate and cofactor selectivity and has the advantage of clarifying the order of functional expression. The highly mutated variants generated by ML design may successfully select consensus sequences since they were well expressed following their change in function. Even in the comparative model generated by Rosetta, the RMSD values averaged 1.22 Å, and the energy values did not significantly increase up to trcMaeB70 (Figure S6). These results indicate that ML can extract amino acid residues contributing to structure and function expression from homologous sequences in a ranking format. However, this method is not a panacea in protein engineering. Similar to the challenges of conventional consensus design, enzyme activity cannot be controlled. The k_{cat}/K_M of the designed trcMaeB was reduced to approximately 1/5 of that of the original version. Improvements in enzyme activity might require traditional directed evolution techniques or recent ML methods to assist in directed evolution. Although we only used amino acid sequences with cofactor specificity labels as inputs, the addition of more parameters, including catalytic activity and denaturation midpoints, for each enzyme, would allow us to weigh each enzyme with regard to those values. This would ultimately facilitate building an influential ML model that can, for example, alter substrate specificity while increasing enzyme activity.

In synthetic biology and metabolic engineering, altering the cofactor preference of metabolic enzymes has been recognized as one of the key strategies for controlling metabolic pathways and maximizing target substance production since 1990.[29] The only structural difference is that NAD(H) and NADP(H) are phosphorylated at the 2′ position of the ribose moiety, well away from the hydride acceptor atom of the nicotinamide ring. However, redox enzymes in nature strongly prefer one of the derivatives and significantly reduce the activity of the other. Studies utilizing enzymes with different redox cofactor selectivity from other organisms were reported; however, protein engineering is often needed because heterologous enzymes may not be expressed or have good catalytic activities. There are reports of narrowing down candidate mutant residues from crystal structures of enzymes with bound cofactors[28−31] and proliferative screening that alters the cellular metabolic state.[32] However, a high-throughput assay that can screen many metabolic enzymes has not been constructed. In this study, we redesigned the cofactor specificity of MaeB from $NADP^+$ to NAD^+ without structure information and screening steps. Therefore, it represents a versatile methodology with potential for extension into the fields of synthetic biology and metabolic engineering, where the redox balance of cells requires control.

In summary, we used the logistic regression model derived from datasets with the same structure but different functions to switch enzyme functions with unknown steric structures without causing fatal destabilization. While directed evolution is generally limited to a few mutations in a protein, our model allows the introduction of dozens of mutations as if guiding a path through a vast search space of the protein. Since this method is a statistical process using a large amount of sequence information, it may be possible to search for hot spot residues that cannot be detected from structural information alone. The accumulation of mutations in additive order of contribution did not disrupt the protein structure. The ability to reliably change the specificity of a cofactor even from sequence data alone is a strength of the logistic regression model.

MATERIALS AND METHODS

Phylogenetic Analysis. BLAST homology search was performed using UniProtKB/Swiss-Prot[48] as a database and utilized three NAD^+-dependent MEs (from *Arabidopsis thaliana*'s mitochondria, *Rhizobium meliloti*, and *Solanum tuberosum*) and three $NADP^+$-dependent MEs (from *E. coli*,

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

Haemophilus influenzae, and *Flaveria pringlei*) as query sequences. The *E*-value was set to $<10^{-7}$. The amino acid sequences of the obtained IDs were extracted from UniProtKB, and those between 280 and 350 amino acid residues were used for subsequent operations. Multiple sequence alignments were created using MAFFT.[49] After removing poorly aligned regions, a phylogenetic tree was estimated by the maximum likelihood method using IQTree v2.[50] The LG+R6 model was selected for phylogenetic tree prediction based on the Bayesian information criterion score using ModelFinder.[51] The reliability of the estimated clade was evaluated by the bootstrap method with UFBoot2[52] using 1500 bootstrap iterations.

Machine Learning. One thousand amino acid sequences of NAD- and NADP-forms of ME were randomly obtained using the KEGG database.[53] The amino acid sequences of the dataset ranged between 289 and 1042 residues in length. Duplicate samples were removed, and multiple sequence alignment was performed to align amino acid sequences of all samples (including gaps) to the same length using Clustal Omega.[54] A binary representation was introduced to process the amino acid sequences in the ML framework, where M is the type of amino acid and N is the number of residues, including gaps after alignment. Amino acid sequences are represented by $M \times N$ {0, 1} binary variables, where M is the type of amino acid and N is the amino acid sequence length. In addition, the NAD- and NADP-types were represented as 0 and 1, respectively, to treat them as teacher labels. A logistic regression model was used for training, with 30% of the total data used for test data and 70% used for training data. The partial regression coefficient β is the weight parameter of the logistic function; it was optimized by the steepest descent method. In particular, the cross-entropy was used as a loss function and the gradient descent method was used to minimize the cross-entropy loss.

Plasmid Construction. Full-length MaeB derived from *E. coli* strain K12 was obtained from the ASKA clone library.[55] This MaeB had a 6×His-tag at the N-terminus and was subcloned into the pCA24N vector. The DNA fragment of trcMaeB (also known as the ME region of MaeB) was made by PCR amplification of the pCA24N-MaeB template using PrimeSTAR GXL DNA polymerase (Takara Bio, Kusatsu, Japan). The amplified fragment was inserted into a pCA24N vector previously digested with *Sfi*I using Gibson assembly (reagents were acquired from New England Biolabs, Ipswich, MA) and then transformed into *E. coli* strain AG1. The DNA fragments of trcMaeB10−100 mutants were codon-optimized for *E. coli* strain K12 (Thermo Fisher Scientific, Waltham, MA), amplified by PCR, gel-purified, and inserted into pCA24N, as described above. pCA24N-trcMaeB K237Q was generated by PCR-based site-directed mutagenesis using mutant primers and trcMaeB as the template to obtain two fragments. Overlap extension PCR used the two obtained fragments as a template to generate the full-length gene, which was cloned into the pCA24N vector, as described above. All strains harboring the constructs used in this study were selected on LB agar plates containing 50 μg mL^{-1} chloramphenicol. Sequences were confirmed using Sanger sequencing.

Protein Expression and Purification. *Escherichia coli* strain AG1 was transformed with the resultant plasmids and spread onto an LB agar plate containing 50 μg mL^{-1} chloramphenicol. Single colonies were randomly picked and grown in an LB liquid media containing 50 μg mL^{-1} chloramphenicol at 37 °C. The overnight cell culture was inoculated into LB media supplemented with 50 μg mL^{-1} chloramphenicol in a baffled Erlenmeyer flask. The cells were cultivated with shaking at 37 °C until the optical density at 600 nm reached 0.5−0.6, followed by the addition of isopropyl-β-D-thiogalactoside to a final concentration of 0.5 mM to induce protein expression, and the cells were cultivated for a further 3 h. Cells were harvested by centrifugation and resuspended in buffer A (50 mM Tris-HCl pH 7.4, 200 mM NaCl, and 5 mM imidazole). The suspension was sonicated eight times (30 s sonication with 30 s intervals between the treatments) using Bioruptor (Sonicbio, Knanagawa, Japan). The supernatant was collected *via* centrifugation. The supernatant was filtered using a membrane filter (pore size 0.22 μm; Merck, Kenilworth, NJ), followed by the immobilization of the expressed 6× His-tagged proteins onto Ni-NTA beads in batches, washed with buffer B (50 mM Tris-HCl pH 7.4, 200 mM NaCl, and 50 mM imidazole), and eluted with buffer C (50 mM Tris-HCl pH 7.4, 200 mM NaCl, 300 mM imidazole). The protein was buffer-exchanged into 50 mM Tris-HCl pH 7.4 using a PD-10 column (Cytiva, Marlborough, MA). All purification and desalting steps were performed at 4 °C in a cold room. Proteins were then concentrated using an Amicon 10-kDa cutoff Ultra centrifugal filter device (Merck). Protein concentrations were measured by recording the absorbance at 280 nm. Purities of the recombinant proteins were determined using SDS-PAGE.

Kinetic Assay. Enzyme activity was determined by dynamically measuring the absorbance at 340 nm of NAD(P)H produced during the oxidative decarboxylation of L-malic acid to pyruvate using the UV−vis spectrophotometer (V-750; Jasco, Tokyo, Japan). The following molar extinction coefficient for NAD(P)H was used for all calculations: 6.22 mM^{-1} cm^{-1}. All compounds of the reaction mixture were pipetted into a 1 cm light path cuvette, and reactions were initiated by the addition of enzyme solution. The reaction media was composed of 1 μM purified ME (full-length MaeB, trcMaeB, trcMaeB10−100 variants, or trcMaeB K273Q), 67 mM Tris-HCl pH 7.4, 5 mM $MnCl_2$, 3.3 mM L-malic acid, and 0.2−3.2 mM NAD(P)$^+$. Apparent Michaelis−Menten parameters were determined for both cofactors (NAD$^+$ and NADP$^+$) by varying their concentrations about the K_M, while other components were at saturating concentrations. The initial rates at variable NAD(P)$^+$ concentrations were fitted to the Michaelis−Menten model using R software (v. 4.1.1).

Homology Modeling. Model structures for trcMaeB and its mutants were generated using the RosettaCM protocol.[56] Template structures against trcMaeB amino acid sequences were searched in Protein BLAST using PDB as the database, with three structures showing the highest % identity selected (PDB IDs: 2DVM, 5CEE, and 6ZN4). 3-mer and 9-mer fragments were prepared on the Robetta fragment server (http://robetta.bakerlab.org/fragmentsubmit.jsp) to fill in for missing residues during the hybridization step. Target sequences were threaded into prealigned templates using Rosetta's partial thread application after obtaining all input files.[56] The threaded models were passed to the hybridization application using Rosetta XML scripts;[57] 1500 models were generated per run, and the model with the lowest total energy was chosen as the final model. The RosettaCM operating procedure followed the RosettaCommons manual (https://www.rosettacommons.org/docs/latest/application_

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

documentation/structure_prediction/RosettaCM). Input files and command lines are listed in the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.2c00315.

Molecular phylogeny of the malic enzyme (Figure S1); histogram of amino acid sequence length of the malic enzyme used in the machine learning dataset (Figure S2); sequence organization and model structures of MaeB from *E. coli* strain K12 (Figure S); sequence alignment of trcMaeB and its mutants (Figure S4); dependence of initial reaction velocity on $NAD(P)^+$ concentrations (Figure S5); scatter plots of the Rosetta total energy values of the variant models created with the RosettaCM protocol and their violin plots (Figure S6); ME dimer crystal structure of MaeB from *B. bacteriovorus* HD100 (Figure S7); RMSD values and TM-scores of pairwise structures (Table S1 and S2); the scores of *k*-fold cross-validation with four different ML models (Table S3); and input files and command lines for homology modeling (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Teppei Niide − *Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan;* orcid.org/0000-0001-7555-2318; Email: tniide@ist.osaka-u.ac.jp

Hiroshi Shimizu − *Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan;* orcid.org/0000-0002-8986-0861; Email: shimizu@ist.osaka-u.ac.jp

Authors

Sou Sugiki − *Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan*

Yoshihiro Toya − *Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.2c00315

Author Contributions

T.N., Y.T., and H.S. conceived the research. S.S. and T.N. performed the machine learning computations, computational modeling, and wet lab experiments. All authors participated in the data interpretation. T.N. and H.S. supervised the project. T.N. wrote the paper. All authors read and approved the final manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Japan Society for the Promotion of Science KAKENHI (Grant Number 19H05626 to H.S. and Y.T. and 22K04841 to T.N.) and ACT-X, Japan Science and Technology Agency (Grant Number 20348865 to T.N.) This work was partly achieved through the use of large-scale computer systems at the Cybermedia Center, Osaka University.

■ REFERENCES

(1) Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **1998**, *31*, 125−131.

(2) Weinreich, D. M.; Delaney, N. F.; DePristo, M. A.; Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **2006**, *312*, 111−114.

(3) Miton, C. M.; Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **2016**, *25*, 1260−1272.

(4) Starr, T. N.; Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **2016**, *25*, 1204−1218.

(5) Baier, F.; Hong, N.; Yang, G.; Pabis, A.; Miton, C. M.; Barrozo, A.; Carr, P. D.; Kamerlin, S. C. L.; Jackson, C. J.; Tokuriki, N. Cryptic genetic variation shapes the adaptive evolutionary potential of enzymes. *eLife* **2019**, *8*, No. e40789.

(6) Dellus-Gur, E.; Elias, M.; Caselli, E.; Prati, F.; Salverda, M. L. M.; De Visser, J. A. G. M.; Fraser, J. S.; Tawfik, D. S. Negative Epistasis and Evolvability in TEM-1 β-Lactamase—The Thin Line between an Enzyme's Conformational Freedom and Disorder. *J. Mol. Biol.* **2015**, *427*, 2396−2409.

(7) Yano, T.; Oue, S.; Kagamiyama, H. Directed evolution of an aspartate aminotransferase with new substrate specificities. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5511−5515.

(8) Oue, S.; Okamoto, A.; Yano, T.; Kagamiyama, H. Redesigning the Substrate Specificity of an Enzyme by Cumulative Effects of the Mutations of Non-active Site Residues. *J. Biol. Chem.* **1999**, *274*, 2344−2349.

(9) Antikainen, N. M.; Martin, S. F. Altering protein specificity: techniques and applications. *Bioorg. Med. Chem.* **2005**, *13*, 2701−2716.

(10) Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Gould, S. M. Q.; Roodveldt, C.; Tawfik, D. S. The "evolvability" of promiscuous protein functions. *Nat. Genet.* **2005**, *37*, 73−76.

(11) Tomatis, P. E.; Rasia, R. M.; Segovia, L.; Vila, A. J. Mimicking natural evolution in metallo-β-lactamases through second-shell ligand mutations. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13761−13766.

(12) Schmidt, M.; Hasenpusch, D.; Kähler, M.; Kirchner, U.; Wiggenhorn, K.; Langel, W.; Bornscheuer, U. T. Directed Evolution of an Esterase from Pseudomonas fluorescens Yields a Mutant with Excellent Enantioselectivity and Activity for the Kinetic Resolution of a Chiral Building Block. *ChemBioChem* **2006**, *7*, 805−809.

(13) Yang, G.; Hong, N.; Baier, F.; Jackson, C. J.; Tokuriki, N. Conformational tinkering drives evolution of a promiscuous activity through indirect mutational effects. *Biochemistry* **2016**, *55*, 4583−4593.

(14) Goldsmith, M.; Aggarwal, N.; Ashani, Y.; Jubran, H.; Greisen, P.; Ovchinnikov, S.; Leader, H.; Baker, D.; Sussman, J. L.; Goldenzweig, A.; Fleishman, S. J.; Tawfik, D. S. Overcoming an optimization plateau in the directed evolution of highly efficient nerve agent bioscavengers. *Protein Eng., Des. Sel.* **2017**, *30*, 333−345.

(15) Tracewell, C. A.; Arnold, F. H. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. *Curr. Opin. Chem. Biol.* **2009**, *13*, 3−9.

(16) Bloom, J. D.; Arnold, F. H. In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 9995−10000.

(17) Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. De novo computational design of retro-aldol enzymes. *Science* **2008**, *319*, 1387−1391.

(18) Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453*, 190−195.

(19) Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; Clair, J. L., St; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* **2010**, *329*, 309−313.
(20) Bedbrook, C. N.; Yang, K. K.; Rice, A. J.; Gradinaru, V.; Arnold, F. H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Comput. Biol.* **2017**, *13*, No. e1005786.
(21) Wu, Z.; Jennifer Kan, S. B.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 8852−8858.
(22) Fox, R. J.; Davis, S. C.; Mundorff, E. C.; Newman, L. M.; Gavrilovic, V.; Ma, S. K.; Chung, L. M.; Ching, C.; Tam, S.; Muley, S.; Grate, J.; Gruber, J.; Whitman, J. C.; Sheldon, R. A.; Huisman, G. W. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **2007**, *25*, 338−344.
(23) Romero, P. A.; Krause, A.; Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, E193−E201.
(24) Bedbrook, C. N.; Yang, K. K.; Robinson, J. E.; Mackey, E. D.; Gradinaru, V.; Arnold, F. H. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **2019**, *16*, 1176−1184.
(25) Saito, Y.; Oikawa, M.; Nakazawa, H.; Niide, T.; Kameda, T.; Tsuda, K.; Umetsu, M. Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. *ACS Synth. Biol.* **2018**, *7*, 2014−2022.
(26) Jakub, O.; Plotkin, J. B. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, E2301−E2309.
(27) Wrenbeck, E. E.; Azouz, L. R.; Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **2017**, *8*, No. 15695.
(28) Khoury, G. A.; Fazelinia, H.; Chin, J. W.; Pantazes, R. J.; Cirino, P. C.; Maranas, C. D. Computational design of Candida boidinii xylose reductase for altered cofactor specificity. *Protein Sci.* **2009**, *18*, 2125−2138.
(29) Scrutton, N. S.; Berry, A.; Perham, R. N. Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. *Nature* **1990**, *343*, 38−43.
(30) Bastian, S.; Liu, X.; Meyerowitz, J. T.; Snow, C. D.; Chen, M. M. Y.; Arnold, F. H. Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in *Escherichia coli*. *Metab. Eng.* **2011**, *13*, 345−352.
(31) Cahn, J. K. B.; Werlang, C. A.; Baumschlager, A.; Brinkmann-Chen, S.; Mayo, S. L.; Arnold, F. H. A General Tool for Engineering the NAD/NADP Cofactor Preference of Oxidoreductases. *ACS Synth. Biol.* **2017**, *6*, 326−333.
(32) Maxel, S.; Saleh, S.; King, E.; Aspacio, D.; Zhang, L.; Luo, R.; Li, H. Growth-Based, High-Throughput Selection for NADH Preference in an Oxygen-Dependent Biocatalyst. *ACS Synth. Biol.* **2021**, *10*, 2359−2370.
(33) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct., Funct., Bioinf.* **2004**, *57*, 702−710.
(34) Harding, C. J.; Cadby, I. T.; Moynihan, P. J.; Lovering, A. L. A rotary mechanism for allostery in bacterial hybrid malic enzymes. *Nat. Commun.* **2021**, *12*, No. 1228.
(35) Huergo, L. F.; Araújo, G. A. T.; Santos, A. S. R.; Gerhardt, E. C. M.; Pedrosa, F. O.; Souza, E. M.; Forchhammer, K. The NADP-dependent malic enzyme MaeB is a central metabolic hub controlled by the acetyl-CoA to CoASH ratio. *Biochim. Biophys. Acta, Proteins Proteomics* **2020**, *1868*, No. 140462.
(36) Bologna, F. P.; Andreo, C. S.; Drincovich, M. F. Escherichia coli malic enzymes: Two isoforms with substantial differences in kinetic properties, metabolic regulation, and structure. *J. Bacteriol.* **2007**, *189*, 5937−5946.
(37) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(38) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Dustin Schaeffer, R.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Christopher Garcia, K.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.
(39) Ryu, J. Y.; Kim, H. U.; Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 13996−14001.
(40) Gainza, P.; Sverrisson, F.; Monti, F.; Rodolà, E.; Boscaini, D.; Bronstein, M. M.; Correia, B. E. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **2020**, *17*, 184−192.
(41) Shroff, R.; Cole, A. W.; Diaz, D. J.; Morrow, B. R.; Donnell, I.; Annapareddy, A.; Gollihar, J.; Ellington, A. D.; Thyer, R. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synth. Biol.* **2020**, *9*, 2927−2935.
(42) Khersonsky, O.; Roodveldt, C.; Tawfik, D. S. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **2006**, *10*, 498−508.
(43) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29*, 7133−7155.
(44) Sternke, M.; Tripp, K. W.; Barrick, D. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 11275−11284.
(45) Porebski, B. T.; Buckle, A. M. Consensus protein design. *Protein Eng. Des. Sel.* **2016**, *29*, 245−251.
(46) Goldenzweig, A.; Goldsmith, M.; Hill, S. E.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; Lieberman, R. L.; Aharoni, A.; Silman, I.; Sussman, J. L.; Tawfik, D. S.; Fleishman, S. J. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2016**, *63*, 337−346.
(47) Peleg, Y.; Vincentelli, R.; Collins, B. M.; Chen, K. E.; Livingstone, E. K.; Weeratunga, S.; Leneva, N.; Guo, Q.; Remans, K.; Perez, K.; Bjerga, G. E. K.; Larsen, Ø.; Vaněk, O.; Skořepa, O.; Jacquemin, S.; Poterszman, A.; Kjær, S.; Christodoulou, E.; Albeck, S.; Dym, O.; Ainbinder, E.; Unger, T.; Schuetz, A.; Matthes, S.; Bader, M.; de Marco, A.; Storici, P.; Semrau, M. S.; Stolt-Bergner, P.; Aigner, C.; Suppmann, S.; Goldenzweig, A.; Fleishman, S. J. Community-Wide Experimental Evaluation of the PROSS Stability-Design Method. *J. Mol. Biol.* **2021**, *433*, No. 166964.
(48) Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T. G.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.;

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985

Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M. C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L. S.; Zhang, J.; et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.

(49) Katoh, K.; Misawa, K.; Kuma, K. I.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059−3066.

(50) Minh, B. Q.; Schmidt, H. A.; Chernomor, O.; Schrempf, D.; Woodhams, M. D.; Von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530−1534.

(51) Kalyaanamoorthy, S.; Minh, B. Q.; Wong, T. K. F.; Von Haeseler, A.; Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587−589.

(52) Hoang, D. T.; Chernomor, O.; Von Haeseler, A.; Minh, B. Q.; Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **2018**, *35*, 518−522.

(53) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27−30.

(54) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.

(55) Kitagawa, M.; Ara, T.; Arifuzzaman, M.; Ioka-Nakamichi, T.; Inamoto, E.; Toyonaga, H.; Mori, H. Complete set of ORF clones of Escherichia coli ASKA library (A Complete Set of E. coli K -12 ORF Archive): Unique Resources for Biological Research. *DNA Res.* **2006**, *12*, 291−299.

(56) Song, Y.; Dimaio, F.; Wang, R. Y. R.; Kim, D.; Miles, C.; Brunette, T.; Thompson, J.; Baker, D. High-Resolution Comparative Modeling with RosettaCM. *Structure* **2013**, *21*, 1735−1742.

(57) Fleishman, S. J.; Leaver-Fay, A.; Corn, J. E.; Strauch, E. M.; Khare, S. D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G.; Meiler, J.; Baker, D. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLoS One* **2011**, *6*, No. e20161.

https://doi.org/10.1021/acssynbio.2c00315
ACS Synth. Biol. 2022, 11, 3973−3985