ACS Catalysis

pubs.acs.org/acscatalysis Research Article

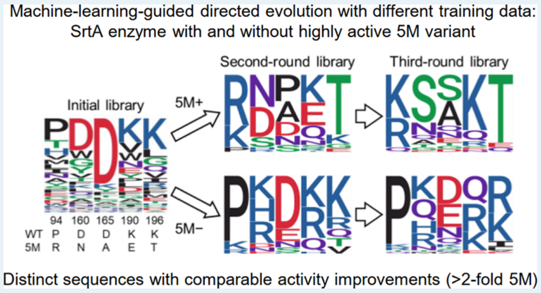
Machine-Learning-Guided Library Design Cycle for Directed Evolution of Enzymes: The Effects of Training Data Composition on Sequence Space Exploration

Yutaka Saito, Misaki Oikawa, Takumi Sato, Hikaru Nakazawa, Tomoyuki Ito, Tomoshi Kameda, Koji Tsuda,* and Mitsuo Umetsu*

Cite This: *ACS Catal.* 2021, 11, 14615−14624 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning (ML) is becoming an attractive tool in mutagenesis-based protein engineering because of its ability to design a variant library containing proteins with a desired function. However, it remains unclear how ML guides directed evolution in sequence space depending on the composition of training data. Here, we present a ML-guided directed evolution study of an enzyme to investigate the effects of a known “highly positive” variant (i.e., variant known to have high enzyme activity) in training data. We performed two separate series of ML-guided directed evolution of Sortase A with and without a known highly positive variant called 5M in training data. In each series, two rounds of ML were conducted: variants predicted by the initial round were experimentally evaluated and used as additional training data for the second-round of prediction. The improvements in enzyme activity were comparable between the two series, both achieving enzyme activity 2.2−2.5 times higher than 5M. Intriguingly, the sequences of the improved variants were largely different between the two series, indicating that ML guided the directed evolution to the distinct regions of sequence space depending on the presence/absence of the highly positive variant in the training data. This suggests that the sequence diversity of improved variants can be expanded not only by conventional ML using the whole training data but also by ML using a subset of the training data even when it lacks highly positive variants. In summary, this study demonstrates the importance of regulating the composition of training data in ML-guided directed evolution.


Machine-learning-guided directed evolution with different training data:
SrtA enzyme with and without highly active 5M variant
Initial library
Second-round library
Third-round library
5M+
5M−
94 160 165 190 196
WT P D D K K
5M R N A E T
Distinct sequences with comparable activity improvements (>2-fold 5M)

KEYWORDS: *machine learning, mutagenesis, protein engineering, directed evolution, library design, training data, sequence space exploration*

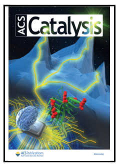
■ INTRODUCTION

The diversity of amino acid sequences generates various protein functions, allowing us to create novel functions not found in nature. While genomics studies have provided vast information on natural proteins, they cover only a small subset of sequence space (defined as the set of all possible amino acid sequences of a protein). Thus, mutagenesis-based protein engineering has been used to create novel protein variants with desired functions by introducing amino acid mutations into natural proteins, which is called directed evolution.[1]

Current experimental approaches for directed evolution often fail to obtain desirable variants due to the difficulty in exploring sequence space. Large libraries with the size of more than 10^9 variants can be generated; the larger the library size becomes, the higher the probability that the library contains the optimal variant is. However, the present screening throughput is far below the large library size, which limits functional improvements that are obtainable in reasonable time.[2−6] Iterative saturation mutagenesis (ISM) is a conventional method where saturation mutagenesis (substitution of one or a few residues for all possible amino acids) and selection of the optimal variant are iteratively conducted.[7] Given the appropriate choice of residues for mutagenesis, this method has the potential to obtain a library containing desirable variants with high probability after multiple iterations.[8] As such, the success of directed evolution crucially depends on preparing a small library with high enrichment of desirable variants.

Recently, machine learning (ML) has been used to accelerate directed evolution of proteins.[9−14] In this method, saturation mutagenesis and/or random mutagenesis are performed to generate an initial library. The variants in the library are experimentally evaluated to obtain their sequences

Received: August 18, 2021
Revised: October 21, 2021
Published: November 19, 2021



and functions and then used as training data to construct a ML model that predicts the function from the sequence. By using the ML model, a second-round library that contains variants predicted to have improved functions is generated. This method enables us to design a library with high enrichment of desirable variants and thus has been successfully applied to directed evolution of various proteins including fluorescent proteins,[15−17] enzymes,[18−20] and others.[21,22] However, it remains unclear how ML-guided directed evolution explores sequence space depending on the composition of training data. Will different training data lead to different regions of sequence space and different levels of functional improvements?

Here, we perform ML-guided directed evolution with different training data to investigate the differences in explored regions of sequence space and resultant functional improvements. As an example, we used Sortase A (SrtA), a transpeptidase enzyme found in Gram-positive bacteria. SrtA conjugates a peptide containing an LPXTG sequence to another peptide containing a GGG sequence.[23] This sequence-specific transpeptidase activity has been utilized for protein modification and conjugation of proteins to proteins, polymers, and solid substrates in in vivo and in vitro.[24−28] Although the wild-type SrtA has poor reaction kinetics that limit its application, several variants with superior activity have been reported.[25,29,30] To make different training data for ML, two typical scenarios encountered in protein engineering were considered: (1) the situation where we already have a "highly positive" variant (variant with a desired function) at least to some degree, and (2) the situation where no such variant is available. Accordingly, we conducted two separate series of ML-guided directed evolution with and without a highly positive SrtA variant called 5M[29] in training data. Interestingly, variants with higher enzyme activity than 5M were discovered by both series with comparable improvements (2.2−2.5 times higher than 5M), while the sequences of the improved variants were largely different between the two series. This demonstrates the ability of ML to guide directed evolution to distinct regions of sequence space depending on the composition of training data, which is useful to expand the sequence diversity of improved variants.

RESULTS

ML-Guided Library Design Cycle. In our previous study,[15] we proposed a ML-guided directed evolution method for fluorescent proteins. In this method, several residues at certain positions in a target fluorescent protein were mutated to make a variant library for training a Bayesian ML model, and then, the ML model was used to rank all possible variants according to their predicted probability of having improved fluorescence performance. Here, we extend this method to enzymes with iterative design of libraries (Figure 1; Methods). In brief, certain residues in a target enzyme are mutated to make an initial library, and functional data (sequence, expression level, and enzyme activity) for approximately 100 variants in this library are obtained. These data are then used to train a ML model, and top-ranked variants predicted by ML are used to design the second-round library. The functional data for dozens of variants in the second-round library are obtained and used as additional training data for the ML model to design the third-round library. We applied this ML-guided library design cycle to SrtA for improving enzyme activity.

Preparation of the Initial Library. To select the residues that should be mutated in SrtA, we referred to the reported high activity variant called 5M[29] that has five mutated residues: P94R, D160N, D165A, K190E, and K196T (Figure 2). At these five residues, point random mutagenesis and site-directed random mutagenesis at all five residues were applied to make the initial library. The expression vectors for this library were prepared as described in our previous study[15] and used to transform *Escherichia coli*. Transformants were cultured in a 96-deep-well plate, and the sequence of each variant was determined. Then, each variant was purified by immobilized metal affinity chromatography (IMAC), and the expression level was measured. Variants with sufficient expression levels were used for enzyme activity assay. In this assay, two substrates—cycle3 green fluorescent protein (GFP) with an LPETG sequence at the C-terminus and Venus yellow fluorescent protein (YFP) with a GGG sequence at the N-terminus—were conjugated with each other by an SrtA variant[25] (Figure S1), and the enzyme activity was estimated by SDS-PAGE analysis. In total, 80 variants were prepared (56 from site-directed random mutagenesis at all five residues and 24 from point random mutagenesis). The enzyme activity of 45 of these variants was measured while the expression levels of the 35 other variants were insufficient for measuring enzyme activity. The sequences of these variants are shown in Table S1. Most variants had low expression levels and enzyme activity (Figure 3); only 4 of the 45 measured variants showed higher enzyme activity than the wild type, and all variants

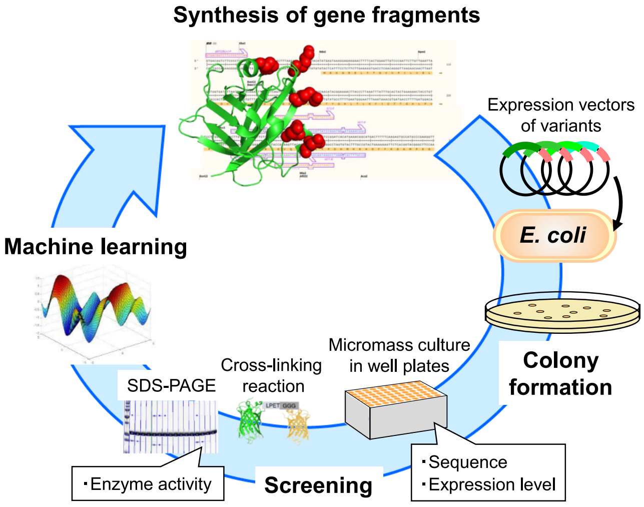


Figure 1. Overview of the ML-guided library design cycle for directed evolution of SrtA. See the main text for the detail explanation.

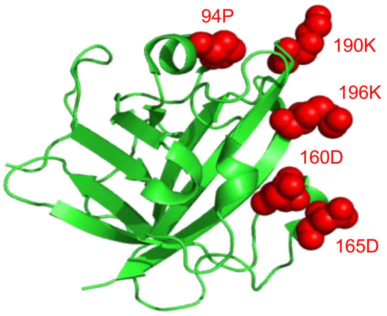


Figure 2. Structure of SrtA (PDB ID: 1T2P). The five mutated residues (94P, 160D, 165D, 190K, and 196K) are colored in red.

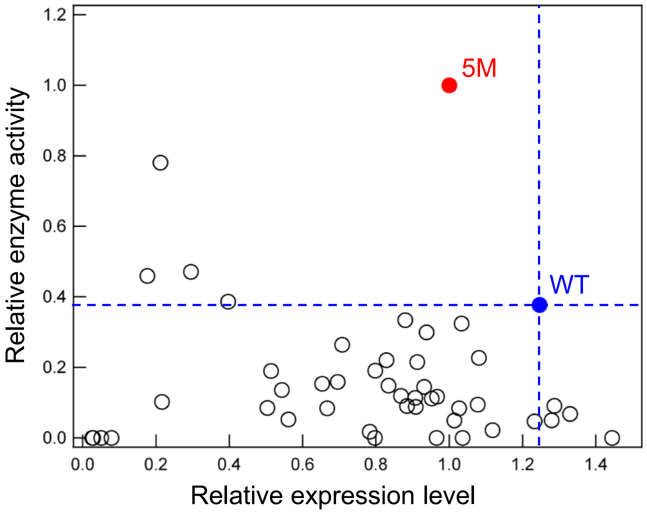


Figure 3. Expression levels and enzyme activity of the SrtA variants in the initial library. Values are normalized by those of 5M. Dashed lines represent the values of the wild type (WT). Only variants with measured enzyme activity (Table 1) are shown.

Table 1. Summary of the Initial Library[a]

	initial 5M+ library	initial 5M− library
activity measured	45	45
low expression	35	35
wild type	1	1
5M	1	0
total	82	81

[a] The numbers of SrtA variants used as training data for ML are shown. The initial 5M+ library contains a highly positive variant 5M while the initial 5M− library does not. Activity measured: variants with measured enzyme activity. Low expression: variants with insufficient expression levels for measuring enzyme activity. See Table S1 for the sequence information.

showed lower enzyme activity than 5M. These data were used to train a ML model.

ML with the Initial Library. We constructed a ML model based on a Gaussian process that predicts the enzyme performance score of an SrtA variant from its sequence (Methods; eqs 1 and 2). The enzyme performance score was defined as follows:

$$y = \sigma(\text{activity}_n - 1)$$

where $\sigma(\cdot)$ is a sigmoidal function, and activity$_n$ is the enzyme activity of an SrtA variant divided by that of the wild type. We set activity$_n$ as zero if the expression level is insufficient for measuring enzyme activity (e.g., the 35 low-expressed variants in Table S1). This definition of the enzyme performance score allows the ML model to predict variants with high enzyme activity while avoiding variants with insufficient expression levels. For feature vectors used in the ML model, we considered a variety of amino acid descriptors based on physicochemical properties, structural topology, and evolutionary information as well as their combinations. In benchmark experiments, we found that the combination of Z-scale[31] and position-specific score matrix (PSSM)[32] achieved the best prediction accuracy (Methods). Z-scale is an amino acid descriptor defined by dimensionality reduction of numerous features including experimentally measured values such as retention times in thin-layer chromatography and shifts in nuclear magnetic resonance, as well as calculated values such as molecular orbital indices, total, polar, and nonpolar surface area, van der Waals volume of the side chain, log P, molecular weight, and the indicators of hydrogen bond donor/acceptor properties, and side chain charge.[31] PSSM was used as features for incorporating evolutionary information from SrtA homologs (Methods). The dimensionality of our final feature vector was 6 per residue. Thus, the number of features used in our final model was 30 (i.e., 6 dimensions × 5 mutated residues).

To train the ML model, two different data sets were prepared (Table 1): one consisted of the 80 variants from the initial library together with the wild type and 5M (initial 5M+ library) and the other consisted of the 80 variants and the wild type but excluded 5M (initial 5M− library). These data sets reflected practical scenarios in protein engineering: the 5M+ data set corresponds to the situation where we already have a highly positive variant for a target protein (e.g., 5M for SrtA) and hope to discover superior variants, and the 5M− data set simulates the situation where no such variant is available. We constructed two ML models by using the 5M+ and 5M− data sets, respectively, as training data. By using each ML model, we ranked all possible variants regarding the five mutated residues (i.e., 20^5 = 3,200,000 variants) excluding those in the training data (Data S1 and S2 for 5M+ and 5M−, respectively). This experimental design allowed us to evaluate the effects of a highly positive variant on prediction results in ML-guided directed evolution.

In the prediction result for 5M+, the six high activity variants reported previously[29] were ranked within the top 2% (Table 2), implying that the ML prediction was reliable. The prediction result for 5M− showed comparable reliability except that 5M, which has been reported to have the highest activity among these six variants,[29] had a slightly low rank within the top 5%.

Table 2. Ranks of the Reported High Activity SrtA Variants by ML Prediction[a]

	prediction 5M+		prediction 5M−	
variants[29]	rank	% with this rank or above	rank	% with this rank or above
5M (RNAET)			161,052	5.0
SNAKT	14,991	1.5	59,093	1.9
SNDKT	10,052	0.3	2121	0.1
SNAKK	34,895	1.1	22,245	0.7
PNDET	27,149	0.8	5798	0.2
SDAKK	42,412	1.3	19,104	0.6

[a] Ranks among all possible variants (20^5 = 3,200,000) excluding those in the training data are shown. "Prediction 5M+" and "prediction 5M−" represent the prediction results by the ML models trained with the initial 5M+ and 5M− libraries, respectively. Variants are represented by amino acids at the five mutated residues (94, 160, 165, 190, and 196).

Screening the in the Second-Round Library. To assess the performance of the predicted variants, the second-round library was made for each of 5M+ and 5M− by means of the mix primer method (Methods). The primer sequences were designed so that the top 50 variants in the ranking list were contained in the library. After cloning, we could prepare 37 variants ranked within the top 50 plus 30 variants ranked within 51−606 for the second-round 5M+ library and 45 variants within the top 50 plus 31 variants ranked within 52−

4238 for the second-round 5M− library (Table S2). Interestingly, most of these variants (except that ranked 29th in the ranking list of 5M−) were prepared with sufficient expression levels for measuring enzyme activity. This suggests that the enzyme performance score used in the ML model, which we designed to incorporate not only enzyme activity but also expression levels (Methods; eq 2), contributed to the prediction of variants with superior expression levels compared to those in the initial library.

For both second-round 5M+ and 5M− libraries, many variants showed higher enzyme activity than the wild type (Figure 4). Three variants in the second-round 5M− library showed higher enzyme activity than 5M, while a few variants in the second-round 5M+ library showed comparable enzyme activity to 5M (Figure S2). These results indicate that ML successfully designed the second-round libraries with high enrichment of high activity variants.

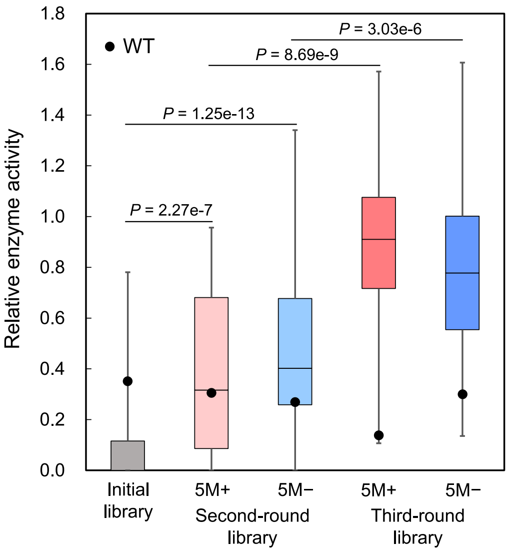


Figure 4. Enzyme activity of the SrtA variants in the initial and second- and third-round libraries. Values are normalized by that of 5M. Black dots represent the values of the wild type (WT). In all box plots, 5M and WT are not included in their distributions. Variants with insufficient expression levels for measuring enzyme activity are plotted as enzyme activity of zero; in the box plots of the initial and second-round libraries, lower bounds are zero due to these variants. *P*-values were calculated by the one-sided Mann−Whitney *U* test.

Iteration of ML-Guided Library Design. The results of the second-round libraries were used as additional training data for the ML models. We used not only variants within the top 50 but also those below the top 50 (Table S2) for incorporating all available information into the ML models. The third-round libraries were prepared by the same method as used for the second-round libraries (Data S3 and S4; Table S3). Among these, only a small fraction of variants showed insufficient expression levels for measuring enzyme activity: 2 out of 50 (4%) for 5M+ and 1 out of 50 (2%) for 5M− (Table S3). Most variants showed higher enzyme activity than the wild type (Figure 4) of which 19 and 12 variants in the case of 5M+ and 5M−, respectively, were also higher than 5M (Figure S3). Thus, the iteration of ML-guided library design dramatically increased the fraction of high activity variants, enabling the discovery of many variants with superior activity to 5M.

We evaluated the enzyme activity of the improved variants in detail. For both 5M+ and 5M− libraries, we selected the top variant from the second-round libraries and the top three variants from the third-round libraries in terms of enzyme activity measured in the above screening process (Table 3).

Table 3. Amino Acid Residues in the Top Variants (in Terms of Measured Enzyme Activity) from the Second- and Third-Round Libraries

	amino acid residues				
	94	160	165	190	196
wild type	P	D	D	K	K
5M	R	N	A	E	T
5M+					
second-round top1	K	N	D	K	T
third-round top1	R	S	D	K	T
third-round top2	K	T	A	K	T
third-round top3	K	S	D	K	T
5M−					
second-round top1	K	N	D	R	K
third-round top1	P	K	N	Q	R
third-round top2	R	K	D	K	K
third-round top3	P	K	E	R	R

These variants were individually prepared by culture in 1 L flasks and purified by IMAC and size exclusion chromatography (SEC). The enzyme activity of the purified variants was measured under various concentrations of the substrate GFP and a single concentration of the substrate YFP (Figure 5); the range of GFP concentrations extended higher than that used in the above screening process. All the variants showed higher enzyme activity than the wild type and 5M. In particular, the variants from the third-round libraries showed the enzyme activity 4.4−5.0 and 2.2−2.5 times higher than the wild type and 5M, respectively, at the highest GFP concentration tested (120 μM). Remarkably, the improvements in enzyme activity were comparable between the 5M+ and 5M− libraries. This suggests that ML-guided directed evolution can discover improved variants even when the training data lack highly positive variants such as 5M for SrtA.

Trajectory of Directed Evolution in Sequence Space. Despite the comparable improvements in enzyme activity, the sequence profiles of the variants were largely different between the 5M+ and 5M− libraries (Figure 6a). In the case of 5M+, the amino acids appearing in the 5M sequence were frequently found in the variants from the second-round library but their frequencies tended to become lower in the third-round library. Concomitantly, in the case of 5M−, the amino acids in the variants from the second-round library were dominated by those appearing in the wild type sequence, and the frequencies of these amino acids tended to be lower in the third-round library.

To further visualize ML-guided directed evolution in sequence space, we conducted principal component analysis of the variants based on their sequence-derived feature vectors used in the ML model (Figure 6b). In the case of 5M+, ML guided the evolution around 5M, and the highest activity variant was discovered at the position slightly distant from 5M. In contrast, in the case of 5M−, the evolution guided by ML allowed us to find a different region of sequence space with high activity variants around the wild type. These results indicate that ML guided the evolution to the distinct regions of

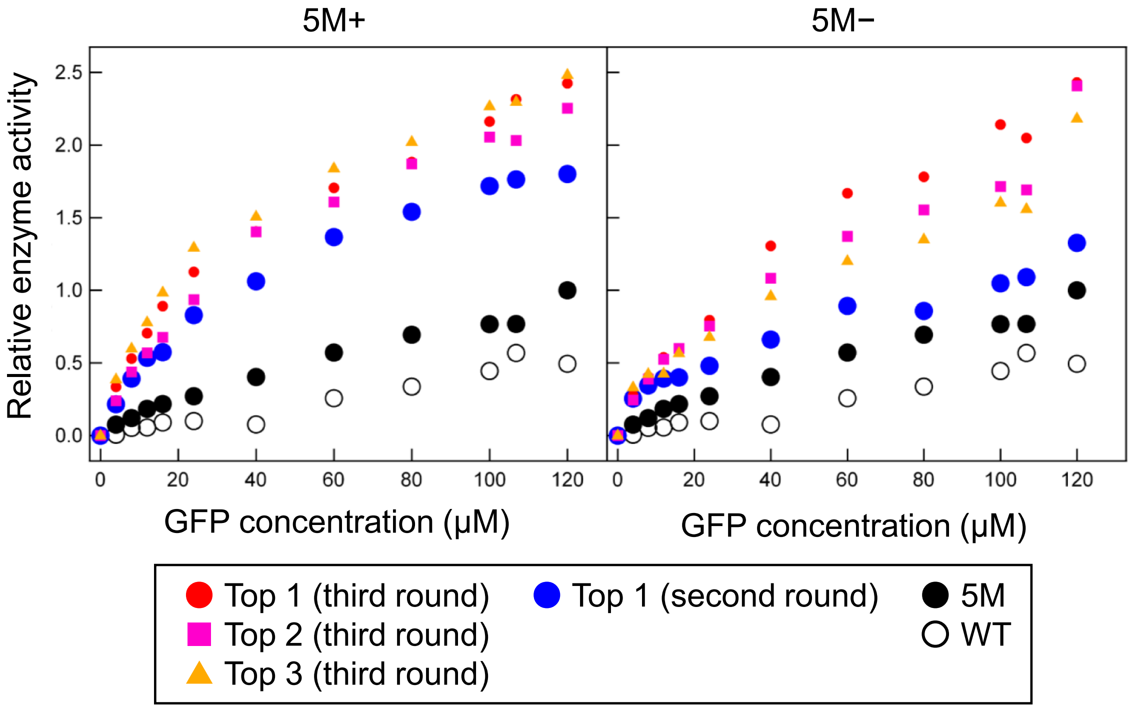


Figure 5. Enzyme activity of the SrtA variants evaluated under different concentrations of the substrate GFP. The concentration of the substrate YFP is fixed at 200 μM. Enzyme activity is normalized by that of 5M at the highest GFP concentration (120 μM). The top variant from the second-round libraries and the top three variants from the third-round libraries (in terms of measured enzyme activity) are shown. WT: wild type.

sequence space depending on the presence/absence of 5M in training data.

DISCUSSION

In this study, ML-guided directed evolution of SrtA was performed to investigate the effects of a highly positive variant 5M in training data. The two series of ML-guided directed evolution with and without 5M explored the distinct regions of sequence space, and both discovered the variants with higher enzyme activity than 5M. These series correspond to typical scenarios encountered in protein engineering: (1) we already have a highly positive variant that has a desired function at least to some degree, and (2) no such variant is available. Our results suggest that ML is useful in both scenarios for discovering improved variants, and each scenario can provide variants with distinct sequences, expanding the sequence diversity of improved variants. The significance of this finding is twofold. First, it is of interest in protein science to understand the fitness landscape of a protein in sequence space. In this study, we revealed that the fitness landscape of SrtA has additional peaks around the wild type in addition to those around the known highly positive variant 5M. Second, expanding the sequence diversity of high activity variants is important for the application of functional proteins. High activity variants may be not usable in application due to their properties other than activity, e.g., solubility, thermostability, immunogenicity, etc. In this case, high activity variants with different sequences serve as alternative choices for application, which may exhibit preferable properties.

Regarding the composition of training data, we focused on the two cases with 5M included and excluded (5M+ and 5M−). To further investigate other compositions, we performed ML using the training data with additional variants removed (Figures S4 and S5). First, we removed the variants whose enzyme activity was the second and third highest next to 5M in the initial library, denoted as 5M−2 and 5M−3, respectively (Figure S4). In sequence space analysis, the top 100 predicted variants were located around the wild type, similarly to the case of 5M−. Nevertheless, 5M−2 and 5M−3 explored the different regions of sequence space not completely overlapped with each other, which was also different from that explored by 5M− (Figure S4b). Next, we removed the 35 variants whose expression levels were insufficient for measuring enzyme activity in the initial library (Figure S5). The top 100 predicted variants showed sequence profiles similar to those obtained by the original ML models (Figure S5a). However, when looking at each variant sequence, these 100 predicted variants missed a substantial fraction of variants in the original second-round libraries (29 out of the 67 variants for 5M+ and 14 out of the 76 variants for 5M−). Remarkably, the variant with the highest enzyme activity in the second-round 5M+ library (the variant "KNDKT" in Table 3) was missed, suggesting the importance of including low-expressed variants in the training data.

In the iteration of ML-guided library design, we conducted the two rounds of ML to obtain the third-round libraries. For testing the saturation of this iterative procedure, the results of the third-round libraries were further used as additional training data for the ML models to predict the fourth-round libraries (Figure S6). We observed that the maximum value of predicted enzyme performance scores was not improved in the fourth round compared to the third round, implying our procedure was already saturated at the third round. Moreover, we also confirmed that the maximum value of predicted enzyme performance scores was increased from the second round to the third round, which was consistent with our experimental evaluation (Figure 4).

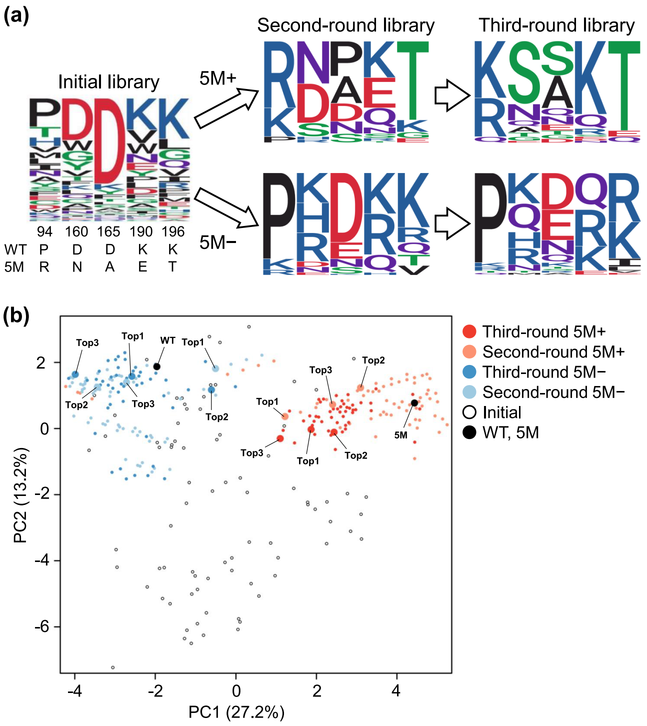


Figure 6. Visualization of the ML-guided directed evolution of SrtA in sequence space. (a) Sequence logo representation of amino acids at the five mutated residues in the initial and second- and third-round libraries. The amino acids in the wild type (WT) and 5M are shown below the sequence logo of the initial library. (b) Principal component analysis of the variants in the initial and second- and third-round libraries. WT, 5M, and the top three variants (Top1, Top2, and Top3) in the second- and third-round libraries (in terms of measured enzyme activity) are tagged at their corresponding positions. The first and the second principal components (PC1 and PC2) are shown with their contribution rates in parentheses.

In the selection of feature vectors, we selected feature vectors that enabled the ML model to find 5M with the smallest amount of training data (Methods; Figure S7). This selection criterion might bias the ML model toward finding 5M. As alternative criteria, it could be considered, e.g., selecting feature vectors that find the variant with the second highest enzyme activity next to 5M or selecting feature vectors that achieve the best model fit in cross-validation.

In ML prediction, we ranked all possible five-point mutants of SrtA; thus, the size of sequence space was 20^5 = 3,200,000. In our previous study,[15] we performed ML-guided directed evolution of fluorescent proteins to explore sequence space of all four-point mutants (20^4 = 160,000). In both studies, the initial training data consisted of approximately 100 variants of which only 3−5% were functionally positive (e.g., the four variants with higher enzyme activity than the wild type in Figure 3). Nonetheless, the libraries proposed by ML were enriched with high performance variants. Therefore, this size of initial training data should be sufficient for discovering high performance variants from the sequence space of 20^4−20^5.

In previous applications of ML in mutagenesis-based protein engineering, training data were prepared by two different methods: (1) each variant is isolated and its function is measured individually[15,18−22] and (2) functions are measured from a mixture of variants via certain information that may correlate with the functions, e.g., read counts in deep sequencing.[16,17,33−35] The quality of training data obtained by the former method can be high while their size is usually limited to tens or hundreds. Thus, this type of training data is suitable when the number of mutated residues is relatively small, as in our present study. The latter method can provide larger training data with the size of thousands or tens of thousands while their quality may be lower due to the indirect measurement of functions such as read counts. In addition, the latter method can only be applied to specific kinds of proteins such as antibodies for which assays from a mixture of variants are established.

The 5M variant was previously discovered by high-throughput screening of SrtA variants using yeast display and fluorescence-activated cell sorting.[29] The library size in this previous study was approximately 10^8. Here, we discovered novel variants with higher enzyme activity than 5M from the region of sequence space around 5M (Figure 6). This indicates that a highly positive variant discovered by massive experimental screening is not always optimal, and ML can discover further improved variants. Moreover, our result using training data without 5M showed an additional ability of ML to discover improved variants from another region of sequence space distant from 5M. These results suggest multiple executions of ML using subsets of the whole training data as a means to expand the sequence diversity of improved variants.

To interpret the molecular nature of the improved SrtA variants, we compared the sequences of the top three variants (in terms of measured enzyme activity) from the third-round libraries (Table 3). At the P94 residue, the amino acids with the NH structure in the side chain were selected in the top variants from both 5M+ and 5M− libraries. At the D160 residue, no acidic amino acids to which the wild type belongs, were selected; instead, polar uncharged and basic amino acids were selected from 5M+ and 5M−, respectively. This may indicate that the acidic side chain is unfavorable for catalysis. For the D165, K190, and K196 residues, the mutations showed different tendencies between 5M+ and 5M−. In 5M+, the top variants had the amino acids of either the wild type or 5M (e.g., the D165 residue showed either the wild-type's D or 5M's A); the directed evolution of SrtA seems punctuated like the jump between the wild type and 5M. On the other hand, in 5M−, the amino acids with similar physicochemical properties to the wild type were selected, which looks like the fine tuning around the wild type. The peptide conjugation by SrtA is two substrate reactions, and SrtA catalysis is reversible; 5M is reported to have the higher affinity for the LPXTG-peptide than the wild type but low affinity for the GGG-peptide,[29] so that 5M shows the higher level of hydrolytic reaction.[36,37] NMR analysis shows the chemical shift change of D165, K190, and K196 by adding the GGG-peptide to the complex of SrtA and LPAT-peptide,[38] and the point mutation of D165N drastically decreases the affinity for the GGG-peptide,[29] suggesting the importance of these residues for the affinity with the GGG-peptide. Thus, we speculate that the back mutations to the wild type amino acids observed at the D165 and K190 residues might improve the affinity for the GGG-peptide.

In this study, we investigated the effect of training data composition using SrtA and a highly positive variant 5M as a model case. To address whether a similar effect of highly positive variants can be observed in other proteins, we conducted computational experiments on the GB1 data set reported in a previous study.[39] This data set provides fitness values (binding activity) of the GB1 protein for all possible variants across four residues (20^4 = 160,000), enabling us to conduct computational experiments mimicking the case of SrtA (Methods). Specifically, we sampled the training data from the GB1 data set including a highly positive variant (HPV, corresponding to 5M for SrtA). Then, the two ML models were constructed with and without this HPV in the training data (HPV+ and HPV−). The top 100 predicted variants were evaluated using the fitness values provided by the GB1 data set (Figure S8). By this computational experiment, we confirmed the effects of HPV similar to those observed in the case of SrtA: the fitness was improved by not only HPV+ but also HPV− compared to the training data, and the sequences of the variants with the highest fitness were largely different between HPV+ and HPV− in most cases. These results suggest that our statement applies to not only SrtA but also GB1. Nonetheless, it remains to be addressed in future study to what extent this observation is general in other proteins.

In conclusion, our study demonstrated the importance of regulating the composition of training data in ML-guided directed evolution. Inclusion of a highly positive variant in training data is not always necessary for discovering improved variants. In addition, multiple executions of ML with subsets of training data may discover distinct regions of sequence space containing improved variants.

METHODS

Preparation of SrtA Variants for Screening. To generate the initial library, the five residues (94, 160, 165, 190, and 196) were mutated by the method described previously.[15] Briefly, point random mutagenesis at each residue and site-directed random mutagenesis at all five residues were performed by means of the 22c-trick method.[40] The gene fragments of the SrtA variants were generated from the plasmid containing the wild-type sequence by overlap extension PCR[41] using external and 22c-trick primers. The second- and third-round libraries proposed by ML were generated in a similar way using appropriate mix primers instead of 22c-trick primers. The variant gene fragments were ligated into pET22b vectors, and *E. coli* (DE3) bacteria were transformed with the resultant vectors. Colonies of transformed bacteria grown on agar media plates containing 100 μg/mL ampicillin were randomly picked up, incubated overnight at 37 °C in 1 mL of LB broth containing 100 μg/mL ampicillin in a deep-well plate (Axygen, CA, USA), and used for gene sequence analysis. A 100 μL aliquot of each cell culture was added into 900 μL of 2× YT broth supplemented with 100 μg/mL ampicillin in a deep-well plate and further cultured. When the optical density of the culture medium reached 0.6−0.8, isopropyl-1-thio-L-D-galactopyranoside was added to each well to a final concentration of 1 mM to induce SrtA variant expression, and the cells were further cultured for 3 h.

Screening of SrtA Variants in the Libraries. The harvested cells in each well were centrifuged, and the cell pellets were lysed in 200 μL of BugBuster Master Mix (Merck Millipore, Darmstadt, Germany). After centrifugation, the expressed SrtA variant in each supernatant was immobilized on an IMAC column, washed with 50 mM imidazole solution, and eluted with buffer A (50 mM Tris−HCl pH 8.0, 200 mM NaCl) containing 300 mM imidazole. Concentrations of the IMAC-purified SrtA variants were estimated by means of the Bradford method. LPETG-fused GFP and GGG-fused YFP were expressed in *E. coli* and purified by means of IMAC and SEC. For the screening assay of enzyme activity, 20 μM concentration of the two substrates (LPETG-fused GFP and GGG-fused YFP) were reacted in buffer A containing 5 mM $CaCl_2$ and 3 ng/μL IMAC-purified SrtA (wild type or variant) for 12 h at 37 °C. Then, 0.5 M ethylenediaminetetraacetic acid was added to stop the reaction, and a 10 μL aliquot of each solution was analyzed by SDS-PAGE. The amount of the conjugate (YFP-linked GFP) relative to the control protein was quantified with ImageQuant TL software.

To evaluate the enzyme activity of the selected SrtA variants in detail, the transformed cells were cultivated in a 1 L shake flask containing 500 mL of 2× YT broth supplemented with 100 μg/mL ampicillin. The expression of the SrtA variant was induced by adding 1 mM isopropyl-1-thio-L-D-galactopyranoside when the optical density of the culture medium reached 0.6, and the cells were grown overnight. The harvested cells were centrifuged, and the pellet was suspended in buffer A and ultrasonicated. The suspension was centrifuged, and the supernatant was purified by means of IMAC and SEC. Then 200 μM GGG-fused YFP was mixed with 0−120 μM LPETG-fused GFP in buffer A containing 5 mM $CaCl_2$ and 3 ng/μL purified SrtA, and the enzyme activity was measured as described for the SrtA screening assay above.

ML Model. We previously used the Bayesian optimization software COMBO[42] to conduct ML-guided directed evolution of fluorescent proteins.[15] Here, we applied a similar procedure for enhancing the enzyme activity of SrtA. COMBO implements a Gaussian process based on a linear regression model using a random feature map:

$$y = w^T\varphi(x) + \varepsilon \tag{1}$$

where y is the enzyme performance score of the protein (defined in the next section), x is a feature vector of the protein, $\varphi(x)$ is a random feature map (the dimensionality is 5000), and ε is an error term. As a kernel function, we used a radial basis function kernel whose hyperparameters were optimized by the type-2 maximum likelihood method implemented in COMBO. Given a training data set $\{(y, x)\}$, COMBO fits a weight vector w so that the enzyme performance score y can be predicted from the feature vector x. For each variant not included in the training data set, COMBO computes the acquisition function by Thompson sampling,[42] which evaluates the probability that the enzyme performance score of the variant is higher than any variant in the training data set. These values were used to rank all possible variants in sequence space.

Here, we performed two rounds of ML for each of the two scenarios (5M+ and 5M−). In the first round, two different training data sets were prepared (initial 5M+ and 5M− libraries in Table 1; Table S1) and separately used to construct two ML models. These ML models were used for ranking all possible variants, obtaining the two ranking lists (Data S1 for 5M+ and Data S2 for 5M−). In the second round, the top-ranked variants from the first-round ML were used as additional training data (Table S2) to update the ML models and the ranking lists (Data S3 for 5M+ and Data S4 for 5M−).

Enzyme Performance Score. To discover improved variants, the ML model needs to predict variants that have not only high enzyme activity but also expression levels sufficient for measuring enzyme activity. To satisfy this requirement, the following enzyme performance score was used in the ML model:

$$y = \sigma(\text{activity}_n - 1) \qquad (2)$$

where $\sigma(\cdot)$ is a sigmoidal function and activity_n is the enzyme activity of the protein divided by that of the wild type. Importantly, we set activity_n as zero if the expression level is insufficient for measuring enzyme activity. Thus, the enzyme performance score takes a high value when the enzyme activity is high while it takes the minimum value for variants with insufficient expression levels. This allows the ML model to predict high activity variants while avoiding variants with insufficient expression levels.

Feature Vector. We defined a feature vector x of the SrtA protein by concatenating the feature vectors of amino acids at the five mutated residues. For a feature vector of each amino acid, we considered a variety of amino acid descriptors including Z-scale,[31] T-scale,[43] ST-scale,[44] FASGAI,[45] MS-WHIM,[46] ProtFP,[47] VHSE,[48] and BLOSUM-based features.[49] The principles and properties of these descriptors are summarized in a review.[47] In addition, we used PSSM as a feature vector to incorporate the evolutionary information from SrtA homologs. Specifically, PSSM was constructed by homology search using PSI-BLAST[32] with the wild-type SrtA sequence as a query against the nr database. The homology search was iterated five times with 200 homologs used for updating PSSM per iteration. This produced PSSM $\{s_{ij}\}$ for each residue i and amino acid j (Data S5). We constituted a PSSM-based feature vector of the SrtA protein by concatenating the PSSM values at the five mutated residues and their amino acids.

To select the optimal feature vector, we performed a feature selection procedure like that in our previous study.[15] Briefly, we performed a benchmark experiment where COMBO was set to find 5M from the initial library by a Bayesian optimization procedure (Figure S7). In this experiment, MS-WHIM, Z-scale, and PSSM achieved better results than the other descriptors in terms of the number of training data needed to find 5M. These three descriptors were further compared in another benchmark experiment where COMBO was trained using the initial library with 5M excluded (i.e., initial 5M− library), and all possible variants in sequence space of 20^5 were ranked. In this experiment, the combination of Z-scale and PSSM achieved the best result in terms of the rank of 5M (Table S4). Therefore, we used the combination of Z-scale and PSSM as the feature vector for our final model in all other parts of this study.

Computational Experiment on the GB1 Data Set. The GB1 data set[39] provides the fitness values (binding activity) for all possible variants across four residues (20^4 = 160,000). Based on this data set, we prepared the training data consisting of the wild type and 100 selected variants with an HPV. These 100 variants also contained all possible point mutants (19 × 4 = 76 variants) and 23 variants randomly selected from the GB1 data set. The HPV was randomly selected from the 221 quadruple mutants whose fitness value is four times higher than the wild type. The ML models were constructed using the training data with and without the HPV (HPV+ and HPV−). For feature vectors, we used Z-scale and PSSM (Data S6), as in the case of SrtA. The fitness values of the predicted variants were obtained from the GB1 data set and compared with those of the training data. This computational experiment was repeated 10 times by different choices of the HPV and the above 23 variants in the training data.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscatal.1c03753.

SrtA variants in the initial library, second-round libraries, and third-round libraries and benchmark of feature vectors (Tables S1−S4), cross-linking reaction of GFP and YFP catalyzed by SrtA, measured enzyme activity of the SrtA variants in the second-round libraries and third-round libraries versus their ranks by ML prediction, sequence space analysis of the SrtA variants predicted by ML, comparison of the predicted enzyme performance scores of the SrtA variants, benchmark of feature vectors, and computational experiment on the GB1 data set (Figures S1−S8), and the legends for Data S1−S6 (PDF)

Ranking list by ML prediction using the initial 5M+ library as training data(XLSX)

Ranking list by ML prediction using the initial 5M− library as training data (XLSX)

Ranking list by ML prediction using the second-round 5M+ library as additional training data (XLSX)

Ranking list by ML prediction using the second-round 5M− library as additional training data (XLSX)

PSSM of SrtA homologs (XLSX)

PSSM of GB1 homologs (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Koji Tsuda − *Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; Center for Advanced Intelligence Project, RIKEN, Chuo-ku, Tokyo 103-0027, Japan; Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan;* orcid.org/0000-0002-4288-1606; Email: tsuda@k.u-tokyo.ac.jp

Mitsuo Umetsu − *Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan; Center for Advanced Intelligence Project, RIKEN, Chuo-ku, Tokyo 103-0027, Japan;* orcid.org/0000-0003-4390-0263; Email: mitsuo@tohoku.ac.jp

Authors

Yutaka Saito − *Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan; AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), Shinjuku-ku, Tokyo 169-8555, Japan; Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; Center for Advanced Intelligence Project, RIKEN, Chuo-ku, Tokyo 103-0027, Japan;* orcid.org/0000-0002-4853-0153

Misaki Oikawa − *Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan*

Takumi Sato − *Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan*

Hikaru Nakazawa − *Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan;* orcid.org/0000-0003-2785-237X

Tomoyuki Ito − *Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan*

Tomoshi Kameda − *Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan; Center for Advanced Intelligence Project, RIKEN, Chuo-ku, Tokyo 103-0027, Japan;* orcid.org/0000-0001-9508-5366

Complete contact information is available at: https://pubs.acs.org/10.1021/acscatal.1c03753

Author Contributions

Y.S. conducted the computational analysis. M.O., T.S., H.N., and T.I. conducted the experiments. Y.S., M.O., T.K., K.T., and M.U. participated in the data interpretation. Y.S. and M.U. wrote the paper. K.T. and M.U. conceived of the study and directed the project. All authors read and approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Ms. Miho Hosoya, Ms. Hiromi Ogata, and Ms. Yuri Ishigaki for experimental support. This work was partly supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) "Technologies for Smart Bio-industry and Agriculture" (funding agency: Bio-oriented Technology Research Advancement Institution, National Agriculture and Food Research Organization (NARO), Japan); Scientific Research Grant from the Japan Society for the Promotion of Science Research Fellowships (JP16H04570, JP16K14483, and JP20H00315 to M.U.) and the project "Development of the Key Technologies for the Next-Generation Artificial Intelligence/Robots" of the Ministry of Economy, Trade and Industry, Japan (to M.U.). Computations were partly performed on the NIG supercomputer at ROIS National Institute of Genetics, AI Bridging Cloud Infrastructure (ABCI) at the National Institute of Advanced Industrial Science and Technology (AIST), and the supercomputer of ACCMS at Kyoto University.

■ REFERENCES

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

(2) Wang, Y.; Xue, P.; Cao, M.; Yu, T.; Lane, S. T.; Zhao, H. Directed Evolution: Methodologies and Applications. *Chem. Rev.* **2021**, *121*, 12384.

(3) Markel, U.; Essani, K. D.; Besirlioglu, V.; Schiffels, J.; Streit, W. R.; Schwaneberg, U. Advances in ultrahigh-throughput screening for directed enzyme evolution. *Chem. Soc. Rev.* **2020**, *49*, 233−262.

(4) Bunzel, H. A.; Garrabou, X.; Pott, M.; Hilvert, D. Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Curr. Opin. Struct. Biol.* **2018**, *48*, 149−156.

(5) Karamitros, C. S.; Konrad, M. Fluorescence-Activated Cell Sorting of Human l-asparaginase Mutant Libraries for Detecting Enzyme Variants with Enhanced Activity. *ACS Chem. Biol.* **2016**, *11*, 2596−2607.

(6) Gianella, P.; Snapp, E. L.; Levy, M. An in vitro compartmentalization-based method for the selection of bond-forming enzymes from large libraries. *Biotechnol. Bioeng.* **2016**, *113*, 1647−1657.

(7) Reetz, M. T.; Wang, L. W.; Bocola, M. Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. *Angew. Chem. Int. Ed. Engl.* **2006**, *45*, 1236−1241.

(8) Acevedo-Rocha, C. G.; Sun, Z.; Reetz, M. T. Clarifying the Difference between Iterative Saturation Mutagenesis as a Rational Guide in Directed Evolution and OmniChange as a Gene Mutagenesis Technique. *ChemBioChem* **2018**, *19*, 2542−2544.

(9) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019**, *16*, 687−694.

(10) Chowdhury, R.; Maranas, C. D. From directed evolution to computational enzyme engineering—A review. *AIChE J.* **2020**, *121*, No. e16847.

(11) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10*, 1210−1223.

(12) Xu, Y.; Verma, D.; Sheridan, R. P.; Liaw, A.; Ma, J.; Marshall, N. M.; McIntosh, J.; Sherer, E. C.; Svetnik, V.; Johnston, J. M. Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* **2020**, *60*, 2773−2790.

(13) Siedhoff, N. E.; Illig, A. M.; Schwaneberg, U.; Davari, M. D. PyPEF—An Integrated Framework for Data-Driven Protein Engineering. *J. Chem. Inf. Model.* **2021**, *61*, 3463−3476.

(14) Zhu, L.; Davari, M. D.; Li, W. Recent Advances in the Prediction of Protein Structural Classes: Feature Descriptors and Machine Learning Algorithms. *Crystals* **2021**, *11*, 324.

(15) 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.

(16) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315−1322.

(17) Biswas, S.; Khimulya, G.; Alley, E. C.; Esvelt, K. M.; Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **2021**, *18*, 389−396.

(18) Liao, J.; Warmuth, M. K.; Govindarajan, S.; Ness, J. E.; Wang, R. P.; Gustafsson, C.; Minshull, J. Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnol.* **2017**, *27*, 16.

(19) 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.

(20) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *2116*, 8852−8858.

(21) Giguère, S.; Laviolette, F.; Marchand, M.; Tremblay, D.; Moineau, S.; Liang, X.; Biron, É.; Corbeil, J. Machine learning assisted design of highly active peptides for drug discovery. *PLoS Comput. Biol.* **2015**, *11*, No. e1004074.

(22) 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.

(23) Ton-That, H.; Liu, G.; Mazmanian, S. K.; Faull, K. F.; Schneewind, O. Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of Staphylococcus aureus at the LPXTG motif. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 12424−12429.

(24) Pishesha, N.; Ingram, J. R.; Ploegh, H. L. Sortase A: A Model for Transpeptidation and Its Biological Applications. *Annu. Rev. Cell Dev. Biol.* **2018**, *34*, 163−188.

(25) Chen, L.; Cohen, J.; Song, X.; Zhao, A.; Ye, Z.; Feulner, C. J.; Doonan, P.; Somers, W.; Lin, L.; Chen, P. R. Improved variants of SrtA for site-specific conjugation on antibodies and proteins with high efficiency. *Sci. Rep.* **2016**, *6*, 31899.

(26) Sellmann, C.; Doerner, A.; Knuehl, C.; Rasche, N.; Sood, V.; Krah, S.; Rhiel, L.; Messemer, A.; Wesolowski, J.; Schuette, M.; Becker, S.; Toleikis, L.; Kolmar, H.; Hock, B. Balancing Selectivity and Efficacy of Bispecific Epidermal Growth Factor Receptor (EGFR) × c-MET Antibodies and Antibody-Drug Conjugates. *J. Biol. Chem.* **2016**, *291*, 25106−25119.

(27) Xiaolin, D.; Böker, A.; Glebe, U. Broadening the scope of sortagging. *RSC Adv.* **2019**, *9*, 4700−4721.

(28) Matsumoto, T.; Furuta, K.; Tanaka, T.; Kondo, A. Sortase A-Mediated Metabolic Enzyme Ligation in Escherichia coli. *ACS Synth. Biol.* **2016**, *5*, 1284−1289.

(29) Chen, I.; Dorr, B. M.; Liu, D. R. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 11399−11404.

(30) Piotukh, K.; Geltinger, B.; Heinrich, N.; Gerth, F.; Beyermann, M.; Freund, C.; Schwarzer, D. Directed evolution of sortase A mutants with altered substrate selectivity profiles. *J. Am. Chem. Soc.* **2011**, *133*, 17536−17539.

(31) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481−2491.

(32) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(33) Liu, G.; Zeng, H.; Mueller, J.; Carter, B.; Wang, Z.; Schilz, J.; Horny, G.; Birnbaum, M. E.; Ewert, S.; Gifford, D. K. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **2020**, *36*, 2126−2133.

(34) Mason, D. M.; Friedensohn, S.; Weber, C. R.; Jordi, C.; Wagner, B.; Meng, S. M.; Ehling, R. A.; Bonati, L.; Dahinden, J.; Gainza, P.; Correia, B. E.; Reddy, S. T. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **2021**, *5*, 600−612.

(35) Bryant, D. H.; Bashir, A.; Sinai, S.; Jain, N. K.; Ogden, P. J.; Riley, P. F.; Church, G. M.; Colwell, L. J.; Kelsic, E. D. Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **2021**, *39*, 691−696.

(36) Heck, T.; Pham, P. H.; Yerlikaya, A.; Thöny-Meyera, L.; Richter, M. Sortase A catalyzed reaction pathways: a comparative study with six SrtA variants. *Catal. Sci. Technol.* **2014**, *4*, 2946−2956.

(37) Antos, J. M.; Truttmann, M. C.; Ploegh, H. L. Recent advances in sortase-catalyzed ligation methodology. *Curr. Opin. Struct. Biol.* **2016**, *38*, 111−118.

(38) Suree, N.; Liew, C. K.; Villareal, V. A.; Thieu, W.; Fadeev, E. A.; Clemens, J. J.; Jung, M. E.; Clubb, R. T. The structure of the Staphylococcus aureus sortase-substrate complex reveals how the universally conserved LPXTG sorting signal is recognized. *J. Biol. Chem.* **2009**, *284*, 24465−24477.

(39) Wu, N. C.; Dai, L.; Olson, C. A.; Lloyd-Smith, J. O.; Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* **2016**, *5*, No. e16965.

(40) Kille, S.; Acevedo-Rocha, C. G.; Parra, L. P.; Zhang, Z. G.; Opperman, D. J.; Reetz, M. T.; Acevedo, J. P. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol.* **2013**, *2*, 83−92.

(41) Sato, K.; Tsuchiya, M.; Saldanha, J.; Koishihara, Y.; Ohsugi, Y.; Kishimoto, T.; Bendig, M. M. Humanization of a mouse anti-human interleukin-6 receptor antibody comparing two methods for selecting human framework regions. *Mol. Immunol.* **1994**, *31*, 371−381.

(42) Ueno, T.; Rhone, T. D.; Hou, Z.; Mizoguchi, T.; Tsuda, K. COMBO: an efficient Bayesian optimization library for materials science. *Mater. Discovery* **2016**, *4*, 18−21.

(43) Tian, F.; Zhou, P.; Li, Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *J. Mol. Struct.* **2007**, *830*, 106−115.

(44) Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids* **2010**, *38*, 805−816.

(45) Liang, G.; Li, Z. Factor analysis scale of generalized amino acid information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides. *QSAR Comb. Sci.* **2007**, *26*, 754−763.

(46) Zaliani, A.; Gancia, E. MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525−533.

(47) van Westen, G. J.; Swier, R. F.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W.; Bender, A. Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *Aust. J. Chem.* **2013**, *5*, 42.

(48) Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers* **2005**, *80*, 775−786.

(49) Georgiev, A. G. Interpretable numerical descriptors of amino acid space. *J. Comput. Biol.* **2009**, *16*, 703−723.